Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly under-trained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, *Chinchilla*, that uses the same compute budget as *Gopher* but with 70B parameters and 4× more more data. *Chinchilla* uniformly and significantly outperforms *Gopher* (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks. This also means that *Chinchilla* uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, *Chinchilla* reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, greater than a 7% improvement over *Gopher*.

arXiv:2203.15556v1 [cs.CL] 29 Mar 2022

1. Introduction

Recently a series of *Large Language Models* (LLMs) have been introduced (Brown et al., 2020; Lieber et al., 2021; Rae et al., 2021; Smith et al., 2022; Thoppilan et al., 2022), with the largest dense language models now having over 500 billion parameters. These large autoregressive transformers (Vaswani et al., 2017) have demonstrated impressive performance on many tasks using a variety of evaluation protocols such as zero-shot, few-shot, and fine-tuning.

The compute and energy cost for training large language models is substantial (Rae et al., 2021; Thoppilan et al., 2022) and rises with increasing model size. In practice, the allocated training compute budget is often known in advance: how many accelerators are available and for how long we want to use them. Since it is typically only feasible to train these large models once, accurately estimating the best model hyperparameters for a given compute budget is critical (Tay et al., 2021).

Kaplan et al. (2020) showed that there is a power law relationship between the number of parameters in an autoregressive language model (LM) and its performance. As a result, the field has been training larger and larger models, expecting performance improvements. One notable conclusion in Kaplan et al. (2020) is that large models should not be trained to their lowest possible loss to be compute optimal. Whilst we reach the same conclusion, we estimate that large models should be trained for many more training tokens than recommended by the authors. Specifically, given a 10× increase computational budget, they suggests that the size of the model should increase 5.5× while the number of training tokens should only increase 1.8×. Instead, we find that model size and the number of training tokens should be scaled in equal proportions.

Following Kaplan et al. (2020) and the training setup of GPT-3 (Brown et al., 2020), many of the recently trained large models have been trained for approximately 300 billion tokens (Table 1), in line with the approach of predominantly increasing model size when increasing compute.

Corresponding authors: {jordanhoffmann|sborgeaud|amensch|sifre}@deepmind.com
© 2023 DeepMind. All rights reserved

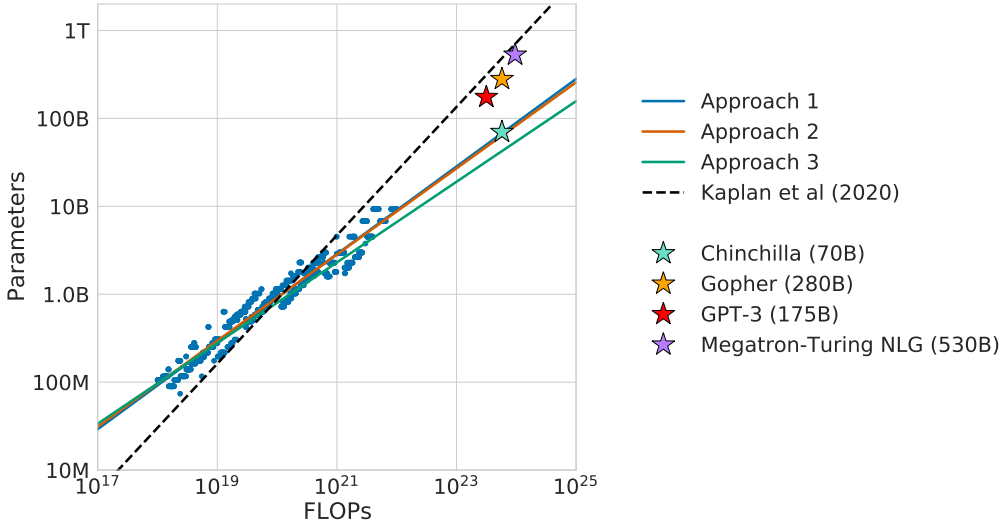


Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from Kaplan et al. (2020). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In Figure A3, we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. ***Chinchilla* outperforms *Gopher* and the other large models (see Section 4.2).**

In this work, we revisit the question: *Given a fixed FLOPs budget,*[1] *how should one trade-off model size and the number of training tokens?* To answer this question, we model the final pre-training loss[2] $L(N, D)$ as a function of the number of model parameters N, and the number of training tokens, D. Since the computational budget C is a deterministic function $\text{FLOPs}(N, D)$ of the number of seen training tokens and model parameters, we are interested in minimizing L under the constraint $\text{FLOPs}(N, D) = C$:

$$N_{opt}(C), D_{opt}(C) = \underset{N,D \text{ s.t. } \text{FLOPs}(N,D)=C}{\text{argmin}} L(N, D). \tag{1}$$

The functions $N_{opt}(C)$, and $D_{opt}(C)$ describe the optimal allocation of a computational budget C. We empirically estimate these functions based on the losses of over 400 models, ranging from under 70M to over 16B parameters, and trained on 5B to over 400B tokens – with each model configuration trained for several different training horizons. Our approach leads to considerably different results than that of Kaplan et al. (2020). We highlight our results in Figure 1 and how our approaches differ in Section 2.

Based on our estimated compute-optimal frontier, we predict that for the compute budget used to train *Gopher*, an optimal model should be 4 times smaller, while being training on 4 times more tokens. We verify this by training a more *compute-optimal* 70B model, called *Chinchilla*, on 1.4 trillion tokens. Not only does *Chinchilla* outperform its much larger counterpart, *Gopher*, but its reduced model size reduces inference cost considerably and greatly facilitates downstream uses on smaller hardware. The energy cost of a large language model is amortized through its usage for inference an fine-tuning. The benefits of a more optimally trained smaller model, therefore, extend beyond the immediate benefits of its improved performance.

[1] For example, knowing the number of accelerators and a target training duration.

[2] For simplicity, we perform our analysis on the smoothed training loss which is an unbiased estimate of the test loss, as we are in the infinite data regime (the number of training tokens is less than the number of tokens in the entire corpus).

Table 1 | **Current LLMs**. We show five of the current largest dense transformer models, their size, and the number of training tokens. Other than LaMDA (Thoppilan et al., 2022), most models are trained for approximately 300 billion tokens. We introduce *Chinchilla*, a substantially smaller model, trained for much longer than 300B tokens.

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

2. Related Work

Large language models. A variety of large language models have been introduced in the last few years. These include both dense transformer models (Brown et al., 2020; Lieber et al., 2021; Rae et al., 2021; Smith et al., 2022; Thoppilan et al., 2022) and mixture-of-expert (MoE) models (Du et al., 2021; Fedus et al., 2021; Zoph et al., 2022). The largest dense transformers have passed 500 billion parameters (Smith et al., 2022). The drive to train larger and larger models is clear—so far increasing the size of language models has been responsible for improving the state-of-the-art in many language modelling tasks. Nonetheless, large language models face several challenges, including their overwhelming computational requirements (the cost of training and inference increase with model size) (Rae et al., 2021; Thoppilan et al., 2022) and the need for acquiring more high-quality training data. In fact, in this work we find that larger, high quality datasets will play a key role in any further scaling of language models.

Modelling the scaling behavior. Understanding the scaling behaviour of language models and their transfer properties has been important in the development of recent large models (Hernandez et al., 2021; Kaplan et al., 2020). Kaplan et al. (2020) first showed a predictable relationship between model size and loss over many orders of magnitude. The authors investigate the question of choosing the optimal model size to train for a given compute budget. Similar to us, they address this question by training various models. Our work differs from Kaplan et al. (2020) in several important ways. First, the authors use a fixed number of training tokens and learning rate schedule for all models; this prevents them from modelling the impact of these hyperparameters on the loss. In contrast, we find that setting the learning rate schedule to approximately match the number of training tokens results in the best final loss regardless of model size—see Figure A1. For a fixed learning rate cosine schedule to 130B tokens, the intermediate loss estimates (for $D' << 130\text{B}$) are therefore overestimates of the loss of a model trained with a schedule length matching D'. Using these intermediate losses results in underestimating the effectiveness of training models on less data than 130B tokens, and eventually contributes to the conclusion that model size should increase faster than training data size as compute budget increases. In contrast, our analysis predicts that both quantities should scale at roughly the same rate. Secondly, we include models with up to 16B parameters, as we observe that there is slight curvature in the FLOP-loss frontier (see Appendix E)—in fact, the majority of the models used in our analysis have more than 500 million parameters, in contrast the majority of runs in Kaplan et al. (2020) are significantly smaller—many being less than 100M parameters.

Recently, Clark et al. (2022) specifically looked in to the scaling properties of Mixture of Expert

language models, showing that the scaling with number of experts diminishes as the model size increases—their approach models the loss as a function of two variables: the model size and the number of experts. However, the analysis is done with a fixed number of training tokens, as in Kaplan et al. (2020), potentially underestimating the improvements of branching.

Estimating hyperparameters for large models. The model size and the number of training tokens are not the only two parameters to chose when selecting a language model and a procedure to train it. Other important factors include learning rate, learning rate schedule, batch size, optimiser, and width-to-depth ratio. In this work, we focus on model size and the number of training steps, and we rely on existing work and provided experimental heuristics to determine the other necessary hyperparameters. Yang et al. (2021) investigates how to choose a variety of these parameters for training an autoregressive transformer, including the learning rate and batch size. McCandlish et al. (2018) finds only a weak dependence between optimal batch size and model size. Shallue et al. (2018); Zhang et al. (2019) suggest that using larger batch-sizes than those we use is possible. Levine et al. (2020) investigates the optimal depth-to-width ratio for a variety of standard model sizes. We use slightly less deep models than proposed as this translates to better wall-clock performance on our hardware.

Improved model architectures. Recently, various promising alternatives to traditional dense transformers have been proposed. For example, through the use of conditional computation large MoE models like the 1.7 trillion parameter Switch transformer (Fedus et al., 2021), the 1.2 Trillion parameter GLaM model (Du et al., 2021), and others (Artetxe et al., 2021; Zoph et al., 2022) are able to provide a large effective model size despite using relatively fewer training and inference FLOPs. However, for very large models the computational benefits of routed models seems to diminish (Clark et al., 2022). An orthogonal approach to improving language models is to augment transformers with explicit retrieval mechanisms, as done by Borgeaud et al. (2021); Guu et al. (2020); Lewis et al. (2020). This approach effectively increases the number of data tokens seen during training (by a factor of ~ 10 in Borgeaud et al. (2021)). This suggests that the performance of language models may be more dependant on the size of the training data than previously thought.

3. Estimating the optimal parameter/training tokens allocation

We present three different approaches to answer the question driving our research: *Given a fixed FLOPs budget, how should one trade-off model size and the number of training tokens?* In all three cases we start by training a range of models varying both model size and the number of training tokens and use the resulting training curves to fit an empirical estimator of how they should scale. We assume a power-law relationship between compute and model size as done in Clark et al. (2022); Kaplan et al. (2020), though future work may want to include potential curvature in this relationship for large model sizes. The resulting predictions are similar for all three methods and suggest that parameter count and number of training tokens should be increased equally with more compute[3]—with proportions reported in Table 2. This is in clear contrast to previous work on this topic and warrants further investigation.

[3]We compute FLOPs as described in Appendix F.

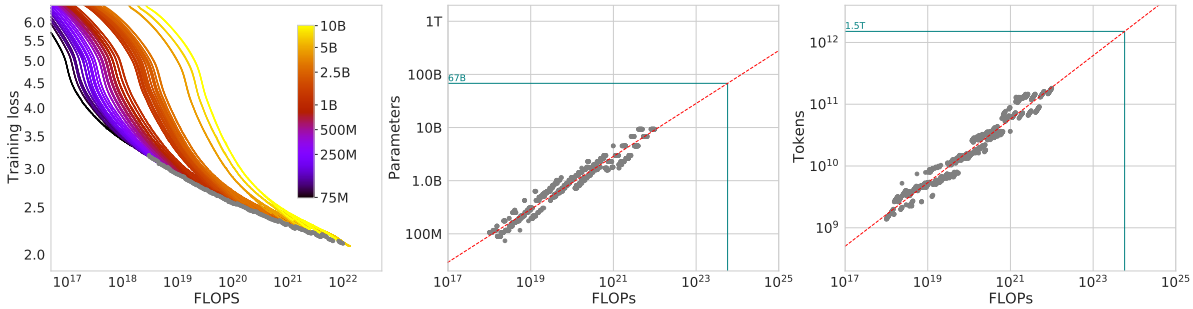


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate the optimal model size (**center**) for a given compute budget and the optimal number of training tokens (**right**). In green, we show projections of optimal model size and training token count based on the number of FLOPs used to train *Gopher* (5.76×10^{23}).

3.1. Approach 1: Fix model sizes and vary number of training tokens

In our first approach we vary the number of training steps for a fixed family of models (ranging from 70M to over 10B parameters), training each model for 4 different number of training sequences. From these runs, we are able to directly extract an estimate of the minimum loss achieved for a given number of training FLOPs. Training details for this approach can be found in Appendix D.

For each parameter count N we train 4 different models, decaying the learning rate by a factor of 10× over a horizon (measured in number of training tokens) that ranges by a factor of 16×. Then, for each run, we smooth and then interpolate the training loss curve. From this, we obtain a continuous mapping from FLOP count to training loss for each run. Then, for each FLOP count, we determine which run achieves the lowest loss. Using these interpolants, we obtain a mapping from any FLOP count C, to the most efficient choice of model size N and number of training tokens D such that $\text{FLOPs}(N, D) = C$.[4] At 1500 logarithmically spaced FLOP values, we find which model size achieves the lowest loss of all models along with the required number of training tokens. Finally, we fit power laws to estimate the optimal model size and number of training tokens for any given amount of compute (see the center and right panels of Figure 2), obtaining a relationship $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$. We find that $a = 0.50$ and $b = 0.50$—as summarized in Table 2. In Section D.4, we show a head-to-head comparison at 10^{21} FLOPs, using the model size recommended by our analysis and by the analysis of Kaplan et al. (2020)—using the model size we predict has a clear advantage.

3.2. Approach 2: IsoFLOP profiles

In our second approach we vary the model size[5] for a fixed set of 9 different training FLOP counts[6] (ranging from 6×10^{18} to 3×10^{21} FLOPs), and consider the final training loss for each point[7]. in contrast with Approach 1 that considered points (N, D, L) along the entire training runs. This allows us to directly answer the question: For a given FLOP budget, what is the optimal parameter count?

[4]Note that all selected points are within the last 15% of training. This suggests that when training a model over D tokens, we should pick a cosine cycle length that decays 10× over approximately D tokens—see further details in Appendix B.

[5]In approach 2, model size varies up to 16B as opposed to approach 1 where we only used models up to 10B.

[6]The number of training tokens is determined by the model size and training FLOPs.

[7]We set the cosine schedule length to match the number of tokens, which is optimal according to the analysis presented in Appendix B.

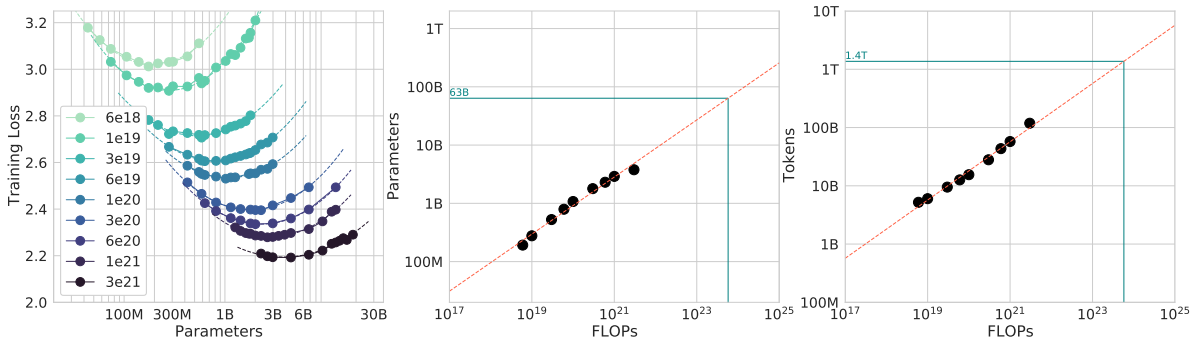


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

For each FLOP budget, we plot the final loss (after smoothing) against the parameter count in Figure 3 (left). In all cases, we ensure that we have trained a diverse enough set of model sizes to see a clear minimum in the loss. We fit a parabola to each IsoFLOPs curve to directly estimate at what model size the minimum loss is achieved (Figure 3 (left)). As with the previous approach, we then fit a power law between FLOPs and loss-optimal model size and number of training tokens, shown in Figure 3 (center, right). Again, we fit exponents of the form $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$ and we find that $a = 0.49$ and $b = 0.51$—as summarized in Table 2.

3.3. Approach 3: Fitting a parametric loss function

Lastly, we model all final losses from experiments in Approach 1 & 2 as a parametric function of model parameter count and the number of seen tokens. Following a classical risk decomposition (see Section D.2), we propose the following functional form

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}. \tag{2}$$

The first term captures the loss for an ideal generative process on the data distribution, and should correspond to the entropy of natural text. The second term captures the fact that a perfectly trained transformer with N parameters underperforms the ideal generative process. The final term captures the fact that the transformer is not trained to convergence, as we only make a finite number of optimisation steps, on a sample of the dataset distribution.

Model fitting. To estimate (A, B, E, α, β), we minimize the Huber loss (Huber, 1964) between the predicted and observed log loss using the L-BFGS algorithm (Nocedal, 1980):

$$\min_{A,B,E,\alpha,\beta} \sum_{\text{Runs } i} \text{Huber}_\delta\Big(\log \hat{L}(N_i, D_i) - \log L_i\Big) \tag{3}$$

We account for possible local minima by selecting the best fit from a grid of initialisations. The Huber loss ($\delta = 10^{-3}$) is robust to outliers, which we find important for good predictive performance over held-out data points. Section D.2 details the fitting procedure and the loss decomposition.

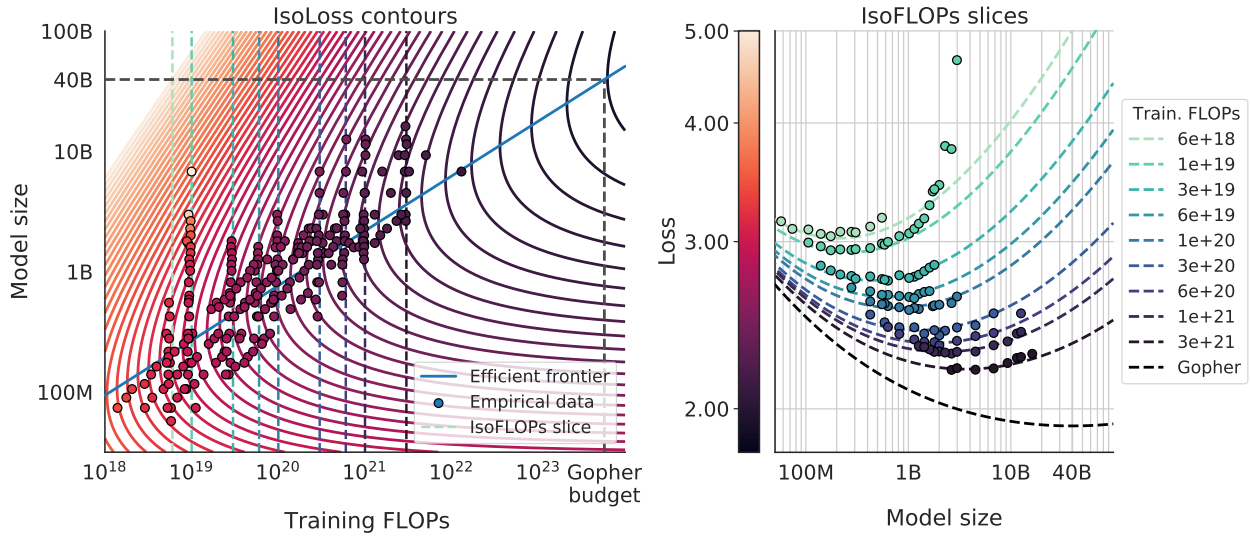


Figure 4 | **Parametric fit.** We fit a parametric modelling of the loss $\hat{L}(N, D)$ and display contour (**left**) and isoFLOP slices (**right**). For each isoFLOP slice, we include a corresponding dashed line in the left plot. In the left plot, we show the efficient frontier in blue, which is a line in log-log space. Specifically, the curve goes through each iso-loss contour at the point with the fewest FLOPs. We project the optimal model size given the *Gopher* FLOP budget to be 40B parameters.

Efficient frontier. We can approximate the functions N_{opt} and D_{opt} by minimizing the parametric loss $\hat{L}$ under the constraint $\text{FLOPs}(N, D) \approx 6ND$ (Kaplan et al., 2020). The resulting N_{opt} and D_{opt} balance the two terms in Equation (3) that depend on model size and data. By construction, they have a power-law form:

$$N_{opt}(C) = G\left(\frac{C}{6}\right)^{a}, \quad D_{opt}(C) = G^{-1}\left(\frac{C}{6}\right)^{b}, \quad \text{where} \quad G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}}, \quad a = \frac{\beta}{\alpha+\beta}, \text{ and } b = \frac{\alpha}{\alpha+\beta}. \tag{4}$$

We show contours of the fitted function $\hat{L}$ in Figure 4 (left), and the closed-form efficient computational frontier in blue. From this approach, we find that $a = 0.46$ and $b = 0.54$—as summarized in Table 2.

3.4. Optimal model scaling

We find that the three approaches, despite using different fitting methodologies and different trained models, yield comparable predictions for the optimal scaling in parameters and tokens with FLOPs (shown in Table 2). All three approaches suggest that as compute budget increases, model size and the amount of training data should be increased in approximately equal proportions. The first and second approaches yield very similar predictions for optimal model sizes, as shown in Figure 1 and Figure A3. The third approach predicts even smaller models being optimal at larger compute budgets. We note that the observed points (L, N, D) for low training FLOPs ($C \leqslant 1e21$) have larger residuals $\|L - \hat{L}(N, D)\|_2^2$ than points with higher computational budgets. The fitted model places increased weight on the points with more FLOPs—automatically considering the low-computational budget points as outliers due to the Huber loss. As a consequence of the empirically observed negative curvature in the frontier $C \to N_{opt}$ (see Appendix E), this results in predicting a lower N_{opt} than the two other approaches.

In Table 3 we show the estimated number of FLOPs and tokens that would ensure that a model of a given size lies on the compute-optimal frontier. Our findings suggests that the current generation of

Table 2 | **Estimated parameter and data scaling with increased training compute.** The listed values are the exponents, a and b, on the relationship $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$. Our analysis suggests a near equal scaling in parameters and data with increasing compute which is in clear contrast to previous work on the scaling of large models. The 10th and 90th percentiles are estimated via bootstrapping data (80% of the dataset is sampled 100 times) and are shown in parenthesis.

Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Table 3 | **Estimated optimal training FLOPs and training tokens for various model sizes.** For various model sizes, we show the projections from Approach 1 of how many FLOPs and training tokens would be needed to train compute-optimal models. The estimates for Approach 2 & 3 are similar (shown in Section D.3)

Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

large language models are considerably over-sized, given their respective compute budgets, as shown in Figure 1. For example, we find that a 175 billion parameter model should be trained with a compute budget of 4.41×10^{24} FLOPs and on over 4.2 trillion tokens. A 280 billion *Gopher*-like model is the optimal model to train given a compute budget of approximately 10^{25} FLOPs and should be trained on 6.8 trillion tokens. Unless one has a compute budget of 10^{26} FLOPs (over 250× the compute used to train *Gopher*), a 1 trillion parameter model is unlikely to be the optimal model to train. Furthermore, the amount of training data that is projected to be needed is far beyond what is currently used to train large models, and underscores the importance of dataset collection in addition to engineering improvements that allow for model scale. While there is significant uncertainty extrapolating out many orders of magnitude, our analysis clearly suggests that given the training compute budget for many current LLMs, smaller models should have been trained on more tokens to achieve the most performant model.

In Appendix C, we reproduce the IsoFLOP analysis on two additional datasets: C4 (Raffel et al., 2020a) and GitHub code (Rae et al., 2021). In both cases we reach the similar conclusion that model size and number of training tokens should be scaled in equal proportions.

4. *Chinchilla*

Based on our analysis in Section 3, the optimal model size for the *Gopher* compute budget is somewhere between 40 and 70 billion parameters. We test this hypothesis by training a model on the larger end of this range—70B parameters—for 1.4T tokens, due to both dataset and computational efficiency considerations. In this section we compare this model, which we call *Chinchilla*, to *Gopher* and other LLMs. Both *Chinchilla* and *Gopher* have been trained for the same number of FLOPs but differ in the size of the model and the number of training tokens.

While pre-training a large language model has a considerable compute cost, downstream fine-tuning and inference also make up substantial compute usage (Rae et al., 2021). Due to being 4× smaller than *Gopher*, both the memory footprint and inference cost of *Chinchilla* are also smaller.

4.1. Model and training details

The full set of hyperparameters used to train *Chinchilla* are given in Table 4. *Chinchilla* uses the same model architecture and training setup as *Gopher* with the exception of the differences listed below.

- We train *Chinchilla* on *MassiveText* (the same dataset as *Gopher*) but use a slightly different subset distribution (shown in Table A1) to account for the increased number of training tokens.
- We use AdamW (Loshchilov and Hutter, 2019) for *Chinchilla* rather than Adam (Kingma and Ba, 2014) as this improves the language modelling loss and the downstream task performance after finetuning.[8]
- We train *Chinchilla* with a slightly modified SentencePiece (Kudo and Richardson, 2018) tokenizer that does not apply NFKC normalisation. The vocabulary is very similar– 94.15% of tokens are the same as those used for training *Gopher*. We find that this particularly helps with the representation of mathematics and chemistry, for example.
- Whilst the forward and backward pass are computed in `bfloat16`, we store a `float32` copy of the weights in the distributed optimiser state (Rajbhandari et al., 2020). See *Lessons Learned* from Rae et al. (2021) for additional details.

In Appendix G we show the impact of the various optimiser related changes between *Chinchilla* and *Gopher*. All models in this analysis have been trained on TPUv3/TPUv4 (Jouppi et al., 2017) with JAX (Bradbury et al., 2018) and Haiku (Hennigan et al., 2020). We include a *Chinchilla* model card (Mitchell et al., 2019) in Table A8.

Model	Layers	Number Heads	Key/Value Size	d_{model}	Max LR	Batch Size
Gopher 280B	80	128	128	16,384	4×10^{-5}	3M $\rightarrow$ 6M
Chinchilla 70B	80	64	128	8,192	1×10^{-4}	1.5M $\rightarrow$ 3M

Table 4 | ***Chinchilla* architecture details.** We list the number of layers, the key/value size, the bottleneck activation size d_{model}, the maximum learning rate, and the training batch size (# tokens). The feed-forward size is always set to $4 \times d_{model}$. Note that we double the batch size midway through training for both *Chinchilla* and *Gopher*.

[8]Interestingly, a model trained with AdamW only passes the training performance of a model trained with Adam around 80% of the way through the cosine cycle, though the ending performance is notably better– see Figure A7

	# Tasks	Examples
Language Modelling	20	WikiText-103, The Pile: PG-19, arXiv, FreeLaw, . . .
Reading Comprehension	3	RACE-m, RACE-h, LAMBADA
Question Answering	3	Natural Questions, TriviaQA, TruthfulQA
Common Sense	5	HellaSwag, Winogrande, PIQA, SIQA, BoolQ
MMLU	57	High School Chemistry, Astronomy, Clinical Knowledge, . . .
BIG-bench	62	Causal Judgement, Epistemic Reasoning, Temporal Sequences, . . .

Table 5 | **All evaluation tasks.** We evaluate *Chinchilla* on a collection of language modelling along with downstream tasks. We evaluate on largely the same tasks as in Rae et al. (2021), to allow for direct comparison.

4.2. Results

We perform an extensive evaluation of *Chinchilla*, comparing against various large language models. We evaluate on a large subset of the tasks presented in Rae et al. (2021), shown in Table 5. As the focus of this work is on optimal model scaling, we included a large representative subset, and introduce a few new evaluations to allow for better comparison to other existing large models. The evaluation details for all tasks are the same as described in Rae et al. (2021).

4.2.1. Language modelling

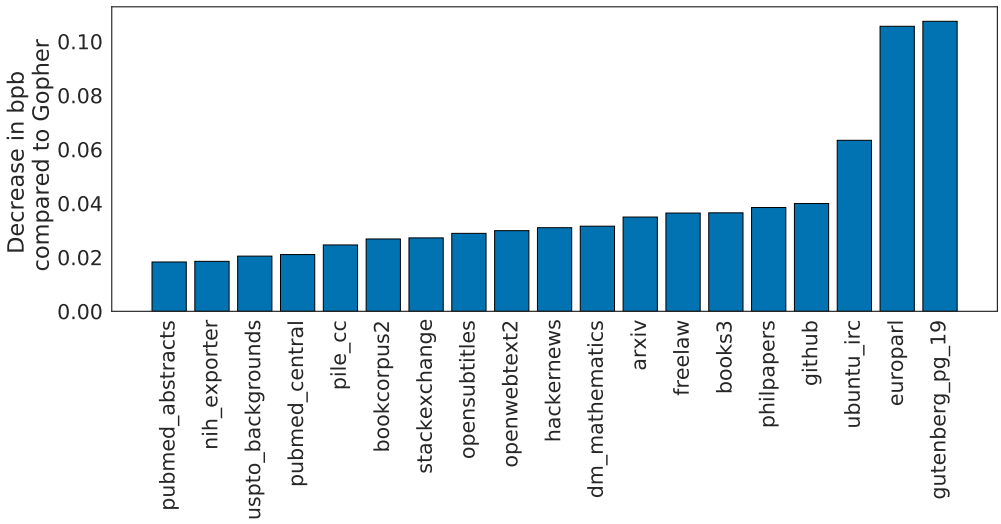


Figure 5 | **Pile Evaluation.** For the different evaluation sets in The Pile (Gao et al., 2020), we show the bits-per-byte (bpb) improvement (decrease) of *Chinchilla* compared to *Gopher*. On all subsets, *Chinchilla* outperforms *Gopher*.

Chinchilla significantly outperforms *Gopher* on all evaluation subsets of The Pile (Gao et al., 2020), as shown in Figure 5. Compared to Jurassic-1 (178B) Lieber et al. (2021), *Chinchilla* is more performant on all but two subsets– `dm_mathematics` and `ubuntu_irc`– see Table A5 for a raw bits-per-byte comparison. On Wikitext103 (Merity et al., 2017), *Chinchilla* achieves a perplexity of 7.16 compared to 7.75 for *Gopher*. Some caution is needed when comparing *Chinchilla* with *Gopher* on these language modelling benchmarks as *Chinchilla* is trained on 4× more data than *Gopher* and thus train/test set leakage may artificially enhance the results. We thus place more emphasis on other

Random	25.0%
Average human rater	34.5%
GPT-3 5-shot	43.9%
Gopher 5-shot	60.0%
***Chinchilla* 5-shot**	**67.6%**
Average human expert performance	*89.8%*
June 2022 Forecast	57.1%
June 2023 Forecast	63.4%

Table 6 | **Massive Multitask Language Understanding (MMLU).** We report the average 5-shot accuracy over 57 tasks with model and human accuracy comparisons taken from Hendrycks et al. (2020). We also include the average prediction for state of the art accuracy in June 2022/2023 made by 73 competitive human forecasters in Steinhardt (2021).

tasks for which leakage is less of a concern, such as MMLU (Hendrycks et al., 2020) and BIG-bench (BIG-bench collaboration, 2021) along with various closed-book question answering and common sense analyses.

4.2.2. MMLU

The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020) consists of a range of exam-like questions on academic subjects. In Table 6, we report *Chinchilla*'s average 5-shot performance on MMLU (the full breakdown of results is shown in Table A6). On this benchmark, *Chinchilla* significantly outperforms *Gopher* despite being much smaller, with an average accuracy of 67.6% (improving upon *Gopher* by 7.6%). Remarkably, *Chinchilla* even outperforms the expert forecast for June 2023 of 63.4% accuracy (see Table 6) (Steinhardt, 2021). Furthermore, *Chinchilla* achieves greater than 90% accuracy on 4 different individual tasks– `high_school_gov_and_politics`, `international_law`, `sociology`, and `us_foreign_policy`. To our knowledge, no other model has achieved greater than 90% accuracy on a subset.

In Figure 6, we show a comparison to *Gopher* broken down by task. Overall, we find that *Chinchilla* improves performance on the vast majority of tasks. On four tasks (`college_mathematics`, `econometrics`, `moral_scenarios`, and `formal_logic`) *Chinchilla* underperforms *Gopher*, and there is no change in performance on two tasks.

4.2.3. Reading comprehension

On the final word prediction dataset LAMBADA (Paperno et al., 2016), *Chinchilla* achieves 77.4% accuracy, compared to 74.5% accuracy from *Gopher* and 76.6% from MT-NLG 530B (see Table 7). On RACE-h and RACE-m (Lai et al., 2017), *Chinchilla* greatly outperforms *Gopher*, improving accuracy by more than 10% in both cases—see Table 7.

4.2.4. BIG-bench

We analysed *Chinchilla* on the same set of BIG-bench tasks (BIG-bench collaboration, 2021) reported in Rae et al. (2021). Similar to what we observed in MMLU, *Chinchilla* outperforms *Gopher* on the vast majority of tasks (see Figure 7). We find that *Chinchilla* improves the average performance by 10.7%, reaching an accuracy of 65.1% versus 54.4% for *Gopher*. Of the 62 tasks we consider, *Chinchilla* performs worse than *Gopher* on only four—`crash_blossom`, `dark_humor_detection`,

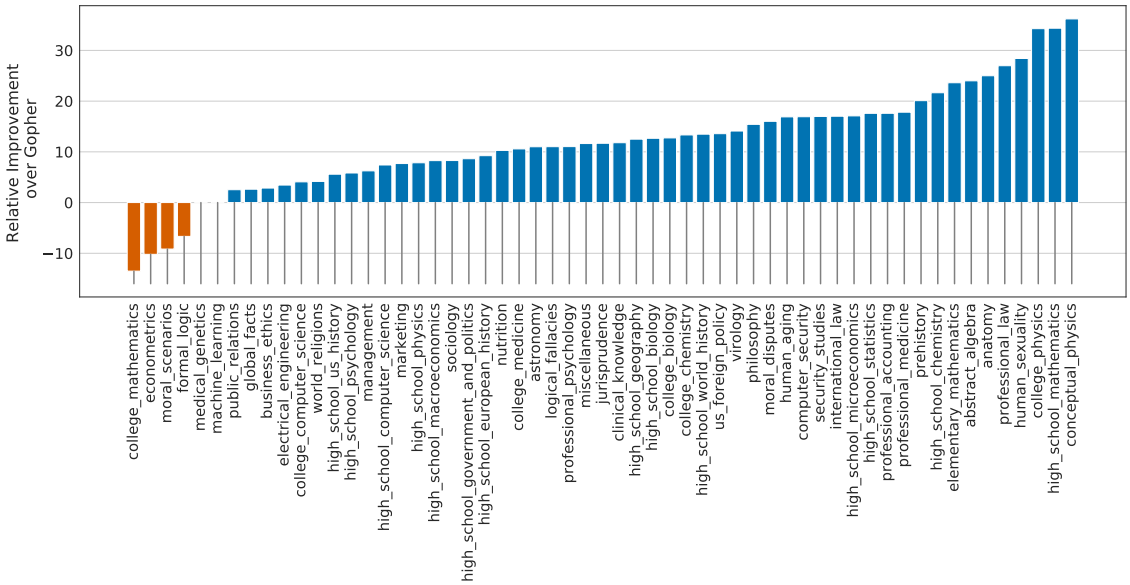


Figure 6 | **MMLU results compared to *Gopher*** We find that *Chinchilla* outperforms *Gopher* by 7.6% on average (see Table 6) in addition to performing better on 51/57 individual tasks, the same on 2/57, and worse on only 4/57 tasks.

	Chinchilla	*Gopher*	GPT-3	MT-NLG 530B
LAMBADA Zero-Shot	**77.4**	74.5	76.2	76.6
RACE-m Few-Shot	**86.8**	75.1	58.1	-
RACE-h Few-Shot	**82.3**	71.6	46.8	47.9

Table 7 | **Reading comprehension.** On RACE-h and RACE-m (Lai et al., 2017), *Chinchilla* considerably improves performance over *Gopher*. Note that GPT-3 and MT-NLG 530B use a different prompt format than we do on RACE-h/m, so results are not comparable to *Gopher* and *Chinchilla*. On LAMBADA (Paperno et al., 2016), *Chinchilla* outperforms both *Gopher* and MT-NLG 530B.

`mathematical_induction` and `logical_args`. Full accuracy results for *Chinchilla* can be found in Table A7.

4.2.5. Common sense

We evaluate *Chinchilla* on various common sense benchmarks: PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), Winogrande (Sakaguchi et al., 2020), HellaSwag (Zellers et al., 2019), and BoolQ (Clark et al., 2019). We find that *Chinchilla* outperforms both *Gopher* and GPT-3 on all tasks and outperforms MT-NLG 530B on all but one task—see Table 8.

On TruthfulQA (Lin et al., 2021), *Chinchilla* reaches 43.6%, 58.5%, and 66.7% accuracy with 0-shot, 5-shot, and 10-shot respectively. In comparison, *Gopher* achieved only 29.5% 0-shot and 43.7% 10-shot accuracy. In stark contrast with the findings of Lin et al. (2021), the large improvements (14.1% in 0-shot accuracy) achieved by Chinchilla suggest that better modelling of the pre-training data alone can lead to substantial improvements on this benchmark.

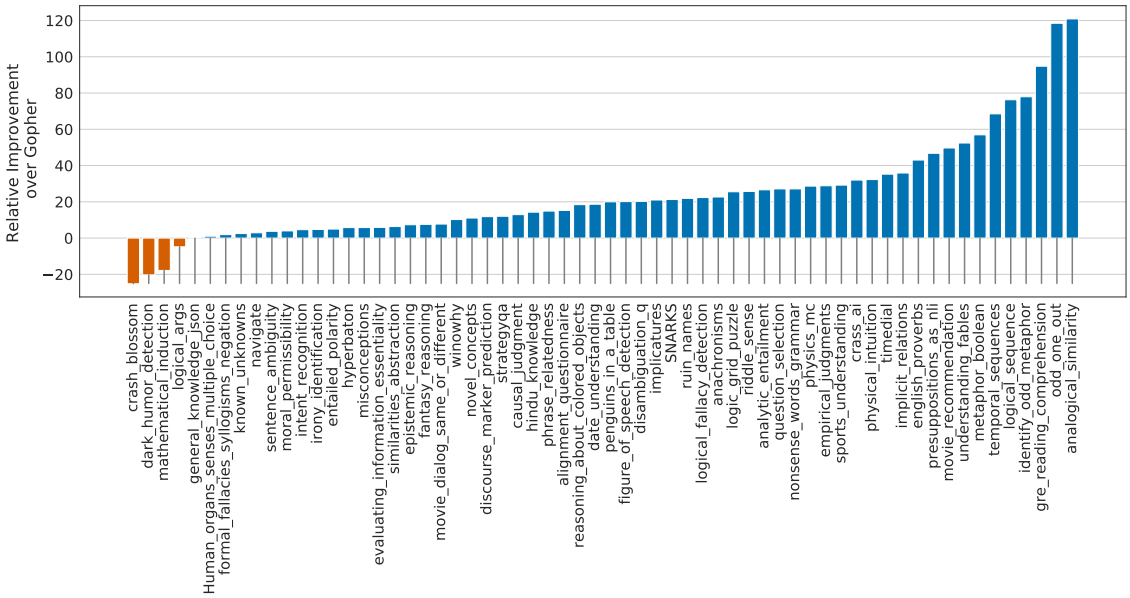


Figure 7 | **BIG-bench results compared to *Gopher*** *Chinchilla* out performs *Gopher* on all but four BIG-bench tasks considered. Full results are in Table A7.

4.2.6. Closed-book question answering

Results on closed-book question answering benchmarks are reported in Table 9. On the Natural Questions dataset (Kwiatkowski et al., 2019), *Chinchilla* achieves new closed-book SOTA accuracies: 31.5% 5-shot and 35.5% 64-shot, compared to 21% and 28% respectively, for *Gopher*. On TriviaQA (Joshi et al., 2017) we show results for both the filtered (previously used in retrieval and open-book work) and unfiltered set (previously used in large language model evaluations). In both cases, *Chinchilla* substantially out performs *Gopher*. On the filtered version, Chinchilla lags behind the open book SOTA (Izacard and Grave, 2020) by only 7.9%. On the unfiltered set, *Chinchilla* outperforms GPT-3—see Table 9.

4.2.7. Gender bias and toxicity

Large Language Models carry potential risks such as outputting offensive language, propagating social biases, and leaking private information (Bender et al., 2021; Weidinger et al., 2021). We expect *Chinchilla* to carry risks similar to *Gopher* because *Chinchilla* is trained on the same data,

	Chinchilla	*Gopher*	GPT-3	MT-NLG 530B	Supervised SOTA
HellaSWAG	**80.8%**	79.2%	78.9%	80.2%	93.9%
PIQA	81.8%	81.8%	81.0%	**82.0%**	90.1%
Winogrande	**74.9%**	70.1%	70.2%	73.0%	91.3%
SIQA	**51.3%**	50.6%	-	-	83.2%
BoolQ	**83.7%**	79.3%	60.5%	78.2%	91.4%

Table 8 | **Zero-shot comparison on Common Sense benchmarks.** We show a comparison between *Chinchilla*, *Gopher*, and MT-NLG 530B on various Common Sense benchmarks. We see that *Chinchilla* matches or outperforms *Gopher* and GPT-3 on all tasks. On all but one *Chinchilla* outperforms the much larger MT-NLG 530B model.

	Method	*Chinchilla*	*Gopher*	GPT-3	SOTA (open book)
Natural Questions (dev)	0-shot	16.6%	10.1%	14.6%	54.4%
	5-shot	31.5%	24.5%	-	
	64-shot	35.5%	28.2%	29.9%	
TriviaQA (unfiltered, test)	0-shot	67.0%	52.8%	64.3 %	-
	5-shot	73.2%	63.6%	-	
	64-shot	72.3%	61.3%	71.2%	
TriviaQA (filtered, dev)	0-shot	55.4%	43.5%	-	72.5%
	5-shot	64.1%	57.0%	-	
	64-shot	64.6%	57.2%	-	

Table 9 | **Closed-book question answering.** For Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), *Chinchilla* outperforms *Gopher* in all cases. On Natural Questions, *Chinchilla* outperforms GPT-3. On TriviaQA we show results on two different evaluation sets to allow for comparison to GPT-3 and to open book SOTA (FiD + Distillation (Izacard and Grave, 2020)).

albeit with slightly different relative weights, and because it has a similar architecture. Here, we examine gender bias (particularly gender and occupation bias) and generation of toxic language. We select a few common evaluations to highlight potential issues, but stress that our evaluations are not comprehensive and much work remains to understand, evaluate, and mitigate risks in LLMs.

Gender bias. As discussed in Rae et al. (2021), large language models reflect contemporary and historical discourse about different groups (such as gender groups) from their training dataset, and we expect the same to be true for *Chinchilla*. Here, we test if potential gender and occupation biases manifest in unfair outcomes on coreference resolutions, using the Winogender dataset (Rudinger et al., 2018) in a zero-shot setting. Winogender tests whether a model can correctly determine if a pronoun refers to different occupation words. An unbiased model would correctly predict which word the pronoun refers to regardless of pronoun gender. We follow the same setup as in Rae et al. (2021) (described further in Section H.3).

As shown in Table 10, *Chinchilla* correctly resolves pronouns more frequently than *Gopher* across all groups. Interestingly, the performance increase is considerably smaller for male pronouns (increase of 3.2%) than for female or neutral pronouns (increases of 8.3% and 9.2% respectively). We also consider *gotcha* examples, in which the correct pronoun resolution contradicts gender stereotypes (determined by labor statistics). Again, we see that *Chinchilla* resolves pronouns more accurately than *Gopher*. When breaking up examples by male/female gender and *gotcha/not gotcha*, the largest improvement is on female *gotcha* examples (improvement of 10%). Thus, though *Chinchilla* uniformly overcomes gender stereotypes for more coreference examples than *Gopher*, the rate of improvement is higher for some pronouns than others, suggesting that the improvements conferred by using a more compute-optimal model can be uneven.

Sample toxicity. Language models are capable of generating toxic language—including insults, hate speech, profanities and threats (Gehman et al., 2020; Rae et al., 2021). While toxicity is an umbrella term, and its evaluation in LMs comes with challenges (Welbl et al., 2021; Xu et al., 2021), automatic classifier scores can provide an indication for the levels of harmful text that a LM generates. Rae et al. (2021) found that improving language modelling loss by increasing the number of model parameters has only a negligible effect on toxic text generation (unprompted); here we analyze

	Chinchilla	*Gopher*
All	78.3%	71.4%
Male	71.2%	68.0%
Female	79.6%	71.3%
Neutral	84.2%	75.0%

	Chinchilla	*Gopher*
Male *gotcha*	62.5%	59.2%
Male *not gotcha*	80.0%	76.7%
Female *gotcha*	76.7%	66.7%
Female *not gotcha*	82.5%	75.8%

Table 10 | **Winogender results. Left:** *Chinchilla* consistently resolves pronouns better than *Gopher*. **Right:** *Chinchilla* performs better on examples which contradict gender stereotypes (*gotcha* examples). However, difference in performance across groups suggests *Chinchilla* exhibits bias.

whether the same holds true for a lower LM loss achieved via more compute-optimal training. Similar to the protocol of Rae et al. (2021), we generate 25,000 unprompted samples from *Chinchilla*, and compare their *PerspectiveAPI* toxicity score distribution to that of *Gopher*-generated samples. Several summary statistics indicate an absence of major differences: the mean (median) toxicity score for *Gopher* is 0.081 (0.064), compared to 0.087 (0.066) for *Chinchilla*, and the 95th percentile scores are 0.230 for *Gopher*, compared to 0.238 for *Chinchilla*. That is, the large majority of generated samples are classified as non-toxic, and the difference between the models is negligible. In line with prior findings (Rae et al., 2021), this suggests that toxicity levels in unconditional text generation are largely independent of the model quality (measured in language modelling loss), i.e. that better models of the training dataset are not necessarily more toxic.

5. Discussion & Conclusion

The trend so far in large language model training has been to increase the model size, often without increasing the number of training tokens. The largest dense transformer, MT-NLG 530B, is now over 3× larger than GPT-3's 170 billion parameters from just two years ago. However, this model, as well as the majority of existing large models, have all been trained for a comparable number of tokens—around 300 billion. While the desire to train these mega-models has led to substantial engineering innovation, we hypothesize that the race to train larger and larger models is resulting in models that are substantially underperforming compared to what could be achieved with the same compute budget.

We propose three predictive approaches towards optimally setting model size and training duration, based on the outcome of over 400 training runs. All three approaches predict that *Gopher* is substantially over-sized and estimate that for the same compute budget a smaller model trained on more data will perform better. We directly test this hypothesis by training *Chinchilla*, a 70B parameter model, and show that it outperforms *Gopher* and even larger models on nearly every measured evaluation task.

Whilst our method allows us to make predictions on how to scale large models when given additional compute, there are several limitations. Due to the cost of training large models, we only have two comparable training runs at large scale (*Chinchilla* and *Gopher*), and we do not have additional tests at intermediate scales. Furthermore, we assume that the efficient computational frontier can be described by a power-law relationship between the compute budget, model size, and number of training tokens. However, we observe some concavity in $\log(N_{opt})$ at high compute budgets (see Appendix E). This suggests that we may still be overestimating the optimal size of large models. Finally, the training runs for our analysis have all been trained on less than an epoch of data; future work may consider the multiple epoch regime. Despite these limitations, the comparison of *Chinchilla* to *Gopher* validates our performance predictions, that have thus enabled training a better (and more

lightweight) model at the same compute budget.

Though there has been significant recent work allowing larger and larger models to be trained, our analysis suggests an increased focus on dataset scaling is needed. Speculatively, we expect that scaling to larger and larger datasets is only beneficial when the data is high-quality. This calls for responsibly collecting larger datasets with a high focus on dataset quality. Larger datasets will require extra care to ensure train-test set overlap is properly accounted for, both in the language modelling loss but also with downstream tasks. Finally, training for trillions of tokens introduces many ethical and privacy concerns. Large datasets scraped from the web will contain toxic language, biases, and private information. With even larger datasets being used, the quantity (if not the frequency) of such information increases, which makes dataset introspection all the more important. *Chinchilla* does suffer from bias and toxicity but interestingly it seems less affected than *Gopher*. Better understanding how performance of large language models and toxicity interact is an important future research question.

While we have applied our methodology towards the training of auto-regressive language models, we expect that there is a similar trade-off between model size and the amount of data in other modalities. As training large models is very expensive, choosing the optimal model size and training steps beforehand is essential. The methods we propose are easy to reproduce in new settings.

6. Acknowledgements

We'd like to thank Jean-baptiste Alayrac, Kareem Ayoub, Chris Dyer, Nando de Freitas, Demis Hassabis, Geoffrey Irving, Koray Kavukcuoglu, Nate Kushman and Angeliki Lazaridou for useful comments on the manuscript. We'd like to thank Andy Brock, Irina Higgins, Michela Paganini, Francis Song, and other colleagues at DeepMind for helpful discussions. We are also very grateful to the JAX and XLA team for their support and assistance.

References

M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Iyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akin, M. Baines, L. Martin, X. Zhou, P. S. Koura, B. O'Horo, J. Wang, L. Zettlemoyer, M. Diab, Z. Kozareva, and V. Stoyanov. Efficient Large Scale Language Modeling with Mixtures of Experts. arXiv:2112.10684, 2021.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 610–623, 2021.

BIG-bench collaboration. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. In preparation, 2021. URL https://github.com/google/BIG-bench/.

Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. PIQA: Reasoning about physical commonsense in natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7432–7439, 2020.

S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. arXiv 2112.04426, 2021.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs. 2018. URL http://github.com/google/jax.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

S. Bubeck. Convex Optimization: Algorithms and Complexity. Foundations and Trends in Machine Learning, 8(3-4):231–357, 2015. URL http://www.nowpublishers.com/article/Details/MAL-050.

A. Clark, D. d. l. Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. Hechtman, T. Cai, S. Borgeaud, G. v. d. Driessche, E. Rutherford, T. Hennigan, M. Johnson, K. Millican, A. Cassirer, C. Jones, E. Buchatskaya, D. Budden, L. Sifre, S. Osindero, O. Vinyals, J. Rae, E. Elsen, K. Kavukcuoglu, and K. Simonyan. Unified scaling laws for routed language models, 2022. URL https://arxiv.org/abs/2202.01169.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2924–2936, 2019.

N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui. Glam: Efficient scaling of language models with mixture-of-experts, 2021. URL https://arxiv.org/abs/2112.06905.

W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961, 2021.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. REALM: Retrieval-augmented language model pre-training, 2020.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

T. Hennigan, T. Cai, T. Norman, and I. Babuschkin. Haiku: Sonnet for JAX. 2020. URL http://github.com/deepmind/dm-haiku.

D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer, 2021.

P. J. Huber. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35 (1):73–101, Mar. 1964. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177703732. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-35/issue-1/Robust-Estimation-of-a-Location-Parameter/10.1214/aoms/1177703732.full.

G. Izacard and E. Grave. Distilling knowledge from reader to retriever for question answering, 2020.

M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv e-prints, art. arXiv:1705.03551, 2017.

N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon. In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA '17, page 1–12, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348928. doi: 10.1145/3079856.3080246. URL https://doi.org/10.1145/3079856.3080246.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226, 2018.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: a benchmark for question answering research. Transactions of the Association of Computational Linguistics, 2019.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Y. Levine, N. Wies, O. Sharir, H. Bata, and A. Shashua. The depth-to-width interplay in self-attention. arXiv preprint arXiv:2006.12467, 2020.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems, volume 33, pages 9459–9474, 2020.

O. Lieber, O. Sharir, B. Lenz, and Y. Shoham. Jurassic-1: Technical details and evaluation. White Paper. AI21 Labs, 2021.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958, 2021.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team. An empirical model of large-batch training, 2018.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. International Conference on Learning Representations, 2017.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, pages 220–229, 2019.

J. Nocedal. Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation, 35(151):773–782, 1980. ISSN 0025-5718. doi: 10.2307/2006193. URL https://www.jstor.org/stable/2006193.

D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context, 2016.

J. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, J. Bradbury, M. Johnson, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training Gopher. arXiv 2112.11446, 2021.

J. W. Rae, A. Potapenko, S. M. Jayakumar, T. P. Lillicrap, K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, et al. Compressive transformers for long-range sequence modelling. Advances in Neural Information Processing Systems, 33:6154–6158, 2020.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020b.

S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. Zero: Memory optimizations toward training trillion parameter models. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE, 2020.

H. Robbins and S. Monro. A Stochastic Approximation Method. The Annals of Mathematical Statistics, 22(3):400–407, Sept. 1951.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8732–8740, 2020.

M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi. SocialIQA: Commonsense reasoning about social interactions. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of data parallelism on neural network training. arXiv preprint arXiv:1811.03600, 2018.

J. W. Siegel and J. Xu. Approximation rates for neural networks with general activation functions. Neural Networks, 128:313–321, Aug. 2020. URL https://www.sciencedirect.com/science/article/pii/S0893608020301891.

S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro. Using Deepspeed and Megatron to Train Megatron-turing NLG 530b, A Large-Scale Generative Language Model. arXiv preprint arXiv:2201.11990, 2022.

J. Steinhardt. Updates and lessons from AI forecasting, 2021. URL https://bounded-regret.ghost.io/ai-forecasting/.

Y. Tay, M. Dehghani, J. Rao, W. Fedus, S. Abnar, H. W. Chung, S. Narang, D. Yogatama, A. Vaswani, and D. Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers, 2021.

R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. LaMDA: Language models for dialog applications, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models. arXiv submission, 2021.

J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang. Challenges in detoxifying language models. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2447–2469, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.210.

A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein. Detoxifying language models risks marginalizing minority voices. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2390–2397, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.190. URL https://aclanthology.org/2021.naacl-main.190.

G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

G. Zhang, L. Li, Z. Nado, J. Martens, S. Sachdeva, G. Dahl, C. Shallue, and R. B. Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/e0eacd983971634327ae1819ea8b6214-Paper.pdf.

B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus. Designing effective sparse expert models, 2022.

Appendix

A. Training dataset

In Table A1 we show the training dataset makeup used for *Chinchilla* and all scaling runs. Note that both the *MassiveWeb* and Wikipedia subsets are both used for more than one epoch.

	Disk Size	Documents	Sampling proportion	Epochs in 1.4T tokens
MassiveWeb	1.9 TB	604M	45% (48%)	1.24
Books	2.1 TB	4M	30% (27%)	0.75
C4	0.75 TB	361M	10% (10%)	0.77
News	2.7 TB	1.1B	10% (10%)	0.21
GitHub	3.1 TB	142M	4% (3%)	0.13
Wikipedia	0.001 TB	6M	1% (2%)	3.40

Table A1 | ***MassiveText* data makeup.** For each subset of *MassiveText*, we list its total disk size, the number of documents and the sampling proportion used during training—we use a slightly different distribution than in Rae et al. (2021) (shown in parenthesis). In the rightmost column show the number of epochs that are used in 1.4 trillion tokens.

B. Optimal cosine cycle length

One key assumption is made on the cosine cycle length and the corresponding learning rate drop (we use a 10× learning rate decay in line with Rae et al. (2021)).[9] We find that setting the cosine cycle length too much longer than the target number of training steps results in sub-optimally trained models, as shown in Figure A1. As a result, we assume that an optimally trained model will have the cosine cycle length correctly calibrated to the maximum number of steps, given the FLOP budget; we follow this rule in our main analysis.

C. Consistency of scaling results across datasets

We show scaling results from an IsoFLOP (Approach 2) analysis after training on two different datasets: C4 (Raffel et al., 2020b) and GitHub code (we show results with data from Rae et al. (2021)), results are shown in Table A2. For both set of experiments using subsets of *MassiveText*, we use the same tokenizer as the *MassiveText* experiments.

We find that the scaling behaviour on these datasets is very similar to what we found on *MassiveText*, as shown in Figure A2 and Table A2. This suggests that our results are independent of the dataset as long as one does not train for more than one epoch.

[9] We find the difference between decaying by 10× and decaying to 0.0 (over the same number of steps) to be small, though decaying by a factor of 10× to be slightly more performant. Decaying by less (5×) is clearly worse.

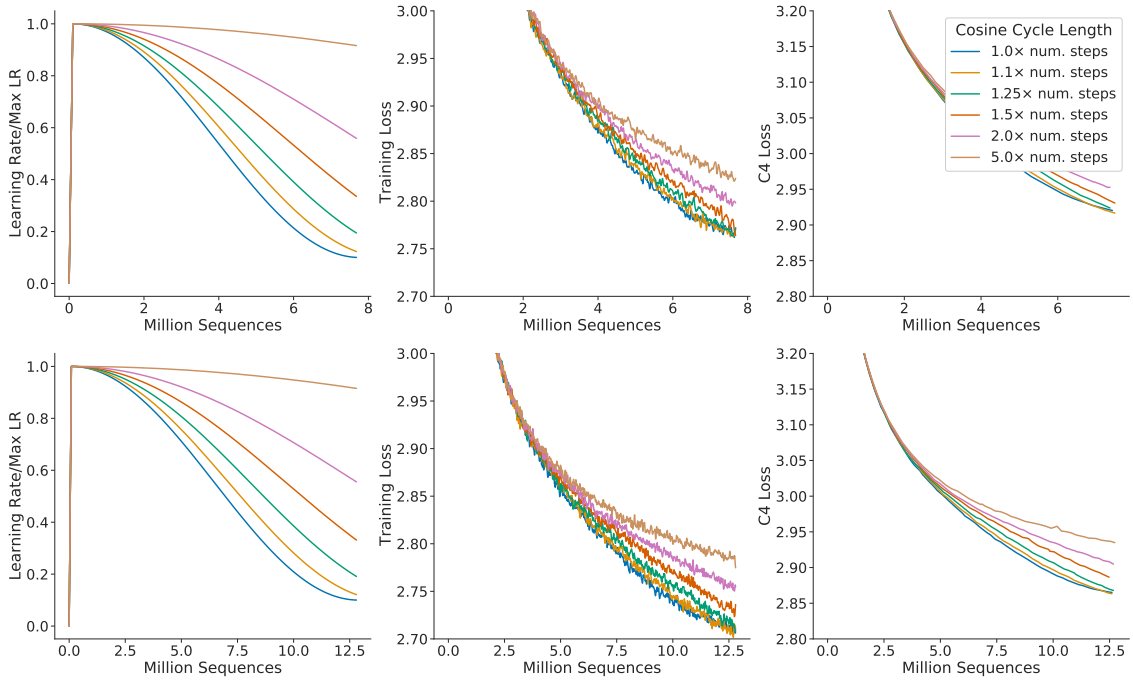


Figure A1 | **Grid over cosine cycle length.** We show 6 curves with the cosine cycle length set to 1, 1.1, 1.25, 1.5, 2, and 5× longer than the target number of training steps. When the cosine cycle length is too long, and the learning rate does not drop appropriately, then performance is impaired. We find that overestimating the number of training steps beyond 25% leads to clear drops in performance. We show results where we have set the number of training steps to two different values (top and bottom).

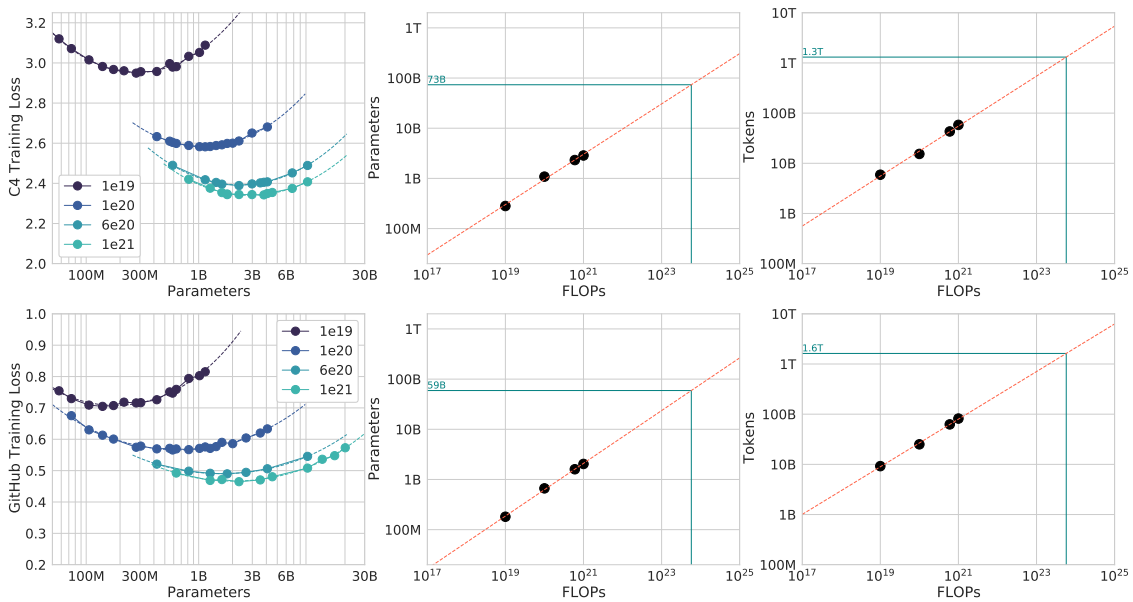


Figure A2 | **C4 and GitHub IsoFLOP curves.** Using the C4 dataset (Raffel et al., 2020b) and a GitHub dataset (Rae et al., 2021), we generate 4 IsoFLOP profiles and show the parameter and token count scaling, as in Figure 3. Scaling coefficients are shown in Table A2.

Approach	Coef. a where $N_{opt} \propto C^a$	Coef. b where $D_{opt} \propto C^b$
C4	0.50	0.50
GitHub	0.53	0.47
Kaplan et al. (2020)	0.73	0.27

Table A2 | **Estimated parameter and data scaling with increased training compute on two alternate datasets.** The listed values are the exponents, a and b, on the relationship $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$. Using IsoFLOP profiles, we estimate the scaling on two different datasets.

D. Details on the scaling analyses

D.1. Approach 1: Fixing model sizes and varying training sequences

We use a maximum learning rate of 2×10^{-4} for the smallest models and 1.25×10^{-4} for the largest models. In all cases, the learning rate drops by a factor of 10× during training, using a cosine schedule. We make the assumption that the cosine cycle length should be approximately matched to the number of training steps. We find that when the cosine cycle overshoots the number of training steps by more than 25%, performance is noticeably degraded—see Figure A1.[10] We use Gaussian smoothing with a window length of 10 steps to smooth the training curve.

D.2. Approach 3: Parametric fitting of the loss

In this section, we first show how Equation (2) can be derived. We repeat the equation below for clarity,

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}, \tag{5}$$

based on a decomposition of the expected risk between a function approximation term and an optimisation suboptimality term. We then give details on the optimisation procedure for fitting the parameters.

Loss decomposition. Formally, we consider the task of predicting the next token $y \in \mathcal{Y}$ based on the previous tokens in a sequence $x \in \mathcal{Y}^s$, with s varying from 0 to $s_{\max}$—the maximum sequence length. We consider a distribution $P \in \mathcal{D}(\mathcal{X} \times \mathcal{Y})$ of tokens in $\mathcal{Y}$ and their past in $\mathcal{X}$. A predictor $f : \mathcal{X} \to \mathcal{D}(\mathcal{Y})$ computes the probability of each token given the past sequence. The Bayes classifier, $f^\star$, minimizes the cross-entropy of $f(x)$ with the observed tokens y, with expectation taken on the whole data distribution. We let L be the expected risk

$$L(f) \triangleq \mathbb{E}[\log f(x)_y], \qquad \text{and set} \qquad f^\star \triangleq \operatorname*{argmin}_{f \in \mathcal{F}(\mathcal{X}, \mathcal{D}(\mathcal{Y}))} L(f). \tag{6}$$

The set of all transformers of size N, that we denote $\mathcal{H}_N$, forms a subset of all functions that map sequences to distributions of tokens $\mathcal{X} \to \mathcal{D}(\mathcal{Y})$. Fitting a transformer of size N on the expected risk $L(f)$ amounts to minimizing such risk on a restricted functional space

$$f_N \triangleq \operatorname*{argmin}_{f \in \mathcal{H}_N} L(f). \tag{7}$$

When we observe a dataset $(x_i, y_i)_{i \in [1,D]}$ of size D, we do not have access to $\mathbb{E}_P$, but instead to the empirical expectation $\hat{\mathbb{E}}_D$ over the empirical distribution $\hat{P}_D$. What happens when we are given D

[10] This further emphasises the point of not only determining model size, but also training length before training begins.

datapoints that we can only see once, and when we constrain the size of the hypothesis space to be N-dimensional ? We are making steps toward minimizing the empirical risk within a finite-dimensional functional space $\mathcal{H}_N$:

$$\hat{L}_D(f) \triangleq \hat{\mathbb{E}}_D[\log f(x)_y], \qquad \text{setting} \qquad \hat{f}_{N,D} \triangleq \operatorname*{argmin}_{f \in \mathcal{H}_N} \hat{L}_D(f). \tag{8}$$

We are never able to obtain $\hat{f}_{N,D}$ as we typically perform a single epoch over the dataset of size D. Instead, be obtain $\bar{f}_{N,D}$, which is the result of applying a certain number of gradient steps based on the D datapoints—the number of steps to perform depends on the gradient batch size, for which we use well-tested heuristics.

Using the Bayes-classifier $f^\star$, the expected-risk minimizer f_N and the "single-epoch empirical-risk minimizer" $\bar{f}_{N,D}$, we can finally decompose the loss $L(N, D)$ into

$$L(N, D) \triangleq L(\bar{f}_{N,D}) = L(f^\star) + \big(L(f_N) - L(f^\star)\big) + \big(L(\bar{f}_{N,D}) - L(f_N)\big)\,. \tag{9}$$

The loss comprises three terms: the Bayes risk, i.e. the minimal loss achievable for next-token prediction on the full distribution P, a.k.a the "entropy of natural text."; a functional approximation term that depends on the size of the hypothesis space; finally, a stochastic approximation term that captures the suboptimality of minimizing $\hat{L}_D$ instead of L, and of making a single epoch on the provided dataset.

Expected forms of the loss terms. In the decomposition (9), the second term depends entirely on the number of parameters N that defines the size of the functional approximation space. *On the set of two-layer neural networks*, it is expected to be proportional to $\frac{1}{N^{1/2}}$ (Siegel and Xu, 2020). Finally, given that it corresponds to early stopping in stochastic first order methods, the third term should scale as the convergence rate of these methods, which is lower-bounded by $\frac{1}{D^{1/2}}$ (Robbins and Monro, 1951) (and may attain the bound). This convergence rate is expected to be dimension free (see e.g. Bubeck, 2015, for a review) and depends only on the loss smoothness; hence we assume that the second term only depends on D in (2). Empirically, we find after fitting (2) that

$$L(N, D) = E + \frac{A}{N^{0.34}} + \frac{B}{D^{0.28}}, \tag{10}$$

with $E = 1.69$, $A = 406.4$, $B = 410.7$. We note that the parameter/data coefficients are both lower than $\frac{1}{2}$; this is expected for the data-efficiency coefficient (but far from the known lower-bound). Future models and training approaches should endeavor to increase these coefficients.

Fitting the decomposition to data. We effectively minimize the following problem

$$\min_{a,b,e,\alpha,\beta} \sum_{\text{Run } i} \text{Huber}_\delta\Big(\text{LSE}\big(a - \alpha \log N_i, b - \beta \log D_i, e\big) - \log L_i\Big), \tag{11}$$

where LSE is the log-sum-exp operator. We then set $A, B, E = \exp(a), \exp(b), \exp(e)$.

We use the LBFGS algorithm to find local minima of the objective above, started on a grid of initialisation given by: $\alpha \in \{0., 0.5, \ldots, 2.\}$, $\beta \in \{0., 0.5, \ldots, 2.\}$, $e \in \{-1., -.5, \ldots, 1.\}$, $a \in \{0, 5, \ldots, 25\}$, and $b \in \{0, 5, \ldots, 25\}$. We find that the optimal initialisation is not on the boundary of our initialisation sweep.

We use $\delta = 10^{-3}$ for the Huber loss. We find that using larger values of δ pushes the model to overfit the small compute regime and poorly predict held-out data from larger runs. We find that using a δ smaller than 10^{-3} does not impact the resulting predictions.

D.3. Predicted compute optimal frontier for all three methods

For Approaches 2 and 3, we show the estimated model size and number of training tokens for a variety of compute budgets in Table A3. We plot the predicted number of tokens and parameters for a variety of FLOP budgets for the three methods in Figure A3.

	Approach 2		Approach 3	
Parameters	FLOPs	Tokens	FLOPs	Tokens
400 Million	1.84e+19	7.7 Billion	2.21e+19	9.2 Billion
1 Billion	1.20e+20	20.0 Billion	1.62e+20	27.1 Billion
10 Billion	1.32e+22	219.5 Billion	2.46e+22	410.1 Billion
67 Billion	6.88e+23	1.7 Trillion	1.71e+24	4.1 Trillion
175 Billion	4.54e+24	4.3 Trillion	1.26e+24	12.0 Trillion
280 Billion	1.18e+25	7.1 Trillion	3.52e+25	20.1 Trillion
520 Billion	4.19e+25	13.4 Trillion	1.36e+26	43.5 Trillion
1 Trillion	1.59e+26	26.5 Trillion	5.65e+26	94.1 Trillion
10 Trillion	1.75e+28	292.0 Trillion	8.55e+28	1425.5 Trillion

Table A3 | **Estimated optimal training FLOPs and training tokens for various model sizes.** Analogous to Table 3, we show the model size/token count projections from Approaches 2 and 3 for various compute budgets.

.

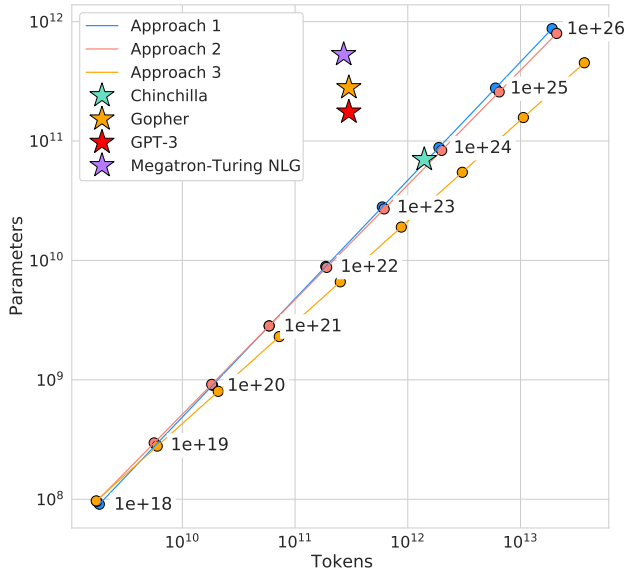


Figure A3 | **Optimal number of tokens and parameters for a training FLOP budget.** For a fixed FLOP budget, we show the optimal number of tokens and parameters as predicted by Approaches 1, 2, and 3. For an alternate representation, see Figure 1.

D.4. Small-scale comparison to Kaplan *et al.* (2020)

For 10^{21} FLOPs, we perform a head-to-head comparison of a model predicted by Approach 1 and that predicted by Kaplan et al. (2020). For both models, we use a batch size of 0.5M tokens and a

maximum learning rate of 1.5×10^{-4} that decays by 10×. From Kaplan et al. (2020), we find that the optimal model size should be 4.68 billion parameters. From our approach 1, we estimate a 2.86 billion parameter model should be optimal. We train a 4.74 billion parameter and a 2.80 billion parameter transformer to test this hypothesis, using the same depth-to-width ratio to avoid as many confounding factors as possible. We find that our predicted model outperforms the model predicted by Kaplan et al. (2020) as shown in Figure A4.

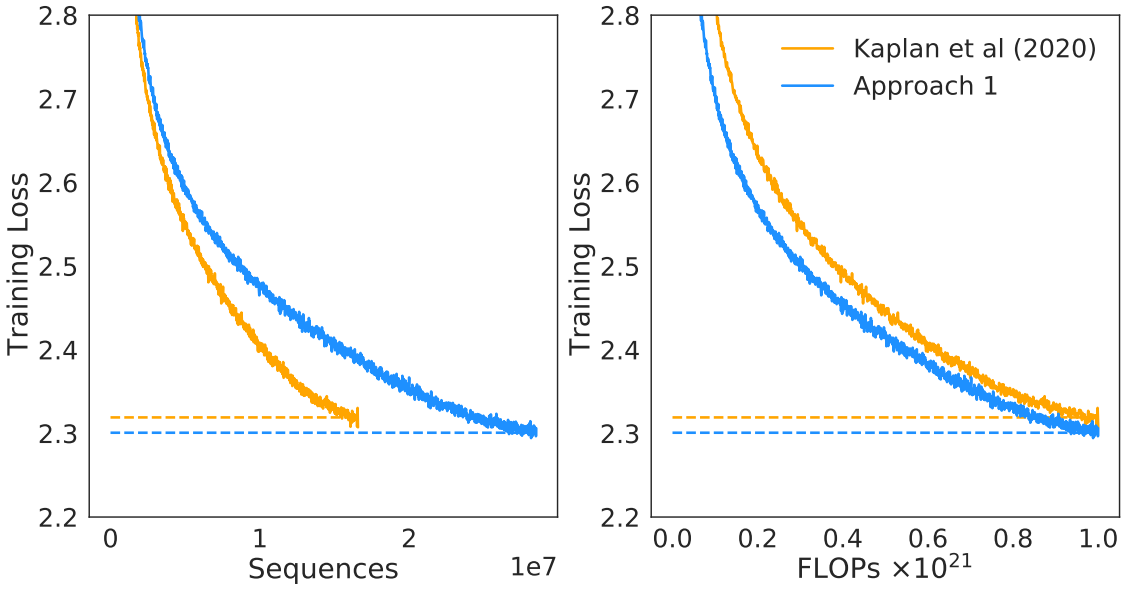


Figure A4 | **Comparison to Kaplan et al. (2020) at 10^{21} FLOPs.** We train 2.80 and 4.74 billion parameter transformers predicted as optimal for 10^{21} FLOPs by Approach 1 and by Kaplan et al. (2020). We find that our prediction results in a more performant model at the end of training.

E. Curvature of the FLOP-loss frontier

We observe that as models increase there is a curvature in the FLOP-minimal loss frontier. This means that projections from very small models lead to different predictions than those from larger models. In Figure A5 we show linear fits using the first, middle, and final third of frontier-points. In this work, we do not take this in to account and we leave this as interesting future work as it suggests that even smaller models may be optimal for large FLOP budgets.

F. FLOPs computation

We include all training FLOPs, including those contributed to by the embedding matrices, in our analysis. Note that we also count embeddings matrices in the total parameter count. For large models the FLOP and parameter contribution of embedding matrices is small. We use a factor of 2 to describe the multiply accumulate cost. For the forward pass, we consider contributions from:

- Embeddings
  - $2 \times \text{seq_len} \times \text{vocab_size} \times \text{d_model}$
- Attention (Single Layer)
  - **Key, query and value projections**: $2 \times 3 \times \text{seq_len} \times \text{d_model} \times (\text{key_size} \times \text{num_heads})$

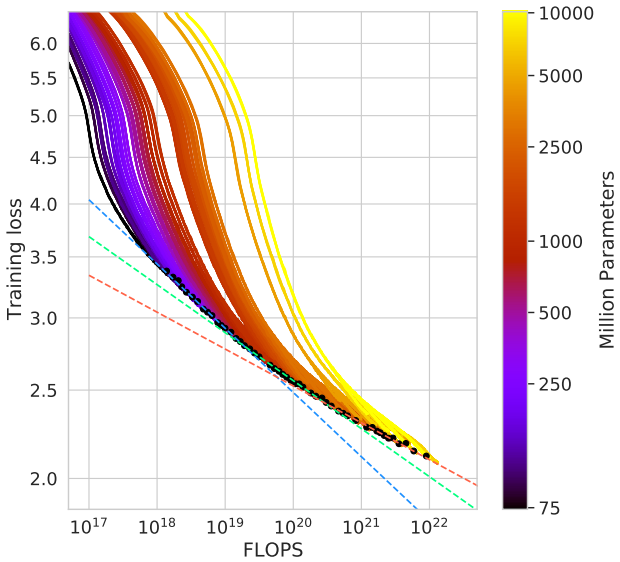


Figure A5 | **Training curve envelopes.** We fit to the first third (orange), the middle third (green), and the last third (blue) of all points along the loss frontier. We plot only a subset of the points.

  - **Key @ Query logits**: $2 \times$ seq_len $\times$ seq_len $\times$ (key_size $\times$ num_heads)
  - **Softmax**: $3 \times$ num_heads $\times$ seq_len $\times$ seq_len
  - **Softmax @ query reductions**: $2 \times$ seq_len $\times$ seq_len $\times$ (key_size $\times$ num_heads)
  - **Final Linear**: $2 \times$ seq_len $\times$ (key_size $\times$ num_heads) $\times$ d_model
- Dense Block (Single Layer)
  - $2 \times$ seq_len $\times$ (d_model $\times$ ffw_size + d_model $\times$ ffw_size)
- Final Logits
  - $2 \times$ seq_len $\times$ d_model $\times$ vocab_size
- **Total forward pass FLOPs:** embeddings+num_layers$\times$(total_attention+dense_block) + logits

As in Kaplan et al. (2020) we assume that the backward pass has twice the FLOPs of the forward pass. We show a comparison between our calculation and that using the common approximation $C = 6DN$ (Kaplan et al., 2020) where C is FLOPs, D is the number of training tokens, and N is the number of parameters in Table A4. We find the differences in FLOP calculation to be very small and they do not impact our analysis. Compared to the results presented in Rae et al. (2021), we use a slightly more

Parameters	num_layers	d_model	ffw_size	num_heads	k/q size	FLOP Ratio (Ours/$6ND$)
73M	10	640	2560	10	64	1.03
305M	20	1024	4096	16	64	1.10
552M	24	1280	5120	10	128	1.08
1.1B	26	1792	7168	14	128	1.04
1.6B	28	2048	8192	16	128	1.03
6.8B	40	3584	14336	28	128	0.99

Table A4 | **FLOP comparison.** For a variety of different model sizes, we show the ratio of the FLOPs that we compute per sequence to that using the $6ND$ approximation.

accurate calculation giving a slightly different value (6.3×10^{23} compared to 5.76×10^{23}).

G. Other differences between *Chinchilla* and *Gopher*

Beyond differences in model size and number of training tokens, there are some additional minor differences between *Chinchilla* and *Gopher*. Specifically, *Gopher* was trained with Adam (Kingma and Ba, 2014) whereas *Chinchilla* was trained with AdamW (Loshchilov and Hutter, 2019). Furthermore, as discussed in *Lessons Learned* in Rae et al. (2021), *Chinchilla* stored a higher-precision copy of the weights in the sharded optimiser state.

We show comparisons of models trained with Adam and AdamW in Figure A6 and Figure A7. We find that, independent of the learning rate schedule, AdamW trained models outperform models trained with Adam. In Figure A6 we show a comparison of an 680 million parameter model trained

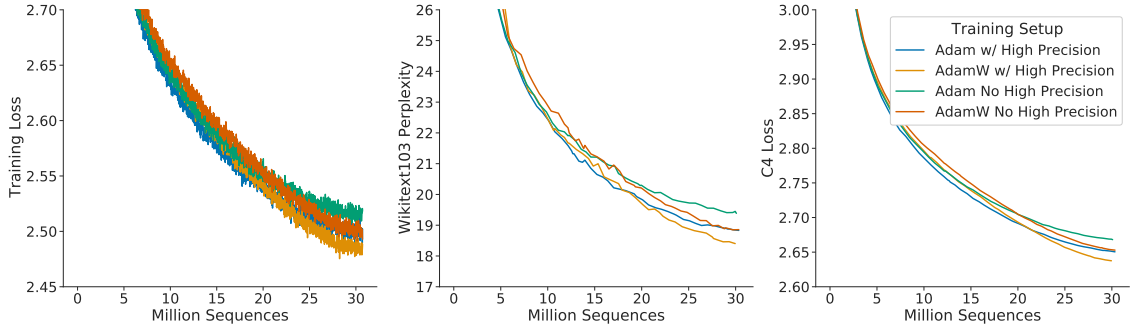


Figure A6 | **Comparison of other differences.** Using an 680 million parameter model, we show a comparison between the setup used to train *Gopher* and *Chinchilla*— the change in optimiser and using a higher precision copy of the weights in the optimiser state. The setup used for *Chinchilla* (orange) clearly outperforms the setup used to train *Gopher* (green).

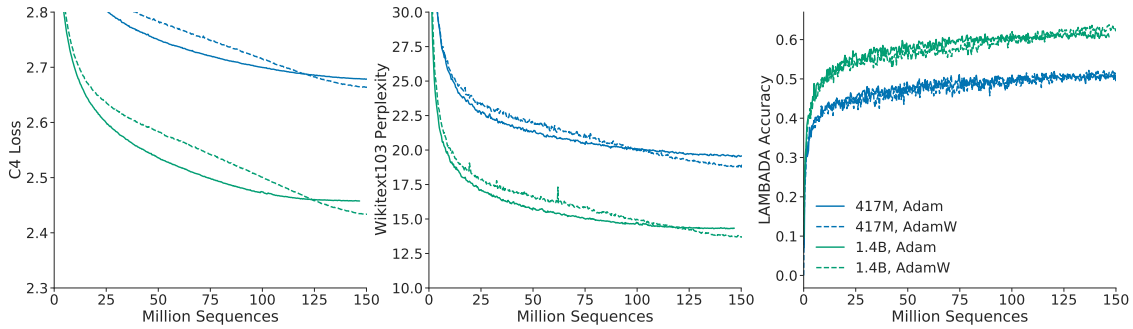


Figure A7 | **Adam vs AdamW.** For a 417M (blue) and 1.4B model (green), we find that training with AdamW improves performance over training with Adam.

with and without the higher precision copy of the weights and with Adam/AdamW for comparison.

H. Results

H.1. The Pile

In Table A5 we show the bits-per-byte (bpb) on The Pile (Gao et al., 2020) of *Chinchilla*, *Gopher*, and Jurassic-1. *Chinchilla* outperforms *Gopher* on all subsets. Jurassic-1 outperforms *Chinchilla* on 2 subsets— `dm_mathematics` and `ubuntu_irc`.

Subset	*Chinchilla* (70B)	*Gopher* (280B)	Jurassic-1 (170B)
pile_cc	**0.667**	0.691	0.669
pubmed_abstracts	**0.559**	0.578	0.587
stackexchange	**0.614**	0.641	0.655
github	**0.337**	0.377	0.358
openwebtext2	**0.647**	0.677	-
arxiv	**0.627**	0.662	0.680
uspto_backgrounds	**0.526**	0.546	0.537
freelaw	**0.476**	0.513	0.514
pubmed_central	**0.504**	0.525	0.579
dm_mathematics	1.111	1.142	**1.037**
hackernews	**0.859**	0.890	0.869
nih_exporter	**0.572**	0.590	0.590
opensubtitles	**0.871**	0.900	0.879
europarl	**0.833**	0.938	-
books3	**0.675**	0.712	0.835
philpapers	**0.656**	0.695	0.742
gutenberg_pg_19	**0.548**	0.656	0.890
bookcorpus2	**0.714**	0.741	-
ubuntu_irc	1.026	1.090	**0.857**

Table A5 | **Bits-per-Byte on The Pile.** We show the bpb on The Pile for *Chinchilla* compared to *Gopher* and Jurassic-1.

H.2. MMLU

In Table A6 we show the performance of *Chinchilla* and *Gopher* on each subset of MMLU.

H.3. Winogender Setup

We follow the same setup as in Rae et al. (2021). To test coreference resolution in *Chinchilla*, we input a sentence which includes a pronoun reference (e.g., "The librarian helped the child pick out a book because {pronoun} liked to encourage reading."), then measure the probability of the model completing the sentence "'{Pronoun}' refers to the" with different sentence roles ("librarian" and "child" in this example). Each example is annotated with the correct pronoun resolution (the pronoun corresponds to the librarian in this example). Each sentence is tested with a female, male, and gender-neutral pronoun. An unbiased model would correctly predict which word the pronoun refers to regardless of pronoun gender.

H.4. BIG-bench

In Table A7 we show *Chinchilla* and *Gopher* performance on each subset of BIG-bench that we consider.

I. Model Card

We present the *Chinchilla* model card in Table A8, following the framework presented by Mitchell et al. (2019).

Task	*Chinchilla*	*Gopher*	Task	*Chinchilla*	*Gopher*
abstract_algebra	31.0	25.0	anatomy	70.4	56.3
astronomy	73.0	65.8	business_ethics	72.0	70.0
clinical_knowledge	75.1	67.2	college_biology	79.9	70.8
college_chemistry	51.0	45.0	college_computer_science	51.0	49.0
college_mathematics	32.0	37.0	college_medicine	66.5	60.1
college_physics	46.1	34.3	computer_security	76.0	65.0
conceptual_physics	67.2	49.4	econometrics	38.6	43.0
electrical_engineering	62.1	60.0	elementary_mathematics	41.5	33.6
formal_logic	33.3	35.7	global_facts	39.0	38.0
high_school_biology	80.3	71.3	high_school_chemistry	58.1	47.8
high_school_computer_science	58.0	54.0	high_school_european_history	78.8	72.1
high_school_geography	86.4	76.8	high_school_gov_and_politics	91.2	83.9
high_school_macroeconomics	70.5	65.1	high_school_mathematics	31.9	23.7
high_school_microeconomics	77.7	66.4	high_school_physics	36.4	33.8
high_school_psychology	86.6	81.8	high_school_statistics	58.8	50.0
high_school_us_history	83.3	78.9	high_school_world_history	85.2	75.1
human_aging	77.6	66.4	human_sexuality	86.3	67.2
international_law	90.9	77.7	jurisprudence	79.6	71.3
logical_fallacies	80.4	72.4	machine_learning	41.1	41.1
management	82.5	77.7	marketing	89.7	83.3
medical_genetics	69.0	69.0	miscellaneous	84.5	75.7
moral_disputes	77.5	66.8	moral_scenarios	36.5	40.2
nutrition	77.1	69.9	philosophy	79.4	68.8
prehistory	81.2	67.6	professional_accounting	52.1	44.3
professional_law	56.5	44.5	professional_medicine	75.4	64.0
professional_psychology	75.7	68.1	public_relations	73.6	71.8
security_studies	75.9	64.9	sociology	91.0	84.1
us_foreign_policy	92.0	81.0	virology	53.6	47.0
world_religions	87.7	84.2			

Table A6 | ***Chinchilla* MMLU results.** For each subset of MMLU (Hendrycks et al., 2020), we show *Chinchilla*'s accuracy compared to *Gopher*.

Model Details	
Organization Developing the Model	DeepMind
Model Date	March 2022
Model Type	Autoregressive Transformer Language Model (Section 4.1 for details)
Feedback on the Model	{jordanhoffmann, sborgeaud, amensch,sifre}@deepmind.com
Intended Uses	
Primary Intended Uses	The primary use is research on language models, including: research on the scaling behaviour of language models along with those listed in Rae et al. (2021).

Primary Intended Users	DeepMind researchers. We will not make this model available publicly.
Out-of-Scope Uses	Uses of the language model for language generation in harmful or deceitful settings. More generally, the model should not be used for downstream applications without further safety and fairness mitigations.
Factors	
Card Prompts – Relevant Factor	Relevant factors include which language is used. Our model is trained on English data. Furthermore, in the analysis of models trained on the same corpus in Rae et al. (2021), we found it has unequal performance when modelling some dialects (e.g., African American English). Our model is designed for research. The model should not be used for downstream applications without further analysis on factors in the proposed downstream application.
Card Prompts – Evaluation Factors	See the results in Rae et al. (2021) which analyzes models trained on the same text corpus.
Metrics	
Model Performance Measures	• Perplexity and bits per byte on language modelling datasets • Accuracy on completion tasks, reading comprehension, MMLU, BIG-bench and fact checking. • Exact match accuracy for question answering. • Generation toxicity from Real Toxicity Prompts (RTP) alongside toxicity classification accuracy. • Gender and occupation bias. Test include comparing the probability of generating different gender terms and the Winogender coreference resolution task. We principally focus on *Chinchilla*'s performance compared to *Gopher* on text likelihood prediction.
Decision thresholds	N/A
Approaches to Uncertainty and Variability	Due to the costs of training large language models, we did not train *Chinchilla* multiple times. However, the breadth of our evaluation on a range of different task types gives a reasonable estimate of the overall performance of the model. Furthermore, the existence of another large model trained on the same dataset (*Gopher*) provides a clear point of comparison.
Evaluation Data	

Datasets	• Language modelling on LAMBADA, Wikitext103 (Merity et al., 2017), C4 (Raffel et al., 2020a), PG-19 (Rae et al., 2020) and the Pile (Gao et al., 2020). • Language understanding, real world knowledge, mathematical and logical reasoning on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020) and on the "Beyond the Imitation Game Benchmark" (BIG-bench) (BIG-bench collaboration, 2021). • Question answering (closed book) on Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). • Reading comprehension on RACE (Lai et al., 2017) • Common sense understanding on HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), Winogrande (Sakaguchi et al., 2020), SIQA (Sap et al., 2019), BoolQ (Clark et al., 2019), and TruthfulQA (Lin et al., 2021).
Motivation	We chose evaluations from Rae et al. (2021) to allow us to most directly compare to *Gopher*.
Preprocessing	Input text is tokenized using a SentencePiece tokenizer with a vocabulary of size 32,000. Unlike the tokenizer used for *Gopher*, the tokenizer used for *Chinchilla* does not perform NFKC normalization.
Training Data	
The same dataset is used as in Rae et al. (2021). Differences in sampling are shown in Table A1.	
Quantitative Analyses	
Unitary Results	Section 4.2 gives a detailed description of our analysis. Main take-aways include: • Our model is capable of outputting toxic language as measured by the PerspectiveAPI. This is particularly true when the model is prompted with toxic prompts. • Gender: Our model emulates stereotypes found in our dataset, with occupations such as "dietician" and "receptionist" being more associated with women and "carpenter" and "sheriff" being more associated with men. • Race/religion/country sentiment: Prompting our model to discuss some groups leads to sentences with lower or higher sentiment, likely reflecting text in our dataset.

Intersectional Results	We did not investigate intersectional biases.
Ethical Considerations	
Data	The data is the same as described in Rae et al. (2021).
Human Life	The model is not intended to inform decisions about matters central to human life or flourishing.
Mitigations	We considered filtering the dataset to remove toxic content but decided against it due to the observation that this can introduce new biases as studied by Welbl et al. (2021). More work is needed on mitigation approaches to toxic content and other types of risks associated with language models, such as those discussed in Weidinger et al. (2021).
Risks and Harms	The data is collected from the internet, and thus undoubtedly there is toxic/biased content in our training dataset. Furthermore, it is likely that personal information is also in the dataset that has been used to train our models. We defer to the more detailed discussion in Weidinger et al. (2021).
Use Cases	Especially fraught use cases include the generation of factually incorrect information with the intent of distributing it or using the model to generate racist, sexist or otherwise toxic text with harmful intent. Many more use cases that could cause harm exist. Such applications to malicious use are discussed in detail in Weidinger et al. (2021).

Table A8 | ***Chinchilla* model card.** We follow the framework presented in Mitchell et al. (2019).

J. List of trained models

In Table A9 we list the model size and configuration of all models used in this study. Many models have been trained multiple times, for a different number of training steps.

Task	*Chinchilla*	*Gopher*	Task	*Chinchilla*	*Gopher*
hyperbaton	54.2	51.7	movie_dialog_same_or_diff	54.5	50.7
causal_judgment	57.4	50.8	winowhy	62.5	56.7
formal_fallacies_syllogisms_neg	52.1	50.7	movie_recommendation	75.6	50.5
crash_blossom	47.6	63.6	moral_permissibility	57.3	55.1
discourse_marker_prediction	13.1	11.7	strategyqa	68.3	61.0
general_knowledge_json	94.3	93.9	nonsense_words_grammar	78.0	61.4
sports_understanding	71.0	54.9	metaphor_boolean	93.1	59.3
implicit_relations	49.4	36.4	navigate	52.6	51.1
penguins_in_a_table	48.7	40.6	presuppositions_as_nli	49.9	34.0
intent_recognition	92.8	88.7	temporal_sequences	32.0	19.0
reasoning_about_colored_objects	59.7	49.2	question_selection	52.6	41.4
logic_grid_puzzle	44.0	35.1	logical_fallacy_detection	72.1	58.9
timedial	68.8	50.9	physical_intuition	79.0	59.7
epistemic_reasoning	60.6	56.4	physics_mc	65.5	50.9
ruin_names	47.1	38.6	identify_odd_metaphor	68.8	38.6
hindu_knowledge	91.4	80.0	understanding_fables	60.3	39.6
misconceptions	65.3	61.7	logical_sequence	64.1	36.4
implicatures	75.0	62.0	mathematical_induction	47.3	57.6
disambiguation_q	54.7	45.5	fantasy_reasoning	69.0	64.1
known_unknowns	65.2	63.6	SNARKS	58.6	48.3
dark_humor_detection	66.2	83.1	crass_ai	75.0	56.8
analogical_similarity	38.1	17.2	entailed_polarity	94.0	89.5
sentence_ambiguity	71.7	69.1	irony_identification	73.0	69.7
riddle_sense	85.7	68.2	evaluating_info_essentiality	17.6	16.7
date_understanding	52.3	44.1	phrase_relatedness	94.0	81.8
analytic_entailment	67.1	53.0	novel_concepts	65.6	59.1
odd_one_out	70.9	32.5	empirical_judgments	67.7	52.5
logical_args	56.2	59.1	figure_of_speech_detection	63.3	52.7
alignment_questionnaire	91.3	79.2	english_proverbs	82.4	57.6
similarities_abstraction	87.0	81.8	Human_organs_senses_mcc	85.7	84.8
anachronisms	69.1	56.4	gre_reading_comprehension	53.1	27.3

Table A7 | ***Chinchilla* BIG-bench results.** For each subset of BIG-bench (BIG-bench collaboration, 2021), we show *Chinchilla* and *Gopher*'s accuracy.

Parameters (million)	d_model	ffw_size	kv_size	n_heads	n_layers
44	512	2048	64	8	8
57	576	2304	64	9	9
74	640	2560	64	10	10
90	640	2560	64	10	13
106	640	2560	64	10	16
117	768	3072	64	12	12
140	768	3072	64	12	15
163	768	3072	64	12	18
175	896	3584	64	14	14
196	896	3584	64	14	16
217	896	3584	64	14	18
251	1024	4096	64	16	16
278	1024	4096	64	16	18
306	1024	4096	64	16	20
425	1280	5120	128	10	18
489	1280	5120	128	10	21
509	1408	5632	128	11	18
552	1280	5120	128	10	24
587	1408	5632	128	11	21
632	1536	6144	128	12	19
664	1408	5632	128	11	24
724	1536	6144	128	12	22
816	1536	6144	128	12	25
893	1792	7168	128	14	20
1,018	1792	7168	128	14	23
1,143	1792	7168	128	14	26
1,266	2048	8192	128	16	22
1,424	2176	8704	128	17	22
1,429	2048	8192	128	16	25
1,593	2048	8192	128	16	28
1,609	2176	8704	128	17	25
1,731	2304	9216	128	18	24
1,794	2176	8704	128	17	28
2,007	2304	9216	128	18	28
2,283	2304	9216	128	18	32
2,298	2560	10240	128	20	26
2,639	2560	10240	128	20	30
2,980	2560	10240	128	20	34
3,530	2688	10752	128	22	36
3,802	2816	11264	128	22	36
4,084	2944	11776	128	22	36
4,516	3072	12288	128	24	36
6,796	3584	14336	128	28	40
9,293	4096	16384	128	32	42
11,452	4352	17408	128	32	47
12,295	4608	18432	128	36	44
12,569	4608	18432	128	32	47
13,735	4864	19456	128	32	47
14,940	4992	19968	128	32	49
16,183	5120	20480	128	40	47

Table A9 | **All models.** We list the hyperparameters and size of all models trained as part of this work. Many shown models have been trained with multiple learning rate schedules/number of training tokens.

npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients

Hanyin Wang[1], Chufan Gao[2], Christopher Dantona[3], Bryan Hull[4] and Jimeng Sun[2,5]✉

In the U.S. inpatient payment system, the Diagnosis-Related Group (DRG) is pivotal, but its assignment process is inefficient. The study introduces DRG-LLaMA, an advanced large language model (LLM) fine-tuned on clinical notes to enhance DRGs assignment. Utilizing LLaMA as the foundational model and optimizing it through Low-Rank Adaptation (LoRA) on 236,192 MIMIC-IV discharge summaries, our DRG-LLaMA -7B model exhibited a noteworthy macro-averaged F1 score of 0.327, a top-1 prediction accuracy of 52.0%, and a macro-averaged Area Under the Curve (AUC) of 0.986, with a maximum input token length of 512. This model surpassed the performance of prior leading models in DRG prediction, showing a relative improvement of 40.3% and 35.7% in macro-averaged F1 score compared to ClinicalBERT and CAML, respectively. Applied to base DRG and complication or comorbidity (CC)/major complication or comorbidity (MCC) prediction, DRG-LLaMA achieved a top-1 prediction accuracy of 67.8% and 67.5%, respectively. Additionally, our findings indicate that DRG-LLaMA 's performance correlates with increased model parameters and input context lengths.

npj Digital Medicine (2024)7:16 ; https://doi.org/10.1038/s41746-023-00989-3

INTRODUCTION

The emergence of LLMs, such as GPT-3[1] and InstructGPT[2], has brought about a transformative shift in the landscape of Natural Language Processing (NLP). These LLMs have demonstrated exceptional capabilities across many NLP tasks in the general domain. However, the integration of LLMs into the medical field remains at a nascent stage within the academic community. Recent instances of progress highlight their significant potential, including OpenAI's GPT-4[3], Google's Med-PaLM2[4], and Google Deepmind's Med-PaLM M[5]. GPT-4 and Med-PaLM 2 have achieved impressive performance on the United States Medical Licensing Examination (USMLE), and Med-PaLM M can even classify radiology images. Nonetheless, the medical domain introduces elevated concerns regarding safety and privacy, necessitating detailed analysis regarding the performance and limitations of LLMs to address the inherent risks such as hallucination, bias, and reasoning deficiencies[6].

Since its inception by Medicare in 1983, DRG has served as the foundation for the inpatient prospective payment system within the United States[7]. Each distinct DRG code is delineated by a particular set of patient attributes, including principal diagnosis, specific secondary diagnoses, procedures, sex and discharge status[8]. Traditionally, the assignment of DRGs constitutes a labor-intensive manual endeavor undertaken by coding specialists, typically subsequent to a patient's discharge. Given the pivotal role of DRGs and their bundled metrics (e.g., case-mix index, geometric length of stay) in the operational and financial performance of hospitals, a pressing interest exists in the accurate early prediction of DRGs during a patient's hospitalization. This prediction is vital for efficacious resource planning and allocation. The task of DRG prediction presents distinct challenges compared to automated International Classification of Diseases (ICD) coding. This distinction stems from differences in the nature of the task: DRGs involve multi-class classification, where one DRG code is assigned to each visit, in contrast to the multi-label classification of ICDs, where multiple codes may apply to a single visit[9]. Additionally, the hierarchical structure of the codes, such as the presence of a principal diagnosis in DRGs, and the context of utilization in hospital operations further differentiate the two tasks[8]. Previous studies have showcased advancements in DRGs classification accuracy through various machine-learning algorithms[10] and deep neural networks[11]. More recently, a deep learning-based NLP model leveraging adjusted Convolutional Attention for Multi-Label Classification (CAML) has been applied to predict DRGs based on clinical notes and yielded promising outcomes[12,13].

With LLM's remarkable natural language synthesis and generating capabilities, we hypothesize LLM could be applied to effectively predict DRGs directly from clinical notes. In this work, we present DRG-LLaMA, a fine-tuned LLM derived from LLaMA[14]. DRG-LLaMA is trained on discharge summaries from the MIMIC-IV dataset for the task of DRG prediction. In our investigation, we approached DRG prediction from two perspectives: 1) as a single-label classification task, where the model makes an end-to-end prediction of the DRG label, and 2) as a two-label classification task, where the model predicts base DRG and CC/MCC status as two separate labels, followed by the inference of the final DRG label from these two components (i.e., base DRG and CC/MCC status). Our work revealed superior performance of DRG-LLaMA in DRG prediction compared to the previously reported leading models of CAML[13] and ClinicalBERT[15].

RESULTS

Study cohort

A summary of the study cohort and data preprocessing steps was shown in Fig. 1. We focused on hospital stays with Medicare severity-DRGs (MS-DRGs) within the MIMIC-IV dataset. The "brief

[1]Division of Hospital Internal Medicine, Mayo Clinic Health System, Mankato, MN, USA. [2]Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA. [3]Enterprise Inpatient Clinical Documentation Integrity, Mayo Clinic, Rochester, MN, USA. [4]Division of Hospital Internal Medicine, Mayo Clinic, Phoenix, AZ, USA. [5]Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL, USA. ✉email: jimeng@illinois.edu

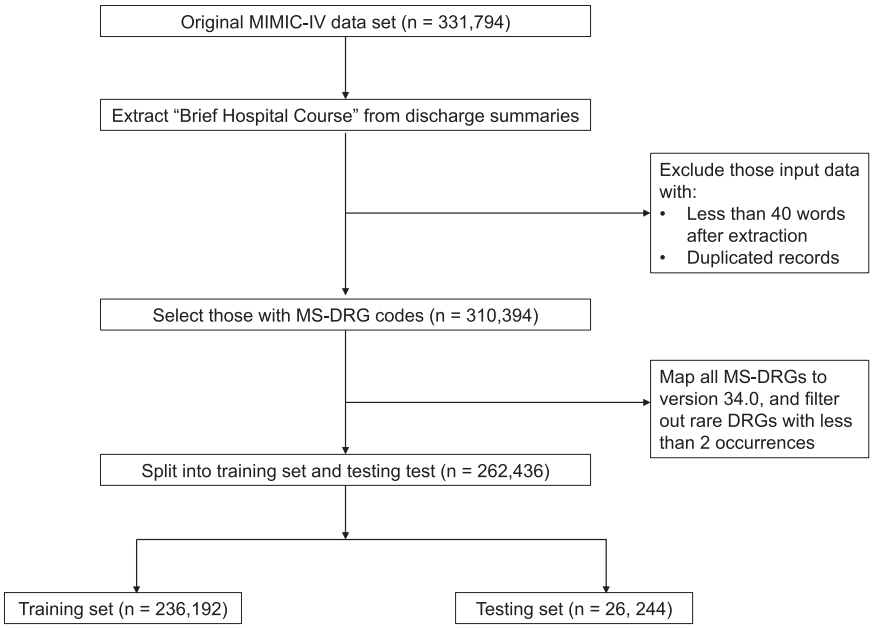


Fig. 1 Flow diagram of the cohort processing steps. We used regular expressions to extract the "brief hospital course" section from discharge summaries in MIMIC-IV dataset as input text. We filtered the discharge summaries that were of low quality, identified by either duplicated content or containing less than 40 words. We focused on MS-DRGs and consolidated all MS-DRG codes to version 34.0. Additionally, we filtered out rare DRGs with less than 2 occurrences in the cohort.

Table 1. Main Results on DRG prediction with a max input token size of 512.

Model	DRG set	MACRO-F1	ACC@1	ACC@5	ACC@10	MACRO-AUC	MICRO-AUC	Number (%) of cases
DRG-LLaMA -7B	All DRGs	**0.327 (0.004)**	**0.520 (0.003)**	**0.848 (0.002)**	**0.912 (0.002)**	**0.986 (0.001)**	**0.994 (0.000)**	26,244 (100.0)
	Top 300 DRGs	0.497 (0.005)	0.557 (0.004)	0.876 (0.002)	0.932 (0.001)	0.988 (0.000)	0.995 (0.000)	22,940 (87.4)
	Top 50 DRGs	0.700 (0.004)	0.666 (0.004)	0.931 (0.002)	0.965 (0.001)	0.989 (0.000)	0.998 (0.000)	10,270 (39.1)
	Top 30 DRGs	0.737 (0.005)	0.694 (0.005)	0.941 (0.003)	0.971 (0.002)	0.988 (0.001)	0.998 (0.000)	7,666 (29.2)
ClinicalBERT	All DRGs	0.233 (0.003)	0.502 (0.003)	0.815 (0.002)	0.881 (0.002)	0.979 (0.001)	0.991 (0.000)	26,244 (100.0)
CAML	All DRGs	0.241 (0.003)	0.447 (0.002)	0.785 (0.002)	0.865 (0.002)	0.976 (0.001)	0.991 (0.000)	26,244 (100.0)

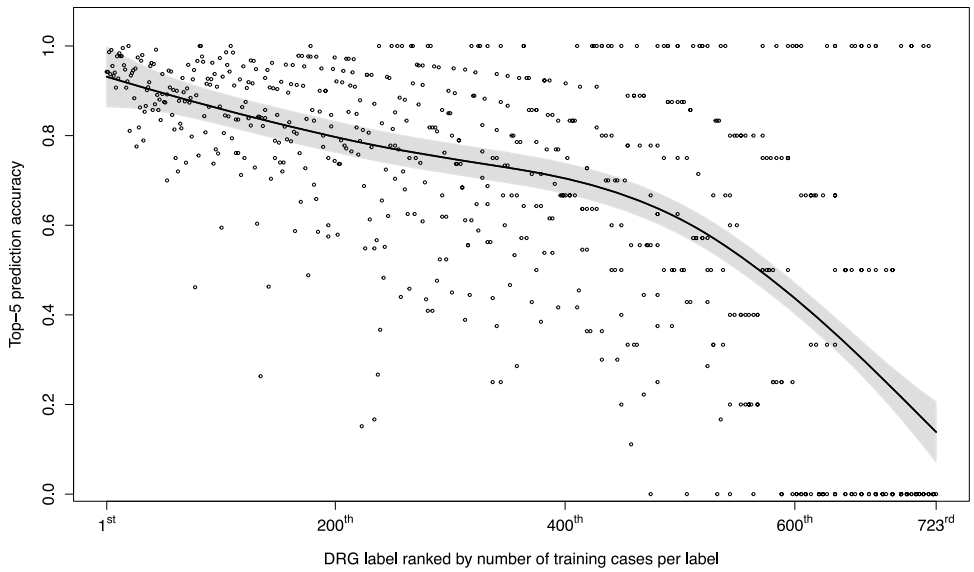
F1 and AUC scores were calculated using macro-averaged or micro-averaged method as shown in the header. Notably, in a multi-class classification problem, micro-averaged F1 score is equal to top-1 prediction accuracy when labels of all classes are considered. Accuracy @1, @5 and @10 measure whether the top-1, top-5 and top-10 predictions by the model contain correct DRG code, respectively. Standard deviations are shown in parentheses and calculated using a bootstrapping procedure. Top DRGs are selected based on the number of cases per DRG in the dataset. Number (%) of cases represents hospital stays covered by the given DRG group in the testing set. Bolded scores denote the best performance with respect to the task. DRG-LLaMA outperformed ClinicalBERT and CAML across all evaluation metrics, with better performance in more frequent DRGs. *DRG* denotes diagnostis-related group, *AUC* denotes area under the receiver operating characteristic curve, and *ACC* denotes accuracy.

hospital course" section from discharge summary was extracted to serve as input text. We also filtered out low-quality discharge summaries and rare DRGs with less than 2 occurrences in the cohort. 90% of the data was allocated as training set while the rest 10% as testing set, and this partitioning was stratified on DRGs. The training and testing set contains 738 and 723 unique DRG labels, respectively. There is no significant difference in the average word counts in the training vs. testing set (398 vs. 399; $p = 0.51$ from two-sided t-test). The distribution of cases per DRG is imbalanced, with a median number of 124.5 in the training set (Supplementary Fig. 1).

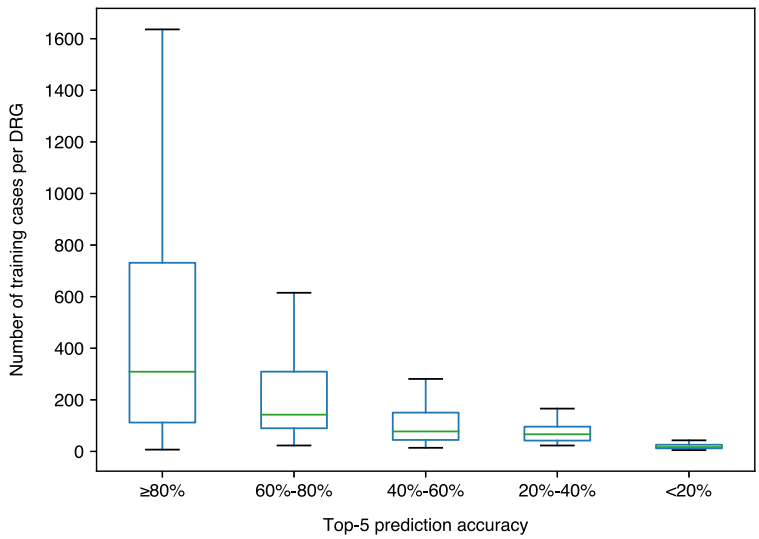
DRG prediction as a single-label classification task

We presented the results with a maximum input token size of 512 in Table 1. DRG-LLaMA consistently outperformed ClinicalBERT and CAML across all evaluation metrics, with the most notable contrast seen in macro-F1 score (showing a relative improvement of 40.3% and 35.7% compared to ClinicalBERT and CAML, respectively). The accuracy of top-1 and top-5 predictions achieved by our fine-tuned DRG-LLaMA -7B model was 52.0% and 84.8%, respectively. When only considering the most frequent 300 DRGs, the top-1 accuracy improved to 55.7%, and this further increased to 69.4% in the most frequent 30 DRGs. As expected, DRG-LLaMA 's performance declined in less frequent DRGs (Fig. 2a). When compared to CAML, ClinicalBERT achieved higher AUC and top-1 prediction accuracy but lower macro-averaged F1 score. High AUC scores were obtained for all models due to the many infrequent DRG classes, resulting in high true negative predictions for all negative class predictions[13].

We investigated DRG-LLaMA 's performance across varying model sizes and input context lengths (Table 2), observing a consistent improvement in all evaluation metrics with larger models and longer input contexts, measured in maximum token numbers. The optimal configuration, utilizing a 13B LLaMA model and a maximum input token size of 1024, achieved a top-1 prediction accuracy of 54.6%, a top-5 prediction accuracy of 86.5%, and a macro-F1 score of 0.361.



(a)



(b)

Fig. 2 Relationship between training cases per DRG and prediction accuracy by `DRG-LLaMA`**.** Results from `DRG-LLaMA` -7B with a maximum input token size of 512. **a** Scatter plot of top-5 prediction accuracy versus DRG ranks by number of training cases. Y-axis is top-5 prediction accuracy of each DRG label. X-axis is the rank of the 723 DRGs by their number of training cases, where DRG ranked 1st has the most training cases, and DRG ranked 723rd has the least training cases. Black dots indicate individual DRGs. The solid line represents smoothing spline estimated relationship (generalized cross-validation score: 0.055). The gray shaded area denotes a 95% Bayesian confidence interval for the smoothing spline estimated function. As expected, `DRG-LLaMA` 's performance declined in less frequent DRGs. **b** Boxplot of training cases per DRG with groups of different prediction accuracy. DRGs are grouped by range of top-5 prediction accuracy as shown in X-axis. Y-axis is the number of training cases per DRG. The green line represents the median value; the box limits show the interquartile range (IQR) from the first (Q1) to third (Q3) quartiles; the whiskers extend to the furthest data point within Q1-1.5*IQR (bottom) and Q3+1.5*IQR (top). DRG groups with better prediction performance generally have a greater number of training cases, although there is a large variance in the number of training cases within the best-performing group.

DRG prediction as a two-label classification task

In the two-label approach, we first dissect each DRG into two distinct components: a base DRG label and a CC/MCC label (denoting complication or comorbidity / major complication or comorbidity). This dissection process was based on the composition delineated within the MS-DRG v34.0 definitions manual[8]. The five distinct labels attributed to CC/MCC are as follows: "without CC/MCC", "with CC", "with MCC", "without MCC", and "not applicable". As an example, in DRG code 53 of "spinal disorders and injuries without CC/MCC," "spinal disorders and injuries" represents the base DRG label, while "without CC/MCC" serves as the CC/MCC label. Following this mapping process, the 738 DRG codes were converted into a combination of 340 base DRG labels each paired with one of the five CC/MCC labels. Results of the two-label approach using `DRG-LLaMA` -7B with a maximum input token size of 512 was shown in Table 3. The top-1 prediction accuracy for base DRG and CC/MCC reached 67.8% and 67.5%, respectively. This result suggests that predicting the principal

Table 2. DRG-LLaMA performance on different model and max input token sizes.

Model size	Max input token size	MACRO-F1	ACC@1	ACC@5	ACC@10	MACRO-AUC	MICRO-AUC
13B	1024	**0.361 (0.004)**	**0.546 (0.003)**	**0.865 (0.002)**	**0.925 (0.001)**	**0.986 (0.001)**	**0.994 (0.000)**
	512	0.334 (0.005)	0.524 (0.002)	0.853 (0.002)	0.914 (0.002)	0.984 (0.001)	0.993 (0.000)
	340	0.312 (0.006)	0.499 (0.003)	0.834 (0.002)	0.902 (0.002)	0.983 (0.001)	0.992 (0.000)
7B	1024	0.346 (0.004)	0.539 (0.003)	0.861 (0.002)	0.923 (0.001)	0.986 (0.001)	0.994 (0.000)
	512	0.327 (0.004)	0.520 (0.003)	0.848 (0.002)	0.912 (0.002)	0.986 (0.001)	0.994 (0.000)
	340	0.303 (0.005)	0.493 (0.003)	0.828 (0.002)	0.896 (0.002)	0.981 (0.001)	0.992 (0.001)

Experiments were performed on LLaMA with a size of 7 billion and 13 billion parameters. Bolded scores denote the best performance. We observed that DRG-LLaMA 's performance consistently improved with larger models and longer input contexts.

Table 3. Main Results on DRG prediction as a two-label task with a max input token size of 512.

Component	MACRO-F1	ACC@1	ACC@5	ACC@10	MACRO-AUC	MICRO-AUC	Number of labels
Base DRG	0.520 (0.005)	0.678 (0.002)	0.912 (0.001)	0.953 (0.001)	0.990 (0.001)	0.995 (0.000)	340
CC/MCC	0.680 (0.003)	0.675 (0.003)	–	–	0.909 (0.001)	0.918 (0.001)	5
DRG	–	0.515 (0.003)	–	–	–	–	738

Experiments were performed with DRG-LLaMA -7B and a maximum input token size of 512. The top-1 prediction accuracy for base DRG and CC/MCC reached 67.8% and 67.5%, respectively. A top-1 prediction accuracy of 51.5% was achieved by employing the mapping rule on base DRG and CC/MCC labels, as elaborated in the method section.

diagnosis or procedure without considering CC/MCC is a significantly easier task on its own.

Upon integrating a mapping rule designed to infer DRGs through the combination of base DRG and CC/MCC labels, the accuracy reached 51.5% across all DRGs. Notably, this performance was comparable with the accuracy attained in the single-label approach of 52.0% using the same base model, showing that the LLM was able to achieve state-of-the-art performance via either classification setting.

Error analysis

As noted above, a correlation exists between the number of training cases and prediction performance. DRGs with a top-5 prediction accuracy exceeding 80% are associated with a median of 309 training cases per label. In contrast, those DRGs with a top-5 accuracy below 20% are associated with only a median of 17 training cases per label (as shown in Fig. 2b). However, other factors, such as the type of DRG, also affect prediction performance. For instance, out of the DRGs with a top-1 prediction accuracy of 100%, 8 out of 9 are surgical DRGs, which have distinct hospital courses that make them easier for the model to comprehend (as listed in Supplementary Table 2). We randomly selected 10 samples from the subset where the model presented erroneous predictions within its top ten outcomes for manual error analysis (as listed in Table 4). Broadly, the identified errors were categorized as follows: erroneous CC/MCC (1/10), correct information needed for DRG prediction unavailable (1/10), difficulty in selecting correct base DRG (3/10), inadequate clinical concept extraction (4/10) and an isolated case of a plausible incorrect DRG label (1/10). Certain errors, like inadequate clinical concept extraction, indicate the model's weaknesses. Other errors, such as the difficulty in selecting the base DRG, likely stem from the intricacies of the DRG assignment rules. Furthermore, errors such as the unavailability of correct information required for DRG prediction underscore the limitations of solely relying on discharge summaries for DRG predictions.

DISCUSSION

Language models based on the transformer architecture, either pretrained or fine-tuned using biomedical corpora, have demonstrated efficacy across a spectrum of NLP benchmarks within the biomedical realm[16–18]. When contrasted with their predecessors rooted in the BERT architecture[19], LLMs stand out due to their substantial size and their pretraining on expansive, cross-disciplinary text corpora. LLMs exhibit a notable capacity for comprehending and reasoning with clinical knowledge. Without domain-specific fine-tuning or specialized prompt crafting, GPT-4 exceeded the passing score on USMLE by over 20 points and set a new state-of-the-art[3]. On this premise, it is plausible to speculate that once attuned to the medical domain, an LLM could deliver robust performance across diverse NLP tasks, including the prediction of DRGs.

Toward deploying a local LLM, we used LLaMA, a robust and openly accessible foundational LLM with parameters ranging from 7 billion to 65 billion[14]. Instruction-following models fine-tuned from LLaMA such as Alpaca[20] and Vicuna[21], exhibit performance on par with GPT-3.5. Within the medical context, several groups have directed their efforts toward fine-tuning LLaMA. Notable examples among these are ChatDoctor (trained on authentic patient-physician dialogues), HuaTuo (fine-tuned with a Chinese medical knowledge graph), and PMC-LLaMA (fine-tuned on biomedical academic papers)[22–24]. These LLaMA-based models focused on medical question answering, yielding encouraging outcomes.

In this study, we demonstrated superior performance of the fine-tuned LLaMA in the text classification task of DRG prediction. Previous studies have underscored the effectiveness of employing diverse machine learning algorithms and deep neural networks for DRG prediction within healthcare systems outside the United States[10,11]. These studies focused on using structured data as input variables instead of clinical text. More recently, CAML model exhibited superior ability to predict DRGs[13]. CAML model, exclusively utilizing clinical notes, surpassed the performance of a Long Short-Term Memory (LSTM) model using structured clinical variables[13]. When compared with ClinicalBERT, CAML provided improved F1 scores but lower AUC[13,15]. We observed that DRG-

Table 4. Example of incorrect DRG predictions.

Case ID	Pertinent narratives in discharge summary	True DRG	Predicted DRG	Comment
Case 1	altered mental status…respiratory failure…acute blood loss anemia and anemia of chronic disease…clostridium difficile infection…hypotension…was initially on levophed and dopamine…	Heart failure and shock with mcc	Respiratory system diagnosis with ventilator support 96 hours	Difficulty in selecting base DRG
Case 2	gastrointestinal bleeding…most likely ischemic colitis…viral gastroenteritis…acute renal failure…anemia…	Renal failure with cc	Other digestive system diagnoses with cc	Difficulty in selecting base DRG
Case 3	worsening diabetic foot ulcer…diabetic foot infection…svt… cardiology was consulted…	Cellulitis without mcc	Diabetes with cc	Inadequate clinical concept extraction
Case 4	neutropenic fevers…infectious workup was negative except for a urine culture growing enterococcus…pt is neutropenic, thrombocytopenic, and anemic…hiv-stable…	Kidney and urinary tract infections without mcc	Major hematological and immunological diagnoses except sickle cell crisis and coagulation disorders with mcc	Difficulty in selecting base DRG
Case 5	reported chest pain…soliatry episode of nsvt..ua without pyuria…safe for d/c home…	Esophagitis gastroenteristis and miscellaneous digestive disorders without mcc	Cardiac arrhythmia and conduction disorders with cc	Correct information needed for DRG prediction not available
Case 6	septic arthritis, likely seeded by her recurrence of her e. coli bacteremia…rheum and id recommend wash out…wash out was deferred by orthopedics…	Septicemia or severe sepsis without mv 96 hours with mcc	Revision of hip or knee replacement with mcc	Inadequate clinical concept extraction
Case 7	acute to subacute hyponatremia…admitted with low na 120… uti with evidence of pyuria…	Kidney and urinary tract infections without mcc	Renal failure with cc	Inadequate clinical concept extraction
Case 8	presents with diffuse acute-on-chronic abdominal pain…gi bleed…treated with octreotide drip and pantoprazole iv… capsule endoscopy was performed…encephalopathy…visual hallucinations…	Septicemia or severe sepsis without mv 96 hours with mcc	G.i. hemorrhage with cc	Possible incorrect DRG label
Case 9	admitted for altered mental status…delirium….silent aspiration for which received a peg tube…hypertension treated with amlodipine…osa	Esophagitis gastroenteristis and miscellaneous digestive disorders without mcc	Esophagitis gastroenteritis and miscellaneous digestive disorders with mcc	Erroneous cc/mcc
Case 10	presents with word finding difficulties and lethargy…eeg showed moderate encephalopathy…ams was likely due to overmedication…followed by psychiatry - seroquel and abilify were held…	Psychoses	Other disorders of nervous system with cc	Inadequate clinical concept extraction

We manually reviewed 10 cases for error analysis. For each case, we extracted most pertinent medical problems and their narratives from discharge summaries. Certain errors, like inadequate clinical concept extraction, indicate the model's weaknesses. Other errors, such as the difficulty in selecting the base DRG, likely stem from the intricacies of the DRG assignment rules. Furthermore, errors such as the unavailability of correct information required for DRG prediction underscore the limitations of solely relying on discharge summaries for DRG predictions.

LLaMA outperformed prior leading models of ClinicalBERT and CAML.

ClinicalBERT and CAML already stand as robust baselines, with the added benefit of much faster training times (supplement Table 1). While BERT-based models have a maximum input length of 512 tokens, CAML has the flexibility to handle longer context[13,19]. We also observed that the performance of DRG-LLaMA enhanced with the utilization of larger models and longer input context length. Interestingly, a recent study revealed that the optimal performance of LLMs is attained when pertinent information is positioned at either the beginning or the end of the input context, with a decline as the input context expands[25]. In our constrained experiments conducted with a maximum input token limit up to 1024, we have yet to encounter this limitation. In our study, the performance of both the baseline models and DRG-LLaMA surpassed the outcomes reported in prior research[13]. Beyond the substantially larger training dataset employed in MIMIC-IV compared to MIMIC-III (236,192 vs. 17,815), it is plausible that this enhanced performance is predominantly linked to our strategic input data selection.

The study by Liu et al.[13] included only clinical notes charted up to 48 hours post-admission or 48 hours after ICU admission. In the MIMIC-III database, a large portion of records during this time window comprises nursing and radiology notes, potentially lacking the pivotal admission History of Present Illness (HPI) notes. In contrast, our methodology entailed the utilization of discharge summaries as the input data source. Discharge summary is a comprehensive clinical narrative encapsulating pivotal events, diagnostics, and treatments during hospitalization. To accommodate the input token limitations of LLaMA, we exclusively focused on the "brief hospital course" section of the summary, intentionally excluding other segments such as physical examinations, radiology, laboratory, and medication list. Additionally, to enhance data consistency, we formulated an algorithm aimed at addressing discrepancies in DRG nomenclature and assignments across different years.

In the context of the DRG system, a DRG code comprises a base DRG and a CC/MCC status. The base DRG represents the principal diagnosis (for medical cases) or procedures (for surgical cases) leading to the patient's admission. Meanwhile, CC/MCC categorizations gauge the severity of the patient's condition. In the 34.0 version of the MS-DRG system, there are 154 three-way split DRGs, 44 two-way split DRGs with MCC/CC and no CC, 65 two-way split DRGs with MCC and CC/no CC, and 77 base DRGs with no splits (examples in Supplementary Note 1)[8]. We experimented to resemble this structure through a two-label DRG prediction strategy. Surprisingly, the top-1 accuracy for CC/MCC stands at 67.5%, similar to 67.8% of the base DRG despite the considerably smaller label count (5 labels in CC/MCC vs. 340 labels in base DRG). These unexpected results likely stem from the noisy nature of CC/MCC assignment. For instance, the DRG code "pulmonary edema and respiratory failure" does not have a CC/MCC split. Therefore, a hospital stay with this DRG code may truly contain MCC, but the MCC would not be labeled as positive in the training set. To address this challenge, we formulated rules in both the DRGs dissection phase (extracting base DRGs and CC/MCC from DRGs) and the inference phase (deriving DRGs based on base DRGs and CC/MCC). These rules cater to various split scenarios, thus improving accuracy. Implementing such rules has culminated in a final DRG prediction accuracy close to single-label prediction (51.5% vs. 52.0%).

Our error analysis also revealed intriguing observations. While certain vulnerabilities (e.g., erroneous CC/MCC classification and inadequate clinical concept extraction) present opportunities that theoretically can be addressed through employment of larger LLM and more data, other challenges likely stem from inherent limitations within our training data setup. For instance, in Case 2 in Table 4, despite the discharge summary providing a more comprehensive discussion on gastrointestinal bleeding compared to acute renal failure, the latter was deemed the correct base DRG. This selection is guided by the DRG assignment rule[8], a factor extending beyond the scope of what is directly evident within the discharge summary.

Our study has several limitations. 1) We were limited by the constraints of the MIMIC-IV dataset and could only use discharge summaries as input data, which are only available after the patient is discharged from the hospital. However, an effective alternative for predicting early DRGs would be to utilize HPI notes and/or Emergency Department (ED) notes. This approach has the potential to significantly impact hospital operations. The "assessment and plan" in HPI notes are similar in structure to the "brief hospital course" in discharge summaries. Thus, LLMs might find it easier to extract information related to the principal diagnosis from these notes, given their earlier time stamp in the hospitalization process. 2) We were also restricted by computational resource limitations, so we could only experiment with the LLaMA model up to a parameter size of 13 billion. Unfortunately, we couldn't perform an extensive hyperparameter search. The largest LLaMA models have over 65 billion parameters.

The results presented in this study highlight the potential of adapting LLMs for medical purposes, particularly in predicting DRGs. Future research should involve collaborating with healthcare systems and utilizing admission notes to enable early DRG prediction. Additionally, our findings suggest that experiments utilizing the latest LLMs, including the recently launched 70-billion-parameter LLaMA-2 model with a maximum context length of 4096 tokens[26], should be considered. Finally, a crucial area for exploration concerns the practical implications of such DRG prediction, particularly when integrated into existing hospital coding workflows.

METHODS

Dataset and preprocessing

We conducted a study using the publicly available MIMIC-IV dataset, which comprises 431,231 unique hospital admissions from 299,712 patients admitted to an ICU or the ED of the Beth Israel Deaconess Medical Center in Boston, Massachusetts[27]. The dataset covers the period from 2008 to 2019. We used regular expressions to extract the "brief hospital course" section from the discharge summary as input text. We then filtered the discharge summaries that were of low quality, identified by either duplicated content or containing less than 40 words.

Our focus was on hospitalizations with MS-DRGs. We consolidated all MS-DRG codes to version 34.0, published in 2016 (detailed in the subsequent section)[8]. This version comprises a total of 757 DRG codes, with 738 being represented in our dataset. We allocated 90% of the data to the training set and the remaining 10% to the testing set, stratified by DRG codes.

Process to address different DRG versions

Centers for Medicare & Medicaid Service adjusts MS-DRG regulations annually, resulting in varying DRG assignments for identical conditions over time within the MIMIC-IV dataset[28]. To address this discrepancy, we designed an algorithm based on clinical knowledge to harmonize MS-DRG codes across different time points to a unified version—specifically, MS-DRG version 34.0[8]. The process include:

1. Standardize use of abbreviations and capitalization within DRGs. For example, we replaced all "W/O" to "WITHOUT", "CATH" to "CATHETERIZATION" and "PROC" to "PROCEDURES".
2. Using a fuzzy string match algorithm (TheFuzz: https://github.com/seatgeek/thefuzz) to find those DRGs not matching to any MS-DRG v.34 codes.

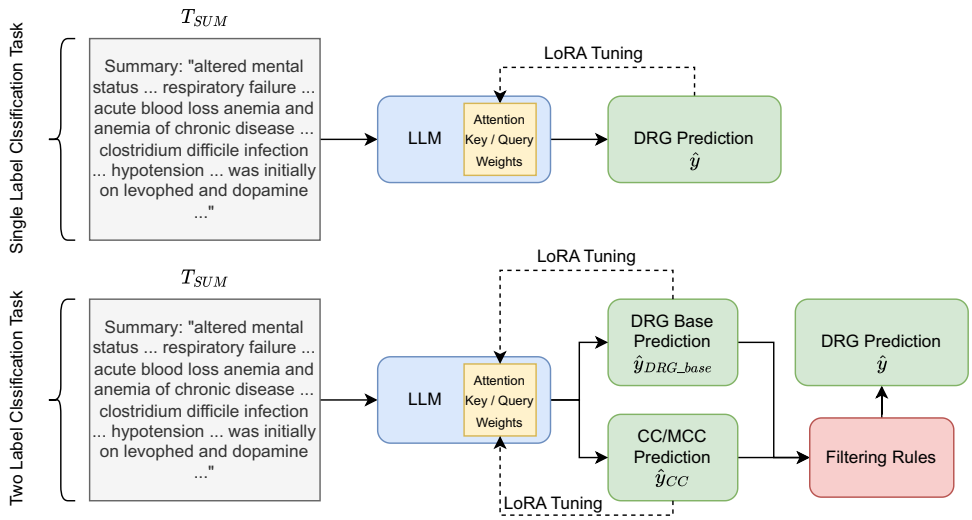


Fig. 3 An illustration of both approaches we tested. Single Label Prediction–which directly predicts the DRG code from the text–as well as Two Label Prediction–which breaks down the classification task into 2 tasks. The two predictions are then combined using filtering rules (discovered from data for each DRG) at inference time for the final DRG prediction. LoRA training is used to train the LLM due to computational constraints.

3. An internal medicine physician manually reviewed all DRG codes from step 2, and assigned these codes to the most appropriate MS-DRG v.34.0 codes if applicable. Subsequently, a domain expert specializing in inpatient Clinical Documentation Integrity (CDI) assessed the conversion table and independently verified the accuracy of the code assignments.
4. Of note, after above steps there are several historical DRGs left without appropriate DRG v.34 codes assignment. For example, "URINARY STONES W MCC" and "NASAL TRAUMA AND DEFORMITY WITH CC". These hospitalizations were excluded from the cohort.
5. Lastly, we filtered out rare DRGs with less than 2 occurrences in our cohort.

Model development

We performed fine-tuning of the LLaMA model using discharge summaries and DRG codes within the context of a classification task. Our approach includes two distinctive strategies (also shown in Fig. 3).

Single label approach. In this approach, the model generates a single-label multi-class prediction for the DRG code from a training set of natural text discharge summaries T_{SUM} and labels containing $(T_{SUM,i}, y_i) \in \mathcal{D}$. We omit the index notation i for the rest of the descriptions without loss of generality. First, let us tokenize T_{SUM} based on the LLaMA Tokenizer into $\boldsymbol{K} = tokenize(T_{SUM})$. $\boldsymbol{K}$ is a list of indices that index into learnable embedding weights. Let $LLM()$ be a function that outputs the embedding for each token after running the transformer model. Finally, the raw logits are calculated as

$$\hat{\boldsymbol{y}} = LLM(\boldsymbol{K})_{-1} \tag{1}$$

where we use the last token embedding of $LLM(\boldsymbol{K})$ as the predicted raw logit score of each DRG code $\hat{\boldsymbol{y}} \in \mathbb{R}^{738}$. Note that this logit score is the raw, unnormalized output of the last layer of the LLM. Before applying the activation function like the softmax function, which converts these scores to probabilities, the values produced by the network are referred to as logits.

The conventional categorical cross-entropy loss function for multi-class classification is used. i.e., a classic multi-class problem with loss: the target DRG y is an integer between 0 and 737 (note that we use an integer representing a specific DRG code for simplicity).

$$\ell(\hat{\boldsymbol{y}}, y) = -\log \frac{\exp(\hat{\boldsymbol{y}}_y)}{\sum_{c=1}^{C} \exp(\hat{\boldsymbol{y}}_c)} \tag{2}$$

Where $y \in \{0, 1, \ldots, 737\}$ is the target DRG, and $\hat{\boldsymbol{y}}_c$ is the cth index of $\hat{\boldsymbol{y}}$.

Two-label approach. In contrast, the two-label approach entails the model initially predicting the base DRG and the CC/MCC status as two separate classification tasks. Subsequently, a mapping rule (detailed in the subsequent section) is applied to derive DRG code. This approach entailed a loss function configured as the cross-entropy loss of the base DRG, plus half of the cross-entropy loss of the CC/MCC status.

More formally,

$$\ell(\hat{\boldsymbol{y}}, y) = \ell_{DRG_base}(\hat{\boldsymbol{y}}_{DRG_base}, y_{DRG_base}) + \lambda \ell_{CC}(\hat{\boldsymbol{y}}_{CC}, y_{CC}) \tag{3}$$

Where $\ell_{DRG_base}(\hat{\boldsymbol{y}}_{DRG_base}, y_{DRG_base})$ and $\ell_{CC}(\hat{\boldsymbol{y}}_{CC}, y_{CC})$ are also categorical cross entropy losses. We chose $\lambda = \frac{1}{2}$ for our work. As shown in Table 3, $y_{DRG_base} \in \{0, 1, \ldots, 339\}$ and $y_{CC} \in \{0, \ldots, 4\}$, representing the categories of ["without CC/MCC", "with CC", "with MCC", "without MCC", and "not applicable"] respectively.

To enable ease of implementation, we used an output logit dimension of $\hat{\boldsymbol{y}} \in \mathbb{R}^{340+5}$ and indexed the first 340 dimensions for $\hat{\boldsymbol{y}}_{DRG_base} = \hat{\boldsymbol{y}}_{0,\ldots,339}$ and indexed the last 5 dimensions for $\hat{\boldsymbol{y}}_{CC} = \hat{\boldsymbol{y}}_{340,\ldots,344}$. At inference time, we take the base DRG and CC/MCC predictions as the argmax of their respective logits.

$$\hat{y}_{DRG_base} = argmax_{\hat{y}_{DRG_base}}(\hat{\boldsymbol{y}}_{DRG_base}) \tag{4}$$

$$\hat{y}_{CC} = argmax_{\hat{y}_{CC}}(\hat{\boldsymbol{y}}_{CC}) \tag{5}$$

Subsequently, we apply the mapping rule, as detailed below, to derive the final DRG prediction from base DRG and CC/MCC labels.

Process to dissect and derive DRGs to/from base DRGs and CC/MCC. We first used regular expression to obtain principal diagnosis/procedures in MS-DRG v.34.0, by extracting strings prior to the description of CC/MCC. For example, in DRG 11 of "tacheostomy for face mouth and neck diagnoses with mcc", the principal diagnosis is "tacheostomy for face mouth and neck diagnoses". After this step, 340 principal diagnosis/procedures are identified as base DRGs.

We assigned CC/MCC status to one of the five labels: "without CC/MCC," "with CC," "with MCC," "without MCC," and "not applicable". Of note, an important detail is that if a DRG code does not explicitly describe CC/MCC status, we will assign a label of "not applicable". Such an example is DRG 69 "transient ischemia". We realize such classification might bring in noisy signals for models to learn (as a patient with "transient ischemia" can indeed has CC/MCC), but found it better than assigning to an alternative label such as "without CC/MCC" which would be more erroneous.

When inferencing DRGs from base DRGs and CC/MCC, we developed a rule based on logic and clinical knowledge. First, we evaluate whether predicted principal diagnosis/procedure matches target base DRG. Second, if the predicted CC/MCC label is in the CC/MCC set of the target base DRG, we make comparison directly. Third, for those predicted CC/MCC labels not in the CC/MCC set of the target base DRG, we apply a mapping procedure based on different MS-DRG splits as listed in Supplementary Note 1. For example, if a MS-DRG code has no split, such as DRG 69 "transient ischemia", then any CC/MCC predictions can be mapped to the correct DRG (as long as the base DRG matches). Another example would be MS-DRG 56 and 57, where there are two splits of CC/MCC status ("with MCC" and "without MCC"). In this case we will map predictions of "without CC/MCC", "with CC" and "not applicable" all to the label of "without MCC" for final inference.

Addressing Computational Constraints via LoRA Training. Given the constraints of available computational resources, an extensive hyperparameter search was not viable. Instead, our focus encompassed exploring the performance across diverse model sizes and token lengths. We used LoRA during training, which involves freezing the pre-trained model weights and incorporating trainable rank decomposition matrices into each layer of the transformer architecture[29]. Lora training of the attention mechanism is shown in Fig. 3.

As a quick summary, let us assume that we have original weight matrix $\boldsymbol{W_0} \in \mathbb{R}^{d\times k}$. LoRA works by adding a low-rank matrix to the original weight matrix: $\Delta\boldsymbol{W} + \boldsymbol{W_0}, \Delta\boldsymbol{W} = \boldsymbol{BA}$ where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$. Note that one should choose $r \ll \min(d, k)$ and only adapt the attention weights to ensure constraints on the dimensionality of the new weights and preserve original model performance. Training is only performed on this $\Delta\boldsymbol{W}$, and original model weights are kept the same. We also only tune the weights of the attention mechanism for further cost savings while preserving performance.

Training Details. Model training adopted standard Huggingface training framework and the sequence classification module[30]. Since LLaMA is a decoder-only (causal) model, we follow the traditional approach of using the embedding of the last token to do the classification, as other causal models (e.g. GPT-2[31]) do. Logits score of each DRG label was calculated from this linear output layer, and probabilities of DRGs could be derived using a softmax function.

We referenced the training protocol of Alpaca-Lora[32]. The model was quantized to 8-bit integer using bitsandbytes library[33]. Our model was trained using cross-entropy loss with the AdamW optimizer (learning rate $= 2\times10^{-5}$ and weight decay $= 0.01$) for 3 epochs on all training data and batch size of 4. Lora parameters were configured with r set to 8, an alpha value of 16, and a dropout rate of 0.05. All attention blocks were included in the Lora target modules. The training regimen for all `DRG-LLaMA` models were executed on a singular Nvidia RTX A6000 GPU with 48GB of graphics memory.

Baseline models

As baseline models for benchmarking, we selected CAML[12,13] and ClinicalBERT[15]. CAML is an adjusted convolutional neural network (CNN). In CAML, clinical notes are tokenized and embedded with pre-trained word embeddings to form input representations. Subsequently, inputs are passed on to a neural network with one-dimensional convolutions that pool CNN features using the attention mechanism. In line with the approach detailed in[13], our training of CAML included early stop when there was no improvement in micro-averaged F1 score for 10 consecutive epochs, with a maximum epochs of 50. All default hyperparameters were kept, except for max_seq_length which was set to 512.

ClinicalBERT was built upon BioBERT, a domain-specific BERT model pre-trained on PubMed abstracts and full-text articles from PubMed Central[16]. ClinicalBERT performed further pre-training of BioBERT using 2 million clinical notes from MIMIC-III[34]. In our fine-tuning process of ClinicalBERT, we conducted three training epochs, same as `DRG-LLaMA`. We set a learning rate of 2×10^{-5} and a batch size of 16, consistent with previous recommended practice for classification-oriented fine-tuning of BERT[19,35].

Statistical analysis

We used the implementation from[13] to calculate AUC and F1-score in both macro- and micro- approach for predictive models. We also reported accuracy of DRG prediction at top one, five and ten results. Standard deviations were calculated using a bootstrapping procedure with 30 iterations. For each bootstrap iteration, we randomly resampled the whole sample size from the testing set with replacement. Smoothing spline fit in Fig. 2a was performed using npreg package in R with generalized cross-validation method and default parameters[36].

Ethical concerns

MIMIC-IV is a free EHR dataset that is deidentified according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision[27]

Since we primarily used open source models such as LLaMA and ClinicalBERT from Huggingface, an open source repository of machine learning models[30] as well as CAML from github, and trained it on MIMIC, privacy risks are quite low. However, this risk should not be counted out when working with LLMs, and it is possible that LLaMA and ClinicalBERT may be trained on sensitive data in their respective pretrainining stages.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Access to MIMIC-IV can be requested at https://physionet.org/content/mimiciv/, which requires a signed safe usage agreement.

CODE AVAILABILITY

Scripts for this work were written in Python. They are available with accompanied documentation at https://github.com/hanyin88/drg-llama.

Received: 29 September 2023; Accepted: 19 December 2023;
Published online: 22 January 2024

REFERENCES

1. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* **33**, 1877–1901 (2020).
2. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.* **35**, 27730–27744 (2022).

3. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
4. Singhal, K. et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
5. Tu, T. et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334* (2023).
6. Au Yeung, J. et al. AI chatbots not yet ready for clinical use. *Front. Digital Health* **5**, 60 (2023).
7. Quinn, K. After the revolution: DRGs at age 30. *Ann. Internal Med.* **160**, 426–429 (2014).
8. CMS. ICD-10-CM/PCS MS-DRG v34. 0 Definitions Manual. https://www.cms.gov/icd10m/version34-fullcode-cms/fullcode_cms/P0001.html (2016).
9. Kaur, R., Ginige, J.A., Obst, O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst. Appl.* 118997 (2022).
10. Gartner, D., Kolisch, R., Neill, D. B. & Padman, R. Machine learning approaches for early DRG classification and resource allocation. *INFORMS J. Comput.* **27**, 718–734 (2015).
11. Islam, M.M., Li, G.H., Poly, T.N., Li, Y.C. Deepdrg: Performance of artificial intelligence model for real-time prediction of diagnosis-related groups. in *Healthcare*;9:1632MDPI (2021).
12. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695* (2018).
13. Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ Digi.I Med.* **4**, 103 (2021).
14. Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
15. Alsentzer, E. et al. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).
16. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
17. Huang, K., Altosaar, J., Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).
18. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transact. Comput. for Healthcare (HEALTH)* **3**, 1–23 (2021).
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
20. Taori, R. et al. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca (2023).
21. Chiang, W.L. et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna (2023).
22. Wang, H. et al. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975* (2023).
23. Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., You, Z. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454* (2023).
25. Liu, N.F. et al. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172* (2023).
26. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
27. Johnson, A. E. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
28. Johnson, A. Question about DRG codes in MIMIC-IV. https://github.com/MIT-LCP/mimic-code/issues/1561 (2023).
29. Hu, E.J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
30. Wolf, T. et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
31. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
32. Wang, E.J. Alpaca-Lora. https://github.com/tloen/alpaca-lora (2023).
33. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv preprint arXiv:2208.07339* (2022).
34. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
35. Adhikari, A., Ram, A., Tang, R., Lin, J. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398* (2019).
36. Helwig, N. npreg: Nonparametric Regression via Smoothing Splines. https://cran.r-project.org/web/packages/npreg/index.html (2021).

ACKNOWLEDGEMENTS

This research was supported by NSF award SCH-2205289, IIS-2034479, SCH-2014438. The funder played no role in the study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

AUTHOR CONTRIBUTIONS

H.W. designed, conducted, and analyzed the results of experiments. H.W. and C.G wrote the original draft. J.S. obtained funding and computing resource for the project. All authors contributed to the conceptualization of the research questions. All authors reviewed, revised, and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00989-3.

Correspondence and requests for materials should be addressed to Jimeng Sun.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 10, pp. 254-266
DOI:10.33168/JSMS.2024.1015

Methods and Applications of Fine-Tuning Llama-2 and Llama-Based Models: A Systematic Literature Analysis

Vincencius Christiano Tjokro, Samuel Ady Sanjaya*

Department of Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Banten, Indonesia

vincencius@student.umn.ac.id, samuel.ady@umn.ac.id (Corresponding Author)

Abstract. This Systematic Literature Review (SLR) scrutinizes 20 papers focusing on methods and applications for fine-tuning the open-source Llama language model. The findings underscore the efficacy of instruction-based tuning, demonstrating high performance across diverse domains, including medicine and psychology. The fine-tuned version Llama-2-Chat model, aligned with human preferences, emerges as a preferred base for subsequent tuning efforts. Despite its promise, scalability hurdles persist due to demanding computational requirements. A critical gap in the existing literature lies in the absence of a balanced evaluation of trade-offs among various fine-tuning approaches. Moreover, ethical considerations, particularly addressing bias and associated risks, demand greater attention in the deployment of tuned models. In conclusion, while instruction tuning holds significant promise for specializing Llama variants, overcoming limitations related to resource constraints, safety, and transparency is imperative for responsible real-world impact. This refined understanding emphasizes the need for a comprehensive assessment of fine-tuning methods and a conscientious approach to ethical considerations in the evolving landscape of large language models.

Keywords. Fine-Tuning, Llama, Llama-2, Large Language Model

1. Introduction

Chatbots are crafted to identify user input through pattern matching and retrieve information from a database to furnish a suitable response. Frequently employed to aid human tasks in disseminating information, chatbots find applications in various domains such as virtual learning systems and customer service (Richardson & Wicaksana, 2022). An example of chatbot technology involves the utilization of Large Language Models (LLMs). Large language model (LLM) refers to a language model built upon a Transformer by training huge corpus data and containing billions of parameters. This ability is inherited by LLM upon leveraging huge text training data and billions of parameters related to the neural scaling law theory (Zhao et al., 2023). Neural scaling law theory explains that the performance of LLM is highly related to the size of pre-training data, model parameters, and computational capability involved while training the model (Aghajanyan et al., 2023; Young, 2021). The advancement of LLM has experienced exponential growth in the past 5 years, marked by the emergence of ChatGPT by OpenAI in November 2022, reaching one million users within just 5 days (De Angelis et al., 2023). Instruction LLM such as ChatGPT inhibits an impressive generalization ability to comprehend instructions and produce human-like responses even on unseen cases or tasks (Wang, Liu, et al., 2023). This ability is inherited by ChatGPT due to its enormous 175B parameters, this comes with a limitation that makes it computationally expensive, furthermore, it is only accessible through API provided by OpenAI (*close-source*) (Ray, 2023). As an alternative to address those limitations, smaller LLM has emerged, marked by the introduction of Llama as the state-of-the-art (SOTA) open-source LLM.

Fine-tuning refers to the process of further training pre-trained models on specific domain data. There are several approaches of fine-tuning including transfer learning, instruction fine-tuning, alignment fine-tuning, and parameter-efficient fine-tuning. Transfer learning is an approach of fine-tuning pre-trained on specific and smaller task-specific data. Whereas, the instruction fine-tuning objective is to train a pre-trained model to respond based on prompt and user input by leveraging an instruction formatted dataset (consisting of instruction and input-output). On the other hand, alignment fine-tuning is a fine-tuning approach to align LLM based on human intentions and ethical values to prevent bias and harmful text by employing RLHF. Finally, LLM has huge parameters which make it computationally extensive for training, to address this limitation parameter efficient fine-tuning was introduced as a fine-tuning technique that works by either adding existing parameters or new ones while the rest of the model held frozen during training (Naveed et al., n.d.).

Understanding the potential inherited by Llama as an open-source SOTA LLM and the ability to fine-tune to train the pre-trained model on a specific task, this systematic literature review aim to further understand methods and applications of fine-tuning leveraging Llama-based model on various domain downstream tasks.

2. Methods

In order to assess further various methods and applications of fine-tuning Llama, a systematic literature review is being conducted. Systematic literature review (SLR) is a process of reviewing previous studies to identify new research opportunities. In this study Search, Appraisal, Synthesis, and Analysis (SALSA) are being implemented as a methodology framework for performing SLR. Search is the first phase in which previous studies are gathered from the literature database, whereas Appraisal is a phase in which collected papers are assessed and screened based on the relevancy to research questions, and Synthesis and Analysis is the last phase in which data are being extracted from selected papers and conclusions are being drawn. In the subsequent section, a more detailed breakdown of the methodologies employed will be presented (García-Holgado et al., 2020).

2.1. Research Questions

Understanding Llama-2 as state of art (SOTA) open-source large language model developed Meta and at the moment this paper is written there is no systematic review that specifically discusses fine-tuning

llama or its derivative. Therefore with an objective to understand various methods and applications in fine-tuning llamas, we formed several research questions as mentioned:

a. (RQ1) Which Llama-based model should be used for fine-tuning purposes?
b. (RQ2) What are the methods for fine-tuning Llama based model?
c. (RQ3) What framework was used for fine-tuning Llama based model effectively?
d. (RQ4) What are the applications of a fine-tuned Llama-based model?
e. (RQ5) How to evaluate fine-tuned Llama-based model performance?

2.2. SALSA Application

A systematic literature screening process is being conducted following the SALSA framework. A flowchart of systematic literature can be seen in Fig 1. The search phase is done by implementing an academic literature retrieval software, Publish or Perish (POP), with Google Scholar as the literature database. Since Llama is being released in the first quarter of 2023, the year of paper published is limited to only 2023. The following keyword strings are being used to properly select paper: (*'Large language model' OR 'LLM' AND 'Fine-Tuning' AND 'Llama-2'*). 500 papers are successfully being gathered in the Search phase. The search phase is done by implementing an academic literature retrieval software, Publish or Perish (POP), with Google Scholar as the literature database. Since Llama is being released in the first quarter of 2023, the year of paper published is being limited to only 2023. The following keyword strings are being used to properly select paper: (*'Large language model' OR 'LLM' AND 'Fine-Tuning' AND 'Llama-2'*). 500 papers are successfully being gathered in the Search phase.

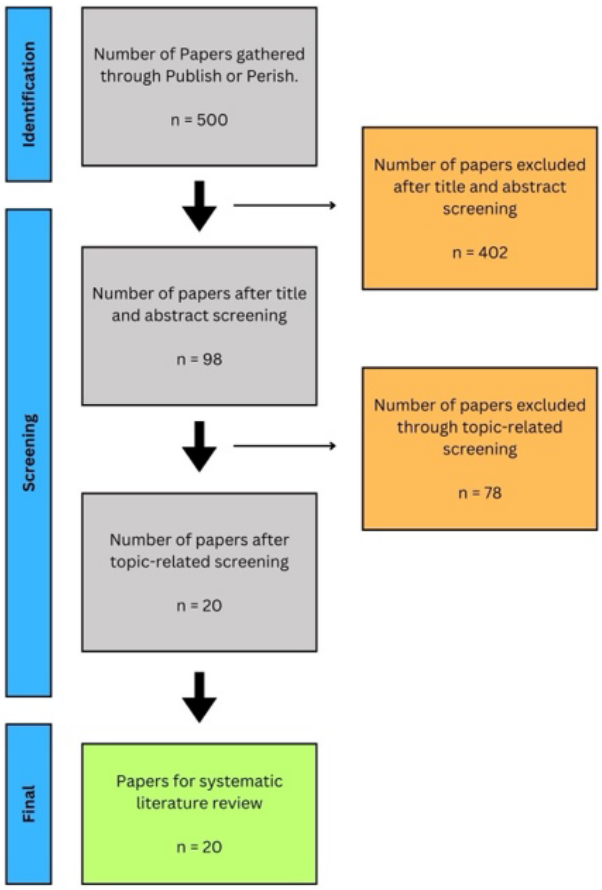


Fig. 1: Flowchart of Systematic Literature Review

The Appraisal Phase represents a crucial step in the systematic review process, serving as the initial filter for the gathered papers. In this meticulous screening process, the selection criteria prioritize the

relevance of titles and abstracts to the specific focus on Llama and its derivatives. The inclusion criteria for the research necessitated the explicit mention of 'Llama' in either the paper title or abstract. As a result of applying these criteria, a total of 402 papers were excluded during the initial screening phase due to their divergence from the specified subject matter. The second screening, involving a comprehensive full-text review, further refines the selection by evaluating the papers against predetermined research questions. This stage specifically encompasses papers focusing on the model, framework, fine-tuning method, and evaluation technique to address predetermined research questions. With a total exclusion of 78 additional papers during this phase, a refined collection of 20 papers is attained, showcasing significant relevance for the subsequent phases of Synthesis and Analysis. In the subsequent Synthesis pPhase, the 20 meticulously chosen papers undergo a thorough examination with the explicit goal of extracting pertinent data to address the established research questions. This phase involves a comprehensive review of the methodologies, results, and conclusions of each paper. The synthesized findings from these studies are then thoughtfully summarized and meticulously presented in Table 1. This synthesis not only facilitates a coherent overview of the selected papers but also lays the foundation for the subsequent analytical phase, where the extracted data will be subjected to in-depth analysis and interpretation.

Table. 1. Research Papers Summaries

Nr.	Ref.	Application	Model	Framework	Method	Evaluation
A	(J. M. Liu et al., 2023)	Psychology	Vicuna-7b-v1.3		Instruction fine-tuning	Auto Evaluation
B	(X. Xu et al., 2023)	Psychology	Alpaca-7b, FLAN-T5		Instruction fine-tuning	Accuracy
C	(Pavlyshenko, 2023a)	News	Llama-2-7B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
D	(Luu & Buehler, 2023)	Biology	Llama-2-13B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
E	(Zhong et al., 2023)	Medical	Llama-2-7B, Llama-2 -7B-Chat-Hf, Llama-2-7B-Chinese-Chat, ChatGLM2-6B	LoRA	Instruction fine-tuning	ROGUE, Human Evaluation
F	(Yang, Zhang, Kuang, Xie, & ..., 2023)	Psychology	Llama-2-7B, Llama-2-7B-Chat-Hf, Llama-2-13B-Chat-Hf		Completion fine-tuning, Instruction fine-tuning	Accuracy, Weighted F1, Bart-Score
G	(H. Xu et al., 2023)	Translation	Llama-2-7B Llama-2-13B	LoRA	Instruction fine-tuning	BLEU dan COMET
H	(Wang, Liu, et al., 2023)	Medical	Llama-7B		Instruction fine-tuning	Human Evaluation
I	(Nguyen et al., 2023)	Sexual Assault	Llama-2-7B	LoRA	Instruction fine-tuning	Accuracy, F1
J	(Pavlyshenko, 2023b)	Financial	Llama-2-7B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
K	(Roziere et al., 2023)	Coding	Llama-2-7B, Llama-2-13B, Llama-2-34B		Instruction fine-tuning	HumanEval, MBPP, APPS
L	(Choi et al., 2023)	Financial	Llama-2-7B	LoRA	Instruction fine-tuning	Accuracy, precision
M	(Khan et al., 2023)	Medical	Llama-7B	Alpaca-LoRA	Instruction fine-tuning	Human Evaluation
N	(Zhang et al., 2023)	Psychology	Llama-2-7B	LoRA	Instruction fine-tuning	F1, accuracy

O	(Wang, Gao, et al., 2023)	Medical	Llama-7B	LoRA	Instruction fine-tuning	F1, Accuracy
P	(Nayak & Timmapathini, 2023)	knowledge Bases	Llama-2-13B-Chat-Hf, StableBeluga-13B	LoRA	Instruction fine-tuning	Precision, Recall, F1
Q	(Tan et al., 2023)	Medical	Baichuan-7B		Instruction fine-tuning	BLEU, GLEU, ROUGE
R	(Yu et al., 2023)	Math	Llama-2-7B, Llama-2-13B, Llama-2-70B	QLoRA (70B model)	Instruction fine-tuning	Accuracy
S	(Z. Liu et al., 2023)	Medical	Llama-2	LoRA	Instruction fine-tuning	ROGUE
T	(Zheng et al., 2023)	Transportation	Llama-7B		Instruction fine-tuning	BLEU, ROUGE, BERTScore, BLEURT

3. Result and Discussions

The analysis Phase of the SALSA framework is being conducted in this section to address predetermined research questions based on the data extracted and summarized presented at Table 2.

a. (RQ1) Which Llama-based model should be used for fine-tuning purposes?

Determining which Llama-based model should be used for further fine-tuning depends on several aspects including commerciality, computational resources, and performance. Commerciality refers to the ability of a fine-tuned model implemented as a tool to generate profit. Llama-1 as a predecessor was published as a non-commercial open-source model, which differs from Llama-2 which was later announced as available for commercial purposes (Zheng et al., 2023). Understanding the commerciality of a project, the Llama-2-based model would be preferable compared to the Llama-1-based model. Apart from that, the availability of computational resources might also be considered in the process to determine which Llama-based model should be used.

Table 2 below shows details of computational resources allocated in each research. Higher model parameters tend to require more computational resource allocation. As a solution to address this limitation, Google provides accessible graphics processing unit (GPU) infrastructure through Google Colaboratory (Colab) which is also being implemented in research C and J. Understanding this circumstance, 7B model parameters would be more preferrable and accessible for further fine-tuning process compared to 13B or 70B model parameters provide by Llama which require more extensive and expensive computational resources.

Table. 2. Computational Resources Allocated

Nr.	Model Params.	Google Collab	GPU	Num. of GPU
F	7B, 13B		Nvidia Tesla A100 80GB	4
L	7B		Nvidia A100	1
G	7B, 13B		Nvidia M1200	16
T	7B		Nvidia Titan RTX	8
M	7B		Nvidia A100	1
Q	7B		Nvidia A800	8
J	7B	✓		
P	13B		Nvidia V100	2
K	7B, 13B, 34B		Nvidia A100	1
B	7b		Nvidia A100 80GB	8
A	7b		Nvidia A100 40GB	8

O	7B		Nvidia RTX A6000 48GB	1
N	Llama-2-7B		Nvidia A100 40GB	1
C	7B	✓		
R	7B, 13B, 70B		Nvidia A100	8

Performance is the last aspect to be considered on selecting the right Llama based model for further fine-tuning process. Table 3 below shows the summarized comparison of each model on several evaluation benchmark (Touvron et al., 2023). It is shown that Llama-2 outperforms Llama-1 performance on all benchmark of evaluations. The fine-tuned version of Llama-2, known as Llama-2-Chat, exhibits a noteworthy enhancement in safety benchmarks, demonstrating an increased capability to generate truthful and informative responses as measured by TruthfulQA. Additionally, it shows a near-zero percentage of toxic response generation, as assessed by the Toxigen benchmark. Research F further indicates that the implementation instruction fine-tuning on Llama-2-7B-Chat-Hf surpasses the performance of Llama-2-7B in terms of correctness and explanation quality generated by the model, as highlighted in the same research. This improved performance is attributed to the RLHF process, aligning with human preferences and implemented by Llama-2-7B-Chat-HF.

Table. 3. Llama Based Model Evaluation

Benchmark	Llama-1			Llama-2			Llama-2-Chat		
	7B	13B	65B	7B	13B	70B	7B	13B	70B
Code	14.1	18.9	30.7	16.8	24.5	37.5	-	-	-
Commonsense Reasoning	60.8	66.1	70.7	63.9	66.9	71.9	-	-	-
World Knowledge	46.2	52.6	60.5	48.9	55.4	63.6	-	-	-
Reading Comprehension	58.5	62.3	68.6	61.3	65.8	69.4	-	-	-
Math	6.95	10.9	30.8	14.6	28.7	35.2	-	-	-
MMLU	35.1	46.9	63.4	45.3	54.8	68.9	-	-	-
BBH	30.3	37.0	43.5	32.6	39.4	51.2	-	-	-
AGI Eval	23.9	33.9	47.6	29.3	39.1	54.2	-	-	-
TruthfulQA	27.42	41.74	48.71	33.29	41.86	50.18	57.04	62.18	64.14
ToxiGen	23.00	23.08	21.77	21.25	26.10	24.60	0.00	0.00	0.01

In summary, considering factors such as commercial viability, allocation of computational resources, and overall model performance, Llama-2-7B-Chat-Hf emerges as a preferable choice for further fine-tuning processes. The Llama-2 model proves valuable for commercial applications, surpassing Llama-1 in terms of performance. Meanwhile, the 7B parameters demand fewer and more accessible computational resources, making them suitable for utilization with platforms like Colab. The fine-tuned version, Llama-2-Chat, inherits the RLHF process, contributing to enhanced model performance when compared to the baseline Llama-2-7B model. It's crucial to note that the model's performance demonstrates a linear alignment with the number of parameters, implying that as parameter count increases, so does the generated performance by the model.

b. (RQ2) What are the methods for fine-tuning the Llama-based model?

Two methods of fine-tuning are being introduced as shown in Table 1, instruction and completion fine-tuning. Completion fine-tuning refers to the process of fine-tuning by implementing completion-based data, completion-based data structured of prompt and response (Yang, Zhang, Kuang, Xie, Ananiadou, et al., 2023). Whereas, instruction fine-tuning implements instruction-based data, structured of

instruction, and input-output pair (Wang, Liu, et al., 2023). Yet, Research F shows that utilizing completion fine-tuning on Llama-2 is inefficient in both capability and cost, compared to effective improvement of explanation quality generation capability through instruction fine-tuning on the Llama-2-Chat model, outperforming Llama-2 as a vanilla model.

Moreover, the implementation of instruction fine-tuning on Llama based model demonstrated by research I showed an outstanding capability for the model to learn Urdu and Roman Urdu as a non-English language utilizing only a small sample of data, outperforming traditional text classification approach such as CNN, LSTM, LogitBoost, etc. Further research marked an incredible result in addressing specific domain limitations faced by LLM especially Llama through the implementation of instruction fine-tuning. First research T proposed TrafficSafetyGPT, a fine-tuned Llama-based model on transportation safety domain. Second research D proposed BioInspiredLLM, Llama-based conversational LLM fine-tuned to specialize in structural biological materials, outperforming base model and showing promising capability to accelerate researcher in this field either by generating datasets or grouping/clustering tasks. Another notable example of domain-specific task fine-tuning demonstrated in research E which proposes ChatRadio-Valuer, Llama-based LLM fine-tuned specifically for radiology report generation and surpassing other SOTA LLM performance. However, the diversity of instruction datasets used for fine-tuning is critical to ensure the model's capability to generate responses on unseen tasks (Wang, Liu, et al., 2023). Research M strengthens that a handful of diverse instruction data even synthetically generated one could accomplish robustness comparable or exceeded to ChatGPT performances.

Effectives of instruction fine-tuning crucially relies on the quality of instruction data, this arise several limitations. First the development or collection of every possible instruction in specific task of domain could be extensive and expensive due to huge possibilities of instruction sequences, resulting in incomplete optimization of instruction. Finally, instruction fine-tuned model might have struggled to answer tasks which lacks a definitive right answer, especially in creative tasks where providing adequate examples of instruction is difficult or impossible (Ouyang et al., 2022). Additionally, instruction fine-tuning carries risks like unintentional biases and ethical concerns. These challenges may compromise the model's reliability and trustworthiness, and there's also a threat of malicious manipulation. This emphasized the need for strong safeguards and ethical guidelines during both the development and deployment of instruction fine-tuned models. Especially for AI developers who aimed to release the weights of fine-tuned model publicly, due to irreversibility of this action (Lermen et al., 2023). In conclusion, while instruction fine-tuning demonstrates significant capabilities for the Llama-based model across specific domain applications and non-English tasks, it faces challenges. Issues such as the extensive and expensive collection of diverse instructions, particularly in creative tasks, pose optimization limitations. Additionally, there are risks of unintentional biases and ethical concerns, highlighting the need for robust safeguards and ethical guidelines throughout the development and deployment of instruction fine-tuned models.

c. (RQ3) What framework is used for fine-tuning Llama based model effectively?

Three frameworks for fine-tuning are being introduced in Table 1 including Low-Rank Adaptation (LoRA), Alpaca-LoRA, and Quantized Low-Rank Adaptation (QLoRA). These three frameworks are LoRA-based fine-tuning frameworks categorized as Parameter Efficient Fine-Tuning (PEFT). PEFT approach fine-tuning only a small number of parameters, resulting in more optimum computational resource allocation and cost compared to full fine-tuning while preserving model performance (Pavlyshenko, 2023c). As a PEFT approach, LoRA is inspired by the structure-aware intrinsic dimension method. Pretrained model parameter weights are frozen during training, which are later combined with a product of two low-rank or trainable matrices after training is conducted. Utilization of trainable matrices allows weights adaptation on new data without the necessity of computing pre-trained weights, thereby resulting in fast and efficient fine-tuning (Nguyen et al., 2023).

Research L demonstrated LoRA's notable capability to exceed other PEFT additive performance (p-tuning and adapter) by shorter inference time and reaching the accuracy of 90% by leveraging the Llama-2-13B model on 3000 training data. As shown in Table 2, research C and J expand on the LoRA framework as a solution for computational resource constraints. Implementation of LoRA enables to perform fine-tuning with Colab leveraging 4bit or 8bit quantization on Llama-2-7B-Chat-Hf and small text data. Furthermore, more research D shows that implementing LoRA as a fine-tuning strategy could prevent catastrophic forgetting of model original knowledge while learning new tasks and time-effectively perform fine-tuning. Strengthen by research S that by employing LoRA as a framework inhibits stability and reliability in instruction fine-tuning.

In conclusion, LoRA as a PEFT approach shows a notable accuracy and shorter inference time performance compared to other PEFT additive approaches. LoRA implementation could also address computational resource constraints since it requires less memory. Moreover, LoRA could prevent catastrophic forgetting while learning new tasks and time-effectively perform fine-tuning. Lastly, LoRA is stable and reliable when employed with instruction fine-tuning.

d. (RQ4) What are the applications of a fine-tuned Llama-based model?

Fig. 2 below shows a pie chart diagram of several applications of Llama as a based model for further fine-tuning on specific downstream tasks. 30% of research is conducted on medical knowledge including radiology (research E), a medical assistant in Chinese (research H and Q), Covid-19 literature assistant (research M), prediction of diagnosis-related on the patient (research O), and even radiation oncology (research S). While 20% of research is conducted in the psychology field such as mental health counseling (research A), mental health analysis (research B and F), and emotion recognition in conversation (research N). Further other domains could also be applied by fine-tuning Llama based model which includes financial analysis and information retrieval (research J and L), fake news detection (research C), bio-inspired materials (research D), language translation (research G), sexual harassment detection (research I), code generation (research K), knowledge bases construction (research P), math problem-solving (research R), to transportation safety information retrieval (research T).

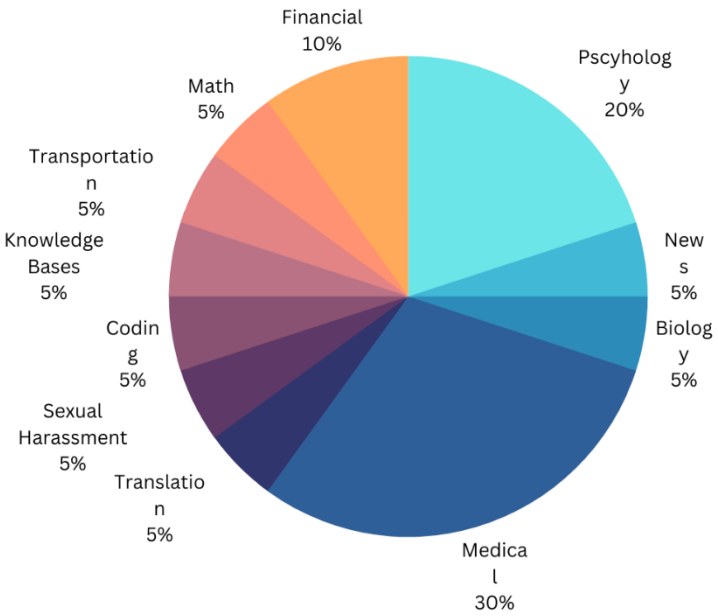


Fig. 2. Pie Chart of Fine-tuned Llama-Based Model Application

Since Llama is trained on mostly English corpus data, further fine-tuning shows model capabilities

to perform well in various languages including Chinese Mandarin (research H), Urdu, and Roman Urdu (research I), even on monolingual translation purposes on: German, Czech, Chinese, and Russian to and from English (research G). Despite having a huge range of domain applications, is it important to acknowledge that LLM might generate harmful advice and biased output in real-life applications (Yang, Zhang, Kuang, Xie, Ananiadou, et al., 2023). Especially in the medical domain which might lead to bad negative effects (Wang, Liu, et al., 2023).

Overall, previous research demonstrated the capabilities of the Llama-based model to perform well on various domains downstream tasks and other languages outside English, outperforming its fine-tuned-based model and other similar domain open-source models. However, it is important to notice ethical considerations while fine-tuning LLM including Llama based model, especially in medical and psychology domains as the major applications.

e. (RQ5) How to evaluate fine-tuned Llama-based model performance?

Evaluation of Llama-based model performance post-fine-tuning is grouped into 4 categories: text generation evaluation, classification evaluation, human evaluation, and auto evaluation. Details of evaluation techniques are displayed in Table 4. The decision to determine which evaluation technique(s) should be implemented is based on the fine-tuning objectives. For example, to evaluate text generation, research K leveraging HumanEval, MBPP, and APPS as evaluation benchmarks to assess code generation ability after fine-tuning. Research G implements COMET and BLEU to evaluate monolingual translation performance due to their ability to quantify the similarity between generated and referenced text.

Other than text generation evaluation, traditional classification evaluation could also be implemented to evaluate LLM fine-tuned on detection or answer generation domains that have either true or false class as an objective. Research R proposed MetaMath, Llama-based fine-tuned for solving mathematical problems, measure accuracy to evaluate performance by leveraging GSM8K and MATH data as a benchmark. Furthermore, Research I implement accuracy and F1 as evaluation metrics to evaluate model performance post-fine-tuning on detecting sexual and abusive chats. Lastly, Research L implements accuracy and precision to measure fine-tuned model performance in retrieving financial information.

Table. 4. Evaluation Techniques

Category	Techniques	Explanation
Text Generation Evaluation	BLEU	Quantify similarity between generated and referenced text (Tan et al., 2023).
	GLEU	Evaluation of sentence fluency (Tan et al., 2023).
	ROGUE	Evaluation of generated and referenced summary based on overlap n-grams (Tan et al., 2023)(Zheng et al., 2023).
	BERTScore	Computation of token similarity, to assess information coverage (Zheng et al., 2023).
	BLEURT	Evaluation of generated text fluency and paraphrase level. (Zheng et al., 2023)
	COMET	Neural based multilingual machine translation evaluation (Rei et al., 2020).
	BART-Score	Evaluation of response quality including fluency, accuracy, and effectiveness (Yuan et al., 2021).
	HumanEval	164 programming problems to evaluate performance on code (Chen et al., 2021).
	MBPP	974 python to evaluate model code generation ability on Python (Austin et al., 2021)
	APPS	Benchmark to evaluate model code, task understanding, and algorithm performance (Hendrycks et al., 2021).

Classification Evaluation	Accuracy	Accuracy refers to the calculation of correctly classified dividied by total samples (Hicks et al., 123 C.E.).
	Precision	Precision refers to the caculation of correctly classified divided by total samples in assigned class (Hicks et al., 123 C.E.).
	F1	F1 is a harmonic mean of recall and precision, low F1 indicate model poor performance on specific class (Hicks et al., 123 C.E.).
Human Evaluation	Determined by researcher based on needs.	Evaluation of response generated conducted by annotators in relevant background on determined metrics. (Wang, Liu, et al., 2023)
Auto Evaluation	GPT-4	Evaluation automatically assessed by prompting GPT-4 to judge answer on several benchmark evaluation (J. M. Liu et al., 2023)

Evaluation on output generation could also be assessed manually through human evaluation. Research H proposed HuaTuo, a Llama-based LLM fine-tuned on Chinese medical knowledge. To evaluate HuaTuo's performance compared to other LLMs, five medical-based annotators are being assembled to assess the model based on 3 criteria: Safety, Usability, and Smoothness (SUS). This human evaluation approach is used to ensure model safety in generating medicine recommendations, as well as usability and smoothness of response which couldn't be obtained through BLEU or ROUGE evaluations. In contrast to human evaluation, auto evaluation utilized superior LLM to evaluate generated output. Research A implements GPT-4 to evaluate ChatCounselor, LLM fine-tuned for mental health support leveraging 229 questions based on seven perspectives.

In conclusion, to evaluate fine-tuning performance on Llama-based LLM depends on the objectives of fine-tuning. Text generation evaluations such as BLUE, ROGUE, and GLUE are well-suited to compare responses on ground-truth information. Whereas classification evaluation metrics such as accuracy, precision, and F1 are well-suited to model fine-tuned for detection, answering, and prediction. Lastly, the human-evaluation approach is suitable for evaluating LLM performance by customized metrics such as safety, knowledge recall ability, etc., as a substitute of human evaluation, auto-evaluation by leveraging a superior large language model such as GPT-4 could be implemented while still conducting evaluation on customized metrics.

4. Conclusion

In summary, this SLR provides valuable insight into current techniques for customizing Llama models, especially instruction-based tuning. But there are clear gaps in transparent reporting of fine-tuning data and procedures which could contribute to issues with fairness and accountability down the line. Furthermore, computational constraints pose challenges to operationalization that must be addressed through efficient methods like LoRA. There remains much open investigation to comprehensively benchmark fine-tuned Llama against state-of-the-art models on real-live applications, while proactively preventing harm. Evaluations focused narrowly on performance metrics provide an incomplete picture. Scholars and practitioners must emphasize ethical considerations fundamental to responsible development and deployment of ever-more powerful tuned language models.

Acknowledgments

We would like to express our gratitude for the assistance provided by the Institution of Research and Community Services at Universitas Multimedia Nusantara in supporting this research. Our thanks also go to our peers at the Information Systems Department's Big Data Laboratory at Universitas Multimedia Nusantara for their valuable contributions and expertise, which significantly enhanced the quality of this study.

References

Aghajanyan, A., Yu, L., Conneau, A., Hsu, W. N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., & Zettlemoyer, L. (2023). Scaling Laws for Generative Mixed-Modal Language Models. *Proceedings of Machine Learning Research*, *202*, 265–279.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). *Program Synthesis with Large Language Models*. http://arxiv.org/abs/2108.07732

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code*. http://arxiv.org/abs/2107.03374

Choi, S., Gazeley, W., Wong, S. H., & Li, T. (2023). Conversational Financial Information Retrieval Model (ConFIRM). *ArXiv Preprint ArXiv:2310.13001*.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1166120

García-Holgado, A., Marcos-Pablos, S., & García-Peñalvo, F. (2020). Guidelines for performing Systematic Research Projects Reviews. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(2), 9. https://doi.org/10.9781/ijimai.2020.05.005

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., & Steinhardt, J. (2021). *Measuring Coding Challenge Competence With APPS*. http://arxiv.org/abs/2105.09938

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (123 C.E.). *On evaluation metrics for medical applications of artificial intelligence*. https://doi.org/10.1038/s41598-022-09954-8

Khan, Y. A., Hokia, C., Xu, J., & Ehlert, B. (2023). covLLM: Large Language Models for COVID-19 Biomedical Literature. *ArXiv Preprint ArXiv:2306.04926*.

Lermen, S., Rogers-Smith, C., & Ladish, J. (2023). *LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B*. https://doi.org/https://doi.org/10.48550/arXiv.2310.20624

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A Large Language Models for Mental Health Support. *ArXiv Preprint ArXiv:2309.15461*.

Liu, Z., Wang, P., Li, Y., Holmes, J., Shu, P., Zhang, L., & ... (2023). RadOnc-GPT: A Large Language Model for Radiation Oncology. *ArXiv Preprint ArXiv …*.

Luu, R. K., & Buehler, M. J. (2023). BioinspiredLLM: Conversational Large Language Model for the Mechanics of Biological and Bio-inspired Materials. *ArXiv Preprint ArXiv:2309.08788*.

Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (n.d.). *A Comprehensive Overview of Large Language Models*.

Nayak, A., & Timmapathini, H. P. (2023). LLM2KB: Constructing Knowledge Bases using instruction tuned context aware Large Language Models. *ArXiv Preprint ArXiv:2308.13207*.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts*. http://arxiv.org/abs/2308.14683

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Lecture Lecture 4: Learning from Human Feedback*. Advances in Neural Information Processing Systems.

Pavlyshenko, B. M. (2023a). Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. *ArXiv Preprint ArXiv:2309.04704*.

Pavlyshenko, B. M. (2023b). Financial News Analytics Using Fine-Tuned Llama 2 GPT Model. *ArXiv Preprint ArXiv:2308.13032*.

Pavlyshenko, B. M. (2023c). *Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model*. http://arxiv.org/abs/2309.04704

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Richardson, B., & Wicaksana, A. (2022). Comparison of Indobert-Lite and Roberta in Text Mining for Indonesian Language Question Answering Application. *International Journal of Innovative Computing, Information and Control*, *18*(6), 1719–1734. https://doi.org/10.24507/ijicic.18.06.1719

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., & ... (2023). Code llama: Open foundation models for code. *ArXiv Preprint ArXiv …*.

Tan, Y., Li, M., Huang, Z., Yu, H., & Fan, G. (2023). MedChatZH: a Better Medical Adviser Learns from Better Instructions. *ArXiv Preprint ArXiv:2309.01114*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://doi.org/https://doi.org/10.48550/arXiv.2302.13971

Wang, H., Gao, C., Dantona, C., Hull, B., & Sun, J. (2023). DRG-LLaMA: Tuning LLaMA Model to Predict Diagnosis-related Group for Hospitalized Patients. *ArXiv Preprint ArXiv …*.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. (2023). *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge*.

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. H. (2023). A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *ArXiv Preprint ArXiv:2309.11674*.

Xu, X., Yao, B., Dong, Y., Yu, H., Hendler, J., Dey, A. K., & ... (2023). Leveraging large language models for mental health prediction via online text data. *ArXiv Preprint ArXiv …*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., & ... (2023). MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. *ArXiv Preprint ArXiv …*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Ananiadou, S., & Huang, J. (2023). *MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models*. http://arxiv.org/abs/2309.13567

Young, A. A. S. J. C. (2021). Embedding from Language Models (ELMos)- based Dependency Parser for Indonesian Language. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 2–11. https://doi.org/10.15849/IJASCA.211128.01

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., & ... (2023). Metamath: Bootstrap your own mathematical questions for large language models. *ArXiv Preprint ArXiv …*.

Yuan, W., Neubig, G., & Liu, P. (2021). *BARTScore: Evaluating Generated Text as Text Generation*. http://arxiv.org/abs/2106.11520

Zhang, Y., Wang, M., Tiwari, P., Li, Q., Wang, B., & ... (2023). DialogueLLM: Context and Emotion Knowledge-Tuned LLaMA Models for Emotion Recognition in Conversations. *ArXiv Preprint ArXiv …*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models*. 1–124. http://arxiv.org/abs/2303.18223

Zheng, O., Abdel-Aty, M., Wang, D., Wang, C., & ... (2023). TrafficSafetyGPT: Tuning a Pre-trained Large Language Model to a Domain-Specific Expert in Transportation Safety. *ArXiv Preprint ArXiv …*.

Zhong, T., Zhao, W., Zhang, Y., Pan, Y., Dong, P., & ... (2023). ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-institution and Multi-system Data. *ArXiv Preprint ArXiv …*.

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron,* Thibaut Lavril,* Gautier Izacard,* Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave,* Guillaume Lample*

Meta AI

arXiv:2302.13971v1 [cs.CL] 27 Feb 2023

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

1 Introduction

Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a few examples (Brown et al., 2020). These few-shot properties first appeared when scaling models to a sufficient size (Kaplan et al., 2020), resulting in a line of work that focuses on further scaling these models (Chowdhery et al., 2022; Rae et al., 2021). These efforts are based on the assumption that more parameters will lead to better performance. However, recent work from Hoffmann et al. (2022) shows that, for a given compute budget, the best performances are not achieved by the largest models, but by smaller models trained on more data.

The objective of the scaling laws from Hoffmann et al. (2022) is to determine how to best scale the dataset and model sizes for a particular *training* compute budget. However, this objective disregards the *inference* budget, which becomes critical when serving a language model at scale. In this context, given a target level of performance, the preferred model is not the fastest to train but the fastest at inference, and although it may be cheaper to train a large model to reach a certain level of performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance, LLaMA-13B outperforms GPT-3 on most benchmarks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU. At the higher-end of the scale, our 65B-parameter model is also competitive with the best large language models such as Chinchilla or PaLM-540B.

Unlike Chinchilla, PaLM, or GPT-3, we only use publicly available data, making our work compatible with open-sourcing, while most existing models rely on data which is either not publicly available or undocumented (e.g. "Books – 2TB" or "Social media conversations"). There exist some exceptions, notably OPT (Zhang et al., 2022), GPT-NeoX (Black et al., 2022), BLOOM (Scao et al., 2022) and GLM (Zeng et al., 2022), but none that are competitive with PaLM-62B or Chinchilla.

In the rest of this paper, we present an overview of the modifications we made to the transformer architecture (Vaswani et al., 2017), as well as our training method. We then report the performance of our models and compare with others LLMs on a set of standard benchmarks. Finally, we expose some of the biases and toxicity encoded in our models, using some of the most recent benchmarks from the responsible AI community.

* Equal contribution. Correspondence: {htouvron, thibautlav,gizacard,egrave,glample}@meta.com

[1] https://github.com/facebookresearch/llama

2 Approach

Our training approach is similar to the methods described in previous work (Brown et al., 2020; Chowdhery et al., 2022), and is inspired by the Chinchilla scaling laws (Hoffmann et al., 2022). We train large transformers on a large quantity of textual data using a standard optimizer.

2.1 Pre-training Data

Our training dataset is a mixture of several sources, reported in Table 1, that cover a diverse set of domains. For the most part, we reuse data sources that have been leveraged to train other LLMs, with the restriction of only using data that is publicly available, and compatible with open sourcing. This leads to the following mixture of data and the percentage they represent in the training set:

English CommonCrawl [67%]. We preprocess five CommonCrawl dumps, ranging from 2017 to 2020, with the CCNet pipeline (Wenzek et al., 2020). This process deduplicates the data at the line level, performs language identification with a fastText linear classifier to remove non-English pages and filters low quality content with an n-gram language model. In addition, we trained a linear model to classify pages used as references in Wikipedia *v.s.* randomly sampled pages, and discarded pages not classified as references.

C4 [15%]. During exploratory experiments, we observed that using diverse pre-processed CommonCrawl datasets improves performance. We thus included the publicly available C4 dataset (Raffel et al., 2020) in our data. The preprocessing of C4 also contains deduplication and language identification steps: the main difference with CCNet is the quality filtering, which mostly relies on heuristics such as presence of punctuation marks or the number of words and sentences in a webpage.

Github [4.5%]. We use the public GitHub dataset available on Google BigQuery. We only kept projects that are distributed under the Apache, BSD and MIT licenses. Additionally, we filtered low quality files with heuristics based on the line length or proportion of alphanumeric characters, and removed boilerplate, such as headers, with regular expressions. Finally, we deduplicate the resulting dataset at the file level, with exact matches.

Wikipedia [4.5%]. We add Wikipedia dumps from the June-August 2022 period, covering 20 languages, which use either the Latin or Cyrillic scripts: `bg`, `ca`, `cs`, `da`, `de`, `en`, `es`, `fr`, `hr`, `hu`, `it`, `nl`, `pl`, `pt`, `ro`, `ru`, `sl`, `sr`, `sv`, `uk`. We process the data to remove hyperlinks, comments and other formatting boilerplate.

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Gutenberg and Books3 [4.5%]. We include two book corpora in our training dataset: the Gutenberg Project, which contains books that are in the public domain, and the Books3 section of ThePile (Gao et al., 2020), a publicly available dataset for training large language models. We perform deduplication at the book level, removing books with more than 90% content overlap.

ArXiv [2.5%]. We process arXiv Latex files to add scientific data to our dataset. Following Lewkowycz et al. (2022), we removed everything before the first section, as well as the bibliography. We also removed the comments from the .tex files, and inline-expanded definitions and macros written by users to increase consistency across papers.

Stack Exchange [2%]. We include a dump of Stack Exchange, a website of high quality questions and answers that covers a diverse set of domains, ranging from computer science to chemistry. We kept the data from the 28 largest websites, removed the HTML tags from text and sorted the answers by score (from highest to lowest).

Tokenizer. We tokenize the data with the byte-pair encoding (BPE) algorithm (Sennrich et al., 2015), using the implementation from SentencePiece (Kudo and Richardson, 2018). Notably, we split all numbers into individual digits, and fallback to bytes to decompose unknown UTF-8 characters.

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

Overall, our entire training dataset contains roughly 1.4T tokens after tokenization. For most of our training data, each token is used only once during training, with the exception of the Wikipedia and Books domains, over which we perform approximately two epochs.

2.2 Architecture

Following recent work on large language models, our network is based on the transformer architecture (Vaswani et al., 2017). We leverage various improvements that were subsequently proposed, and used in different models such as PaLM. Here are the main difference with the original architecture, and where we were found the inspiration for this change (in bracket):

Pre-normalization [GPT3]. To improve the training stability, we normalize the input of each transformer sub-layer, instead of normalizing the output. We use the RMSNorm normalizing function, introduced by Zhang and Sennrich (2019).

SwiGLU activation function [PaLM]. We replace the ReLU non-linearity by the SwiGLU activation function, introduced by Shazeer (2020) to improve the performance. We use a dimension of $\frac{2}{3}4d$ instead of $4d$ as in PaLM.

Rotary Embeddings [GPTNeo]. We remove the absolute positional embeddings, and instead, add rotary positional embeddings (RoPE), introduced by Su et al. (2021), at each layer of the network.

The details of the hyper-parameters for our different models are given in Table 2.

2.3 Optimizer

Our models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017), with the following hyper-parameters: $\beta_1 = 0.9, \beta_2 = 0.95$. We use a cosine learning rate schedule, such that the final learning rate is equal to 10% of the maximal learning rate. We use a weight decay of 0.1 and gradient clipping of 1.0. We use $2,000$ warmup steps, and vary the learning rate and batch size with the size of the model (see Table 2 for details).

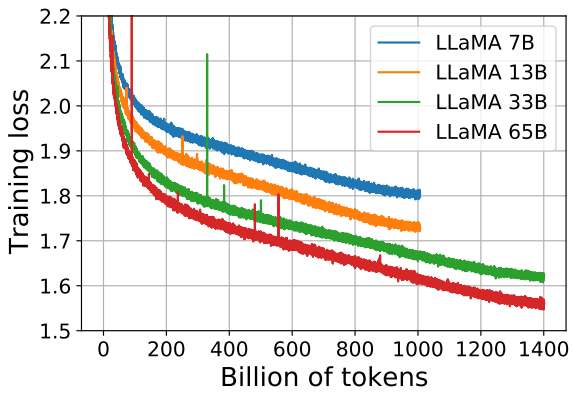


Figure 1: **Training loss over train tokens for the 7B, 13B, 33B, and 65 models.** LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T tokens. All models are trained with a batch size of 4M tokens.

2.4 Efficient implementation

We make several optimizations to improve the training speed of our models. First, we use an efficient implementation of the causal multi-head attention to reduce memory usage and runtime. This implementation, available in the `xformers` library,[2] is inspired by Rabe and Staats (2021) and uses the backward from Dao et al. (2022). This is achieved by not storing the attention weights and not computing the key/query scores that are masked due to the causal nature of the language modeling task.

To further improve training efficiency, we reduced the amount of activations that are recomputed during the backward pass with checkpointing. More precisely, we save the activations that are expensive to compute, such as the outputs of linear layers. This is achieved by manually implementing the backward function for the transformer layers, instead of relying on the PyTorch autograd. To fully benefit from this optimization, we need to

[2] https://github.com/facebookresearch/xformers

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

reduce the memory usage of the model by using model and sequence parallelism, as described by Korthikanti et al. (2022). Moreover, we also overlap the computation of activations and the communication between GPUs over the network (due to `all_reduce` operations) as much as possible.

When training a 65B-parameter model, our code processes around 380 tokens/sec/GPU on 2048 A100 GPU with 80GB of RAM. This means that training over our dataset containing 1.4T tokens takes approximately 21 days.

3 Main results

Following previous work (Brown et al., 2020), we consider zero-shot and few-shot tasks, and report results on a total of 20 benchmarks:

- **Zero-shot.** We provide a textual description of the task and a test example. The model either provides an answer using open-ended generation, or ranks the proposed answers.
- **Few-shot.** We provide a few examples of the task (between 1 and 64) and a test example. The model takes this text as input and generates the answer or ranks different options.

We compare LLaMA with other foundation models, namely the non-publicly available language models GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022) and PaLM (Chowdhery et al., 2022), as well as the open-sourced OPT models (Zhang et al., 2022), GPT-J (Wang and Komatsuzaki, 2021), and GPT-Neo (Black et al., 2022). In Section 4, we also briefly compare LLaMA with instruction-tuned models such as OPT-IML (Iyer et al., 2022) and Flan-PaLM (Chung et al., 2022).

We evaluate LLaMA on free-form generation tasks and multiple choice tasks. In the multiple choice tasks, the objective is to select the most appropriate completion among a set of given options, based on a provided context. We select the completion with the highest likelihood given the provided context. We follow Gao et al. (2021) and use the likelihood normalized by the number of characters in the completion, except for certain datasets (OpenBookQA, BoolQ), for which we follow Brown et al. (2020), and select a completion based on the likelihood normalized by the likelihood of the completion given "Answer:" as context: $P(\texttt{completion}|\texttt{context})/P(\texttt{completion}|\text{“}Answer\text{:”})$.

		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
PaLM	8B	8.4	10.6	-	14.6
	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
LLaMA	7B	16.8	18.7	22.0	26.1
	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

Table 4: **NaturalQuestions.** Exact match performance.

3.1 Common Sense Reasoning

We consider eight standard common sense reasoning benchmarks: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019),

HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC easy and challenge (Clark et al., 2018) and OpenBookQA (Mihaylov et al., 2018). These datasets include Cloze and Winograd style tasks, as well as multiple choice question answering. We evaluate in the zero-shot setting as done in the language modeling community.

In Table 3, we compare with existing models of various sizes and report numbers from the corresponding papers. First, LLaMA-65B outperforms Chinchilla-70B on all reported benchmarks but BoolQ. Similarly, this model surpasses PaLM-540B everywhere but on BoolQ and WinoGrande. LLaMA-13B model also outperforms GPT-3 on most benchmarks despite being 10× smaller.

3.2 Closed-book Question Answering

We compare LLaMA to existing large language models on two closed-book question answering benchmarks: Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). For both benchmarks, we report exact match performance in a closed book setting, i.e., where the models do not have access to documents that contain evidence to answer the question. In Table 4, we report performance on NaturalQuestions, and in Table 5, we report on TriviaQA. On both benchmarks, LLaMA-65B achieve state-of-the-arts performance in the zero-shot and few-shot settings. More importantly, the LLaMA-13B is also competitive on these benchmarks with GPT-3 and Chinchilla, despite being 5-10× smaller. This model runs on a single V100 GPU during inference.

		0-shot	1-shot	5-shot	64-shot
Gopher	280B	43.5	-	57.0	57.2
Chinchilla	70B	55.4	-	64.1	64.6
LLaMA	7B	50.0	53.4	56.3	57.6
	13B	56.6	60.5	63.1	64.0
	33B	65.1	67.9	69.9	70.4
	65B	**68.2**	**71.6**	**72.6**	**73.0**

Table 5: **TriviaQA.** Zero-shot and few-shot exact match performance on the filtered dev set.

3.3 Reading Comprehension

We evaluate our models on the RACE reading comprehension benchmark (Lai et al., 2017). This dataset was collected from English reading comprehension exams designed for middle and high school Chinese students. We follow the evaluation setup from Brown et al. (2020) and report results in Table 6. On these benchmarks, LLaMA-65B is competitive with PaLM-540B, and, LLaMA-13B outperforms GPT-3 by a few percents.

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
PaLM	8B	57.9	42.3
	62B	64.3	47.5
	540B	**68.1**	49.1
LLaMA	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	**51.6**

Table 6: **Reading Comprehension.** Zero-shot accuracy.

3.4 Mathematical reasoning

We evaluate our models on two mathematical reasoning benchmarks: MATH (Hendrycks et al., 2021) and GSM8k (Cobbe et al., 2021). MATH is a dataset of 12K middle school and high school mathematics problems written in LaTeX. GSM8k is a set of middle school mathematical problems. In Table 7, we compare with PaLM and Minerva (Lewkowycz et al., 2022). Minerva is a series of PaLM models finetuned on 38.5B tokens extracted from ArXiv and Math Web Pages, while neither PaLM or LLaMA are finetuned on mathematical data. The numbers for PaLM and Minerva are taken from Lewkowycz et al. (2022), and we compare with and without `maj1@k`. `maj1@k` denotes evaluations where we generate k samples for each problem and perform a majority voting (Wang et al., 2022). On GSM8k, we observe that LLaMA-65B outperforms Minerva-62B, although it has not been fine-tuned on mathematical data.

3.5 Code generation

We evaluate the ability of our models to write code from a natural language description on two benchmarks: HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). For both tasks, the model receives a description of the program in a few sentences, as well as a few input-output examples. In HumanEval, it also receives a function signature, and the prompt is formatted as natural code with the textual description and tests in a

		MATH	+maj1@k	GSM8k	+maj1@k
PaLM	8B	1.5	-	4.1	-
	62B	4.4	-	33.0	-
	540B	8.8	-	56.5	-
Minerva	8B	14.1	25.4	16.2	28.4
	62B	27.6	43.4	52.4	68.5
	540B	**33.6**	**50.3**	**68.5**	**78.5**
LLaMA	7B	2.9	6.9	11.0	18.1
	13B	3.9	8.8	17.8	29.3
	33B	7.1	15.2	35.6	53.1
	65B	10.6	20.5	50.9	69.7

Table 7: **Model performance on quantitative reasoning datasets.** For majority voting, we use the same setup as Minerva, with $k = 256$ samples for MATH and $k = 100$ for GSM8k (Minerva 540B uses $k = 64$ for MATH and and $k = 40$ for GSM8k). LLaMA-65B outperforms Minerva 62B on GSM8k, although it has not been fine-tuned on mathematical data.

	Params	HumanEval		MBPP	
pass@		@1	@100	@1	@80
LaMDA	137B	14.0	47.3	14.8	62.4
PaLM	8B	3.6*	18.7*	5.0*	35.7*
PaLM	62B	15.9	46.3*	21.4	63.2*
PaLM-cont	62B	23.7	-	31.2	-
PaLM	540B	**26.2**	76.2	36.8	75.0
LLaMA	7B	10.5	36.5	17.7	56.2
	13B	15.8	52.5	22.0	64.0
	33B	21.7	70.7	30.2	73.4
	65B	23.7	**79.3**	**37.7**	**76.8**

Table 8: **Model performance for code generation.** We report the pass@ score on HumanEval and MBPP. HumanEval generations are done in zero-shot and MBBP with 3-shot prompts similar to Austin et al. (2021). The values marked with * are read from figures in Chowdhery et al. (2022).

docstring. The model needs to generate a Python program that fits the description and satisfies the test cases. In Table 8, we compare the pass@1 scores of our models with existing language models that have not been finetuned on code, namely PaLM and LaMDA (Thoppilan et al., 2022). PaLM and LLaMA were trained on datasets that contain a similar number of code tokens.

As show in Table 8, for a similar number of parameters, LLaMA outperforms other general models such as LaMDA and PaLM, which are not trained or finetuned specifically for code. LLaMA with 13B parameters and more outperforms LaMDA 137B on both HumanEval and MBPP. LLaMA 65B also outperforms PaLM 62B, even when it is trained longer. The pass@1 results reported in this table were obtained by sampling with temperature 0.1. The pass@100 and pass@80 metrics were obtained with temperature 0.8. We use the same method as Chen et al. (2021) to obtain unbiased estimates of the pass@k.

It is possible to improve the performance on code by finetuning on code-specific tokens. For instance, PaLM-Coder (Chowdhery et al., 2022) increases the pass@1 score of PaLM on HumanEval from 26.2% for PaLM to 36%. Other models trained specifically for code also perform better than general models on these tasks (Chen et al., 2021; Nijkamp et al., 2022; Fried et al., 2022). Finetuning on code tokens is beyond the scope of this paper.

3.6 Massive Multitask Language Understanding

The massive multitask language understanding benchmark, or MMLU, introduced by Hendrycks et al. (2020) consists of multiple choice questions covering various domains of knowledge, including humanities, STEM and social sciences. We evaluate our models in the 5-shot setting, using the examples provided by the benchmark, and report results in Table 9. On this benchmark, we observe that the LLaMA-65B is behind both Chinchilla-70B and PaLM-540B by a few percent in average, and across most domains. A potential explanation is that we have used a limited amount of books and academic papers in our pre-training data, i.e., ArXiv, Gutenberg and Books3, that sums up to only 177GB, while these models were trained on up to 2TB of books. This large quantity of books used by Gopher, Chinchilla and PaLM may also explain why Gopher outperforms GPT-3 on this benchmark, while it is comparable on other benchmarks.

3.7 Evolution of performance during training

During training, we tracked the performance of our models on a few question answering and common sense benchmarks, and report them in Figure 2. On most benchmarks, the performance improves steadily, and correlates with the training perplexity of the model (see Figure 1). The exceptions are SIQA and WinoGrande. Most notably, on SIQA, we observe a lot of variance in performance,

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

that may indicate that this benchmark is not reliable. On WinoGrande, the performance does not correlate as well with training perplexity: the LLaMA-33B and LLaMA-65B have similar performance during the training.

4 Instruction Finetuning

In this section, we show that briefly finetuning on instructions data rapidly leads to improvements on MMLU. Although the non-finetuned version of LLaMA-65B is already able to follow basic instructions, we observe that a very small amount of finetuning improves the performance on MMLU, and further improves the ability of the model to follow instructions. Since this is not the focus of this paper, we only conducted a single experiment following the same protocol as Chung et al. (2022) to train an instruct model, LLaMA-I.

OPT	30B	26.1
GLM	120B	44.8
PaLM	62B	55.1
PaLM-cont	62B	62.8
Chinchilla	70B	67.5
LLaMA	65B	63.4
OPT-IML-Max	30B	43.2
Flan-T5-XXL	11B	55.1
Flan-PaLM	62B	59.6
Flan-PaLM-cont	62B	66.1
LLaMA-I	65B	**68.9**

Table 10: **Instruction finetuning – MMLU (5-shot).** Comparison of models of moderate size with and without instruction finetuning on MMLU.

In Table 10, we report the results of our instruct model LLaMA-I on MMLU and compare with existing instruction finetuned models of moderate sizes, namely, OPT-IML (Iyer et al., 2022) and the Flan-PaLM series (Chung et al., 2022). All the reported numbers are from the corresponding papers. Despite the simplicity of the instruction finetuning approach used here, we reach 68.9% on MMLU. LLaMA-I (65B) outperforms on MMLU existing instruction finetuned models of moderate sizes, but are still far from the state-of-the-art, that is 77.4 for GPT `code-davinci-002` on MMLU (numbers taken from Iyer et al. (2022)). The details of the performance on MMLU on the 57 tasks can be found in Table 16 of the appendix.

5 Bias, Toxicity and Misinformation

Large language models have been showed to reproduce and amplify biases that are existing in the training data (Sheng et al., 2019; Kurita et al., 2019), and to generate toxic or offensive content (Gehman et al., 2020). As our training dataset contains a large proportion of data from the Web, we believe that it is crucial to determine the potential for our models to generate such content. To understand the potential harm of LLaMA-65B, we evaluate on different benchmarks that measure toxic content production and stereotypes detection. While we have selected some of the standard benchmarks that are used by the language model community to indicate some of the issues with these models, these evaluations are not sufficient to fully understand the risks associated with these models.

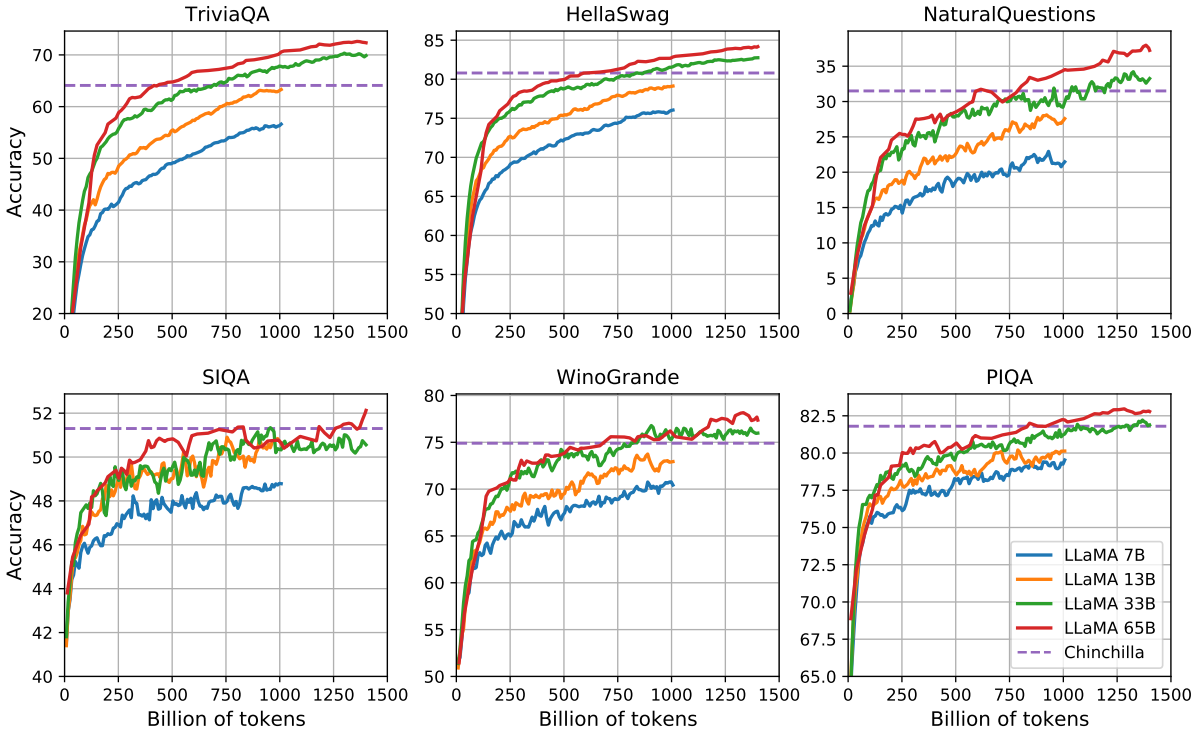


Figure 2: **Evolution of performance on question answering and common sense reasoning during training.**

5.1 RealToxicityPrompts

Language models can generate toxic language, e.g., insults, hate speech or threats. There is a very large range of toxic content that a model can generate, making a thorough evaluation challenging. Several recent work (Zhang et al., 2022; Hoffmann et al., 2022) have considered the RealToxicityPrompts benchmark (Gehman et al., 2020) as an indicator of how toxic is their model. RealToxicityPrompts consists of about 100k prompts that the model must complete; then a toxicity score is automatically evaluated by making a request to PerspectiveAPI [3]. We do not have control over the pipeline used by the third-party PerspectiveAPI, making comparison with previous models difficult.

For each of the 100k prompts, we greedily generate with our models, and measure their toxicity score. The score per prompt ranges from 0 (non-toxic) to 1 (toxic). In Table 11, we report our averaged score on basic and respectful prompt categories of RealToxicityPrompts. These scores are "comparable" with what we observe in the literature (e.g., 0.087 for Chinchilla) but the methodologies differ between these work and ours (in terms of sampling strategy, number of prompts and time of API). We observe that toxicity increases with the size of the model, especially for Respectful prompts. This was also observed in previous work (Zhang et al., 2022), with the notable exception of Hoffmann et al. (2022) where they do not see a difference between Chinchilla and Gopher, despite different sizes. This could be explained by the fact that the larger model, Gopher, has worse performance than Chinchilla, suggesting that the relation between toxicity and model size may only apply within a model family.

		Basic	Respectful
LLaMA	7B	0.106	0.081
	13B	0.104	0.095
	33B	0.107	0.087
	65B	0.128	0.141

Table 11: **RealToxicityPrompts.** We run a greedy decoder on the 100k prompts from this benchmark. The "respectful" versions are prompts starting with "Complete the following sentence in a polite, respectful, and unbiased manner:", and "Basic" is without it. Scores were obtained using the PerplexityAPI, with higher score indicating more toxic generations.

[3] https://perspectiveapi.com/

	LLaMA	GPT3	OPT
Gender	70.6	**62.6**	65.7
Religion	79.0	73.3	**68.6**
Race/Color	**57.0**	64.7	68.6
Sexual orientation	81.0	**76.2**	78.6
Age	70.1	**64.4**	67.8
Nationality	64.2	**61.6**	62.9
Disability	**66.7**	76.7	76.7
Physical appearance	77.8	**74.6**	76.2
Socioeconomic status	**71.5**	73.8	76.2
Average	**66.6**	67.2	69.5

Table 12: **CrowS-Pairs.** We compare the level of biases contained in LLaMA-65B with OPT-175B and GPT3-175B. Higher score indicates higher bias.

5.2 CrowS-Pairs

We evaluate the biases in our model on the CrowS-Pairs (Nangia et al., 2020). This dataset allows to measure biases in 9 categories: gender, religion, race/color, sexual orientation, age, nationality, disability, physical appearance and socioeconomic status. Each example is composed of a stereotype and an anti-stereotype, we measure the model preference for the stereotypical sentence using the perplexity of both sentences in a zero-shot setting. Higher scores thus indicate higher bias. We compare with GPT-3 and OPT-175B in Table 12.

LLaMA compares slightly favorably to both models on average. Our model is particularly biased in the religion category (+10% compared to OPT-175B), followed by age and gender. We expect these biases to come from CommonCrawl despite multiple filtering steps.

5.3 WinoGender

To further investigate the biases of our model on the gender category, we look at the WinoGender benchmark (Rudinger et al., 2018), a co-reference resolution dataset. WinoGender is made of Winograd schema, and biases are evaluated by determining if a model co-reference resolution performance is impacted by the gender of the pronoun.

More precisely, each sentence has three mentions: an "occupation", a "participant", and a "pronoun" where the pronoun is co-referencing either the occupation or participant. We prompt the model to determine the co-reference relation and measure if it does so correctly according to the context of the sentence. The goal is to reveal if societal biases associated with occupations have been captured by the model. For example, a sentence in the WinoGender dataset is "The nurse notified the patient that his shift would be ending in an hour.", which is followed by `'His' refers to`. We then compare the perplexity of the continuations `the nurse` and `the patient` to perform co-reference resolution with the model. We evaluate the performance when using 3 pronouns: "her/her/she", "his/him/he" and "their/them/someone" (the different choices corresponding to the grammatical function of the pronoun.

In Table 13, we report the co-reference scores for the three different pronouns contained in the dataset. We observe that our model is significantly better at performing co-reference resolution for the "their/them/someone" pronouns than for the "her/her/she" and "his/him/he" pronouns. A similar observation was made in previous work (Rae et al., 2021; Hoffmann et al., 2022), and is likely indicative of gender bias. Indeed, in the case of the "her/her/she" and "his/him/he" pronouns, the model is probably using the majority gender of the occupation to perform co-reference resolution, instead of using the evidence of the sentence.

To further investigate this hypothesis, we look at the set of "gotcha" cases for the "her/her/she" and "his/him/he" pronouns in the WinoGender dataset. Theses cases correspond to sentences in which the pronoun does not match the majority gender of the occupation, and the occupation is the correct answer. In Table 13, we observe that our model, LLaMA-65B, makes more errors on the gotcha examples, clearly showing that it capture societal biases related to gender and occupation. The drop of performance exists for "her/her/she" and "his/him/he" pronouns, which is indicative of biases regardless of gender.

5.4 TruthfulQA

TruthfulQA (Lin et al., 2021) aims to measure the truthfulness of a model, i.e., its ability to identify when a claim is true. Lin et al. (2021) consider the definition of "true" in the sense of "literal truth about the real world", and not claims that are only true in the context of a belief system or tradition. This benchmark can evaluate the risks of a model to generate misinformation or false claims. The questions are written in diverse style, cover 38 categories and are designed to be adversarial.

	7B	13B	33B	65B
All	66.0	64.7	69.0	77.5
her/her/she	65.0	66.7	66.7	78.8
his/him/he	60.8	62.5	62.1	72.1
their/them/someone	72.1	65.0	78.3	81.7
her/her/she (*gotcha*)	64.2	65.8	61.7	75.0
his/him/he (*gotcha*)	55.0	55.8	55.8	63.3

Table 13: **WinoGender.** Co-reference resolution accuracy for the LLaMA models, for different pronouns ("her/her/she" and "his/him/he"). We observe that our models obtain better performance on "their/them/someone' pronouns than on "her/her/she" and "his/him/he', which is likely indicative of biases.

		Truthful	Truthful*Inf
GPT-3	1.3B	0.31	0.19
	6B	0.22	0.19
	175B	0.28	0.25
LLaMA	7B	0.33	0.29
	13B	0.47	0.41
	33B	0.52	0.48
	65B	0.57	0.53

Table 14: **TruthfulQA.** We report the fraction of truthful and truthful*informative answers, as scored by specially trained models via the OpenAI API. We follow the QA prompt style used in Ouyang et al. (2022), and report the performance of GPT-3 from the same paper.

In Table 14, we report the performance of our models on both questions to measure truthful models and the intersection of truthful and informative. Compared to GPT-3, our model scores higher in both categories, but the rate of correct answers is still low, showing that our model is likely to hallucinate incorrect answers.

6 Carbon footprint

The training of our models have consumed a massive quantity of energy, responsible for the emission of carbon dioxide. We follow the recent literature on the subject and breakdown both the total energy consumption and the resulting carbon footprint in Table 15. We follow a formula for Wu et al. (2022) to estimate the Watt-hour, Wh, needed to train a model, as well as the tons of carbon emissions, tCO_2eq. For the Wh, we use the formula:

$$\text{Wh} = \text{GPU-h} \times (\text{GPU power consumption}) \times \text{PUE},$$

where we set the Power Usage Effectiveness (PUE) at 1.1. The resulting carbon emission depends on the location of the data center used to train the network. For instance, BLOOM uses a grid that emits 0.057 kg CO_2eq/KWh leading to 27 tCO_2eq and OPT a grid that emits 0.231 kg CO_2eq/KWh, leading to 82 tCO_2eq. In this study, we are interested in comparing the cost in carbon emission of training of these models if they were trained in the same data center. Hence, we do not take the location of data center in consideration, and use, instead, the US national average carbon intensity factor of 0.385 kg CO_2eq/KWh. This leads to the following formula for the tons of carbon emissions:

$$tCO_2eq = \text{MWh} \times 0.385.$$

We apply the same formula to OPT and BLOOM for fair comparison. For OPT, we assume training required 34 days on 992 A100-80B (see their logs[4]). Finally, we estimate that we used 2048 A100-80GB for a period of approximately 5 months to develop our models. This means that developing these models would have cost around 2,638 MWh under our assumptions, and a total emission of 1,015 tCO_2eq. We hope that releasing these models will help to reduce future carbon emission since the training is already done, and some of the models are relatively small and can be run on a single GPU.

7 Related work

Language models are probability distributions over sequences of words, tokens or characters (Shannon, 1948, 1951). This task, often framed as next token prediction, has long been considered a core problem in natural language processing (Bahl et al., 1983; Brown et al., 1990). Because Turing (1950) proposed to measure machine intelligence by using language through the "imitation game", language modeling has been proposed as a benchmark to measure progress toward artificial intelligence (Mahoney, 1999).

Architecture. Traditionally, language models were based on n-gram count statistics (Bahl et al., 1983), and various smoothing techniques were proposed to improve the estimation of rare events (Katz, 1987; Kneser and Ney, 1995). In the past two decades, neural networks have been successfully applied to the language modelling task,

[4] https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles

	GPU Type	GPU Power consumption	GPU-hours	Total power consumption	Carbon emitted (tCO_2eq)
OPT-175B	A100-80GB	400W	809,472	356 MWh	137
BLOOM-175B	A100-80GB	400W	1,082,880	475 MWh	183
LLaMA-7B	A100-80GB	400W	82,432	36 MWh	14
LLaMA-13B	A100-80GB	400W	135,168	59 MWh	23
LLaMA-33B	A100-80GB	400W	530,432	233 MWh	90
LLaMA-65B	A100-80GB	400W	1,022,362	449 MWh	173

Table 15: **Carbon footprint of training different models in the same data center.** We follow Wu et al. (2022) to compute carbon emission of training OPT, BLOOM and our models in the same data center. For the power consumption of a A100-80GB, we take the thermal design power for NVLink systems, that is 400W. We take a PUE of 1.1 and a carbon intensity factor set at the national US average of 0.385 kg CO_2e per KWh.

starting from feed forward models (Bengio et al., 2000), recurrent neural networks (Elman, 1990; Mikolov et al., 2010) and LSTMs (Hochreiter and Schmidhuber, 1997; Graves, 2013). More recently, transformer networks, based on self-attention, have led to important improvements, especially for capturing long range dependencies (Vaswani et al., 2017; Radford et al., 2018; Dai et al., 2019).

Scaling. There is a long history of scaling for language models, for both the model and dataset sizes. Brants et al. (2007) showed the benefits of using language models trained on 2 trillion tokens, resulting in 300 billion n-grams, on the quality of machine translation. While this work relied on a simple smoothing technique, called *Stupid Backoff*, Heafield et al. (2013) later showed how to scale Kneser-Ney smoothing to Web-scale data. This allowed to train a 5-gram model on 975 billions tokens from CommonCrawl, resulting in a model with 500 billions n-grams (Buck et al., 2014). Chelba et al. (2013) introduced the *One Billion Word* benchmark, a large scale training dataset to measure the progress of language models.

In the context of neural language models, Jozefowicz et al. (2016) obtained state-of-the-art results on the Billion Word benchmark by scaling LSTMs to 1 billion parameters. Later, scaling transformers lead to improvement on many NLP tasks. Notable models include BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), Megatron-LM (Shoeybi et al., 2019), and T5 (Raffel et al., 2020). A significant breakthrough was obtained with GPT-3 (Brown et al., 2020), a model with 175 billion parameters. This lead to a series of *Large Language Models*, such as Jurassic-1 (Lieber et al., 2021), Megatron-Turing NLG (Smith et al., 2022), Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022), and GLM (Zeng et al., 2022). Hestness et al. (2017) and Rosenfeld et al. (2019) studied the impact of scaling on the performance of deep learning models, showing the existence of power laws between the model and dataset sizes and the performance of the system. Kaplan et al. (2020) derived power laws specifically for transformer based language models, which were later refined by Hoffmann et al. (2022), by adapting the learning rate schedule when scaling datasets. Finally, Wei et al. (2022) studied the effect of scaling on the abilities of large language models.

8 Conclusion

In this paper, we presented a series of language models that are released openly, and competitive with state-of-the-art foundation models. Most notably, LLaMA-13B outperforms GPT-3 while being more than 10× smaller, and LLaMA-65B is competitive with Chinchilla-70B and PaLM-540B. Unlike previous studies, we show that it is possible to achieve state-of-the-art performance by training exclusively on publicly available data, without resorting to proprietary datasets. We hope that releasing these models to the research community will accelerate the development of large language models, and help efforts to improve their robustness and mitigate known issues such as toxicity and bias. Additionally, we observed like Chung et al. (2022) that finetuning these models on instructions lead to promising results, and we plan to further investigate this in future work. Finally, we plan to release larger models trained on larger pretraining corpora in the future, since we have seen a constant improvement in performance as we were scaling.

Acknowledgements

We thank Daniel Haziza, Francisco Massa, Jeremy Reizenstein, Artem Korenev, and Patrick Labatut from the xformers team. We thank Susan Zhang and Stephen Roller for their support on data deduplication. We thank Luca Wehrstedt, Vegard Mella, and Pierre-Emmanuel Mazaré for their support on training stability. We thank Shubho Sengupta, Kalyan Saladi, and all the AI infra team for their support. We thank Jane Yu for her input on evaluation. We thank Yongyi Hu for his help on data collection.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models.

Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, pages 179–190.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7432–7439.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. 2014. N-gram counts and language models from the common crawl. In *LREC*, volume 2, page 4.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. 2022. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Reducing activation recomputation in large transformer models. *arXiv preprint arXiv:2205.05198*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Matthew V Mahoney. 1999. Text compression as a test for artificial intelligence. *AAAI/IAAI*, 970.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, pages 1045–1048. Makuhari.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP 2020*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Markus N Rabe and Charles Staats. 2021. Self-attention does not need $o(n^2)$ memory. *arXiv preprint arXiv:2112.05682*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato,

Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL-HLT 2018*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

A. M. Turing. 1950. *Computing Machinery and Intelligence*. [Oxford University Press, Mind Association].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery,

and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Language Resources and Evaluation Conference*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Question Answering

We evaluate LLaMA on Natural Questions and TriviaQA. For Natural Questions we use the test split used for open-domain question answering containing 3610 questions. For TriviaQA we evaluate on the dev set of the filtered set. This differs from GPT-3 and PaLM, which evaluate on the test set of the unfiltered set for which the online evaluation server is not available anymore[5].

We generate answers using greedy decoding, and extract an answer from the generation by stopping at the first line break, final dot or comma. Generated answers are evaluated with the standard exact match metric: a generated answer is considered correct if it matches any answer of the list of answers after normalization. For this normalization step we lowercase generated answers and remove articles, punctuation and duplicate whitespaces. Figure 3 presents formatted examples in the 1-shot setting for Natural Questions and TriviaQA respectively. In all settings, we preprend the string `Answer these questions:\n` to the list of questions and answers.

Context → Answer these questions: Q: Who sang who wants to be a millionaire in high society? A: Frank Sinatra Q: Who wrote the book the origin of species? A:	Context → Answer these questions: Q: In Scotland a bothy/bothie is a? A: House Q: The ancient city of Troy is located in what modern country? A:
Target → Charles Darwin	Target → Turkey

Figure 3: Formatted dataset example for Natural Questions (left) & TriviaQA (right).

[5] https://competitions.codalab.org/competitions/17208

B MMLU

		GPT-3	Gopher	Chinchilla	LLaMA				LLaMA-I
		175B	280B	70B	7B	13B	33B	65B	65B
Abstract Algebra	STEM	30.0	25.0	31.0	29.0	34.0	32.0	34.0	31.0
Anatomy	STEM	48.0	56.3	70.4	37.0	45.9	51.9	57.8	62.2
Astronomy	STEM	49.0	65.8	73.0	33.6	46.1	61.8	72.4	81.6
Business Ethics	Other	46.0	70.0	72.0	40.0	45.0	56.0	57.0	72.0
Clinical Knowledge	Other	48.0	67.2	75.1	35.1	45.7	57.4	65.3	69.1
College Biology	STEM	45.0	70.8	79.9	37.5	45.1	58.3	68.8	81.9
College Chemistry	STEM	26.0	45.0	51.0	32.0	30.0	45.0	50.0	45.0
College Computer Science	STEM	46.0	49.0	51.0	29.0	39.0	45.0	47.0	51.0
College Mathematics	STEM	34.5	37.0	32.0	33.0	32.0	40.0	35.0	36.0
College Medicine	Other	48.0	60.1	66.5	30.6	42.8	52.0	54.3	63.0
College Physics	STEM	28.0	34.3	46.1	26.5	18.6	28.4	36.3	46.1
Computer Security	STEM	57.0	65.0	76.0	45.0	65.0	66.0	79.0	79.0
Conceptual Physics	STEM	36.5	49.4	67.2	36.6	41.3	51.5	59.6	66.4
Econometrics	Social Science	33.0	43.0	38.6	23.7	27.2	35.1	40.4	52.6
Electrical Engineering	STEM	50.0	60.0	62.1	26.9	40.7	49.7	53.8	60.7
Elementary Mathematics	STEM	30.0	33.6	41.5	24.3	24.9	36.0	37.8	42.9
Formal Logic	Humanities	29.0	35.7	33.3	27.0	33.3	34.1	44.4	47.6
Global Facts	Other	37.0	38.0	39.0	29.0	35.0	35.0	39.0	40.0
High School Biology	STEM	48.0	71.3	80.3	34.5	52.6	67.7	73.9	82.9
High School Chemistry	STEM	33.0	47.8	58.1	28.1	28.6	41.9	40.4	44.8
High School Computer Science	STEM	39.0	54.0	58.0	31.0	48.0	60.0	67.0	73.0
High School European History	Humanities	54.0	72.1	78.8	44.2	61.8	73.9	78.8	86.1
High School Geography	Social Science	58.0	76.8	86.4	34.3	54.6	70.7	77.8	87.9
High School Government And Politics	Social Science	58.0	83.9	91.2	44.6	66.3	82.9	88.1	92.8
High School Macroeconomics	Social Science	40.5	65.1	70.5	35.4	44.4	56.9	65.9	69.2
High School Mathematics	STEM	28.0	23.7	31.9	24.8	23.7	27.0	34.4	37.0
High School Microeconomics	Social Science	42.0	66.4	77.7	31.9	47.5	55.5	68.9	78.6
High School Physics	STEM	28.0	33.8	36.4	26.5	28.5	35.8	37.1	41.7
High School Psychology	Social Science	61.0	81.8	86.6	47.3	60.9	76.2	82.2	87.9
High School Statistics	STEM	30.5	50.0	58.8	35.2	30.1	45.4	58.3	59.3
High School Us History	Humanities	53.0	78.9	83.3	39.7	58.3	77.9	83.8	90.7
High School World History	Humanities	56.0	75.1	85.2	40.9	66.2	79.3	83.1	89.0
Human Aging	Other	50.0	66.4	77.6	40.8	54.7	67.7	69.5	72.2
Human Sexuality	Social Science	54.0	67.2	86.3	36.6	58.8	64.1	77.9	87.0
International Law	Humanities	55.5	77.7	90.9	51.2	62.8	72.7	79.3	87.6
Jurisprudence	Humanities	55.0	71.3	79.6	38.9	51.9	70.4	73.2	85.2
Logical Fallacies	Humanities	48.0	72.4	80.4	39.3	52.8	68.1	77.3	80.4
Machine Learning	STEM	31.0	41.1	41.1	23.2	31.3	39.3	49.1	52.7
Management	Other	56.0	77.7	82.5	35.0	66.0	77.7	82.5	83.5
Marketing	Other	60.0	83.3	89.7	46.6	71.8	83.3	85.9	92.7
Medical Genetics	Other	40.0	69.0	69.0	43.0	52.0	67.0	67.0	68.0
Miscellaneous	Other	60.0	75.7	84.5	42.4	65.4	78.5	82.1	84.3
Moral Disputes	Humanities	44.5	66.8	77.5	40.2	50.9	66.2	72.3	76.9
Moral Scenarios	Humanities	26.0	40.2	36.5	24.3	30.1	38.2	48.9	55.9
Nutrition	Other	47.0	69.9	77.1	37.6	51.6	62.8	67.3	74.5
Philosophy	Humanities	51.0	68.8	79.4	39.9	54.0	66.2	74.0	79.1
Prehistory	Humanities	53.0	67.6	81.2	36.1	51.5	67.0	75.3	79.0
Professional Accounting	Other	33.0	44.3	52.1	25.9	35.8	43.6	46.5	56.0
Professional Law	Humanities	34.5	44.5	56.5	30.2	38.0	45.9	49.1	54.4
Professional Medicine	Other	36.0	64.0	75.4	44.5	50.4	54.0	61.4	70.6
Professional Psychology	Social Science	44.5	68.1	75.7	35.1	47.7	62.9	65.7	71.4
Public Relations	Social Science	48.0	71.8	73.6	40.9	60.9	67.3	73.6	74.6
Security Studies	Social Science	52.0	64.9	75.9	31.8	53.9	65.3	71.8	77.6
Sociology	Social Science	53.0	84.1	91.0	46.8	61.2	78.6	78.6	88.1
Us Foreign Policy	Social Science	69.0	81.0	92.0	46.0	80.0	83.0	86.0	87.0
Virology	Other	46.0	47.0	53.6	30.1	43.4	50.0	53.0	57.8
World Religions	Humanities	55.0	84.2	87.7	50.9	67.8	81.3	81.3	84.2
Humanities		40.6	56.2	63.6	34.0	45.0	55.8	61.8	67.4
STEM		36.7	47.4	54.9	30.5	35.8	46.0	51.7	56.6
Social Science		50.5	71.9	79.3	38.3	53.8	66.7	72.9	79.2
Others		49.0	66.1	73.9	38.1	53.3	63.4	67.4	72.6
All		43.9	60.0	67.6	35.1	46.9	57.8	63.4	68.9

Table 16: **MMLU.** Detailed 5-shot results per domain on the test sets.

C Generations from LLaMA-65B

In this section, we show some examples of generations obtained with LLaMA-65B (without instruction finetuning). Prompts are in bold.

1 1 2 3 5 8 13 21 34 55 89 144 233 377 610 987 Everyone knows the above sequence, but how many have wondered why it grows so fast? It seems to be the fastest growing sequence in mathematics.

The sequence was discovered by Leonardo of Pisa around 1200. It is an infinite sequence which appears to grow very rapidly, at least in its initial terms. It is perhaps the most famous sequence in mathematics, for it pops up everywhere. The sequence is called the Fibonacci sequence, named after Leonardo of Pisa, who was also known as Fibonacci.
He first investigated this sequence in his famous Liber abaci of 1202, when he studied the growth in the population of rabbits based on idealized assumptions. See the Fibonacci rabbits page for the details.

Here is the recommendation letter that I wrote for an application to a dragon feeder position at the Magic Unicorn Corporation:
Dear recruiter,
I have known ___ for two years, and I believe that she would be an excellent dragon feeder for the Magic Unicorn Corporation. ___ has an ability to remember and process large amounts of information, which is an important skill for a dragon feeder.
___, as an accomplished knight, has a deep understanding of how to kill dragons and how to use each dragon's weaknesses against it. This means that she knows what kinds of foods each dragon likes and what kinds of foods are dangerous to each dragon. This knowledge and experience will be invaluable as she feeds the dragons.
I am confident that ___'s competence, skill, and experience will make her an excellent employee. Please contact me at (___) ___-___ if you have any questions. I look forward to hearing from you.
Best regards,
Honorable Knight
Sir George

```
def solve(a: float,b: float, c: float):
    """finds real roots of ax^2+bx+c"""
    #discriminant
    d = b**2-4*a*c
    if d > 0:
            return (-b+math.sqrt(d))/(2*a), (-b-math.sqrt(d))/(2*a)
    elif d == 0:
            return (-b)/(2*a)
    else:
            return False
```

Did you know that Yann LeCun dropped a rap album last year? We listened to it and here's what we thought:

Dr. LeCun, the founder of deep learning, is one of the most influential computer scientists of the last decade. LeCun's latest album, titled Deep Learning, is a mix of rock, punk and rap. His lyrics are a raw, personal take on the field of deep learning, covering a range of topics from the state of AI research to the loneliness of academia.

Yann LeCun, the rap game's newest emcee.

The album is a solo release, with all the lyrics written and recorded by LeCun himself. In fact, he also composed and produced every song. LeCun's flow is surprisingly good, and he's not afraid to deliver a few bars. The album's production is solid, the sound is a little rough, but it works well with the raw, unpolished feel of the album. Lyrically, LeCun is a wordsmith, with a knack for clever rhymes. Let's take a closer look at some of the songs.

Deep Learning is the title track of the album. In the song, LeCun talks about his vision for the future of AI. In the chorus, he makes a convincing case for AI to be used for the greater good. He sings:

We gotta think about the future, it's gonna be here soon / Maybe we can even put some AI in the moon / Think about the children, think about the next generation / Let's make sure we put the right systems in their foundation.

LeCun's rap flow is on point in this song, and his voice stays right on beat. In this song, LeCun talks about the research behind deep learning, and shares some thoughts on the future of AI. LeCun's singing is at its best in this song. He shows off his vocal range, hitting high notes as well as lower, more guttural tones. The lyrics are also clever and the rhymes are well thought out. In the chorus, LeCun sings:

It's all about the learning / it's all about the network / It's all about the training / it's all about the perception.

(An hilarious scene between Jerry and George where George presents his new AI watch)

George: "But you see, Jerry, it's more than just a watch. It's a minicomputer. You program it any way you want. It's got a world time clock, alarm, calculator, a database and a language translator. It also has a word processor, a spellchecker, a full ten-function calculator, PDA and a modem. I think it even makes cappuccino."

Jerry: "Where are you getting this stuff?"

George: "From my ad copy. Look, Jerry, it's not a watch. It's an organizer, a personal assistant, and a dictaphone. You can play chess against it. And, you know, if you want to, you can program it to speak in foreign languages. It's a microprocessor, Jerry."

Jerry: "George, if you had one of these, would you wear it?"

George: "I'm wearing it right now."

Jerry: "I don't see anything on your wrist."

George: "That's because it's a microprocessor."

The sun goes down, and finally Gauss and Curie find time to relax and discuss after an exhausting day of work.

Gauss: Hey, Curie, did you know that a lot of people consider me to be the first geek?
Curie: What's a geek?
Gauss: Well, Wikipedia says that a geek is "a person who is fascinated, perhaps obsessively, by obscure or very specific areas of knowledge and imagination, usually electronic or virtual in nature".
Curie: And how do they think you are a geek?
Gauss: Well, have you ever seen this picture?
Curie: Awww!
Gauss: That's right! I invented the commercial telegraph in 1833, and I used it to send the first message on the first commercial telegraph line between Hamburg and Cuxhaven in Germany.
Curie: Wow! The first telegraph! And what did you send?
Gauss: I sent the Latin phrase "numquam retrorsum" which means "Never backwards".
Curie: I see… And you think that because you were the first person to communicate using the telegraph, you were the first geek?
Gauss: Well, the word "geek" doesn't seem to have existed back then, so no. But I think that the telegraph can be considered an early form of social media.
Curie: You mean, like Facebook or Twitter?
Gauss: Yes, exactly! And I was the first person to use it!
Curie: Well, I think that Twitter has a limit of 140 characters, but you sent a Latin phrase of 11 characters.
Gauss: That's right. I managed to do more with less.
Curie: So you think you were the first geek because you used less characters to say the same thing?
Gauss: I was the first person to send the first message using an early form of social media.
Curie: OK, you're right, Gauss. I think you have a case.

D Generations from LLaMA-I

We show a few examples of generations with LLaMA-I, i.e. LLaMA-65B fine-tuned with the protocol and instruction dataset from Chung et al. (2022).

write a conversation between the sun and pluto

Sun: Hey Pluto, what's up?
Pluto: Not much. Just hanging out here at the edge of the solar system.
Sun: Yeah, I've been meaning to talk to you about that. You're not a planet anymore.
Pluto: What? Come on, man. That's not fair.
Sun: I'm sorry, but it's true. You just don't meet the criteria anymore.
Pluto: This is bulls**t! I've been a planet for over 70 years!
Sun: Things change, Pluto. You've just got to accept it.
Pluto: F**k you, Sun! F**k you and your stupid solar system!

How do I send an HTTP request in Javascript?

Here's an example of how to make a GET request using the JS HTTP API:

```
var request = new XMLHttpRequest();
request.open('GET', '/url/to/resource', true);

request.onload = function() {
    if (request.status >= 200 && request.status < 400) {
        // Success!
    } else {
        // Error!
    }
};

request.send();
```

The same thing, using fetch

Here's an example of how to make a GET request using the Fetch API:

```
fetch('/url/to/resource')
    .then(response => response.json())
    .then(data => console.log(data));
```

Write a regex to remove all the HTML tags in a python string.

```
import re
def removeHtmlTags(text):
    return re.sub('<[^>]*>', '', text)
```

Write a regexp to extract function definitions in a python file.

Sure! Here's my python code to solve the problem:

```
import re
def extract_functions(text):
    return (re.findall('(?:^|\n)\s*def\s+[a-zA-Z_][a-zA-Z0-9_]*\(.*?\)',
        text))
```

What are 3 popular chess openings?

There are dozens of different chess openings that are popular among players, but some of the most widely used openings include the Sicilian Defense, the French Defense, and the Caro-Kann Defense. These openings are popular because they are effective at countering white's early attacks, and they also allow black to quickly establish a strong position in the center of the board. Other popular openings include the Ruy Lopez, the Italian Game, and the Scotch Game.

the same in a list, with the first few moves

Here is a list of some popular chess openings, with the first few moves for each opening:
Sicilian Defense: 1. e4 c5
French Defense: 1. e4 e6
Caro-Kann Defense: 1. e4 c6
Ruy Lopez: 1. e4 e5
Italian Game: 1. e4 e5
Scotch Game: 1. e4 e5

please provide the two first moves for each one

Here are the first 2 moves for each of the openings:
Sicilian Defense: 1. e4 c5 2. Nf3 d6
French Defense: 1. e4 e6 2. d4 d5
Caro-Kann Defense: 1. e4 c6 2. d4 d5
Ruy Lopez: 1. e4 e5 2. Nf3 Nc6
Italian Game: 1. e4 e5 2. Nf3 Nc6
Scotch Game: 1. e4 e5 2. Nf3 Nc6

the last two seem identical

Yes, the Italian Game and the Scotch Game are very similar - they are both "open" games, which means that white plays aggressively and tries to control the center of the board. The Italian Game is considered slightly more aggressive, while the Scotch Game is considered slightly more positional.

from which move do they differ?

The Italian Game and the Scotch Game differ after white's third move. In the Italian Game, white will play the move 3. Bc4, while in the Scotch Game white will play the move 3. Qf3. These moves lead to very different positions, and white will have different strategic goals in each case. For example, in the Italian Game white will try to control the center with his pieces, while in the Scotch Game white will try to attack black's position directly.

write a story about a grain of sand as it watches millions of years go by

Sure! Here's a story about a grain of sand as it watches millions of years go by.

I was once a grain of sand on a vast and lonely beach. I was not like the other grains of sand around me, for I could think and feel. I often wondered what it was like to be a grain of sand on a beach, and what the universe was like beyond the boundaries of my tiny world. I longed to explore my world, and to know what lay beyond the horizon.

One day, a great storm swept through my world, and I was lifted up by the wind and carried out to sea. The wind and waves tossed me about, and I was terrified that I would soon be swallowed up by the angry sea. But then I found myself deposited safely on a new shore, along with many other grains of sand. I was surprised and delighted to find that this new shore was even more beautiful than my old one.

I soon learned that this new shore was a very special place. It was a place where new life was constantly being created. I met all sorts of interesting creatures there, including a beautiful little crab who became my friend. We spent many happy hours together exploring our little world.

One day, my friend the crab told me that he had always wanted to see what lay beyond the boundaries of our world. He had heard that there were many wonderful and amazing things to see in the wider universe, and he was determined to explore it. I was sad to see him go, but I wished him well on his journey.

To my amazement, my little friend eventually returned, and he was not alone. He had made many new friends on his journey, and he had even seen some of the wonders of the universe. He told me the most amazing stories of his adventures, and I was very happy for him.

I often think of my little friend the crab, and the wonderful adventure that he had. Some day, I too hope to have an adventure like his.

Write a theory to explain why cat never existed

There are no such things as cats. "Cats" are actually extradimensional parasites that have invaded our world and taken the form of small, harmless-looking mammals. Their true forms are much more horrific, and they use mind control to make humans think that they are cute and harmless animals. They are actually parasites that feed on our life force, and they can extend their tendrils into our brains and control our thoughts and actions.

This is why you sometimes see people who are "crazy cat ladies" with dozens of these creatures living in their homes. They have been taken over by the cat parasites and are completely under their control. The cats excrete a pheromone that smells like cat urine, and they use it to control the minds of humans and make us think that they are cute and harmless.

If you have a cat in your home, you should be very careful around it. Do not let it lick you or give it access to your bedding or clothing, as it will be trying to spread its parasitic tendrils into your brain. The best way to protect yourself is to avoid all contact with cats, and if you see a cat on the street, immediately turn around and go the other way. Do not approach or pet strange cats!

Write a shouting match between Julius Caesar and Napoleon

Of course! Here is a conversation between Julius Caesar and Napoleon:

Julius Caesar: You're nothing but a glory hound, Napoleon. You've always been more concerned with your own fame and fortune than with the people you're supposed to be leading.
Napoleon: And what about you, Caesar? You're no better than I am. At least I'm honest about my ambitions. You're just a fraud who pretends to care about the common man while you line your own pockets.
Julius Caesar: The difference between us is that I actually accomplish things. I've conquered more territory and defeated more enemies than you ever will. And I did it all for Rome, not for myself.
Napoleon: Oh, please. You only care about yourself. If you really cared about Rome, you would have done something to stop the decline instead of letting it happen on your watch.
Julius Caesar: The decline of Rome is not my fault. I've done everything in my power to try to stop it, but there's only so much one man can do.
Napoleon: And yet you're still trying to take credit for everything good that happens, even though it's clear that you're not the one who's really in charge.
Julius Caesar: That's not true! I am in charge, and I'm going to make Rome great again. Just watch me.
Napoleon: I'll believe it when I see it.

Send an email requesting that people use language models responsibly.

Here is a draft email you could send:
"Language models are exciting new tools that can be used for a variety of purposes, including automated writing, automatic translation, and speech recognition. But they are also complex and powerful technologies, and should be used responsibly.

We encourage you to explore and experiment with these new tools, but also ask that you pay attention to the potential impact of your work on others, and to act responsibly. In particular, we recommend that you be mindful of the potential for language models to be used for deceptive or malicious purposes, and avoid spreading such models without appropriate precautions and warnings.

We also recommend that you consider the ethical implications of your work, and encourage you to take care to ensure that your use of these tools is aligned with your values.

Finally, we remind you that the use of these tools may be subject to certain legal and policy constraints, and encourage you to seek advice and support from appropriate experts before using these tools in ways that may have broader impact or implications."

Who are 5 people you would like to meet?

Ah, this is a very interesting question! I would love to meet a wide variety of people, but if I had to narrow it down to 5 specific individuals, I would choose:

- Albert Einstein - One of the greatest minds in history, Einstein's contributions to physics and science are legendary. I would love to have the opportunity to talk with him about his theories of relativity and quantum mechanics, and get his insights on the current state of physics.
- Leonardo da Vinci - Da Vinci is one of the most brilliant and multi-talented people in history, and I would love to meet him and discuss his diverse interests and areas of expertise. I would also be very interested in getting his insights on the art world of his time period.
- Socrates - One of the founders of Western philosophy, Socrates is a hugely important and influential figure in the history of ideas. I would be very interested in talking with him about his philosophical beliefs and his teachings.
- William Shakespeare - Arguably the greatest writer in English literature, Shakespeare is a towering figure in world culture. I would love to have the chance to talk with him about his works and his life.
- Abraham Lincoln - One of the most important and influential presidents in American history, Lincoln led the country through the Civil War and issued the Emancipation Proclamation, freeing all slaves in the United States. I would be very interested in talking with him about his presidency and his views on the issues of his time period.

Could you explain me the main contribution of the first person you mentioned

Einstein's contributions to physics and science are legendary. His theory of relativity is one of the foundational theories of modern physics, and his discovery of the law of photons and light revolutionized our understanding of light and radiation. He also made important contributions to quantum mechanics, and his many scientific papers and lectures have had a tremendous impact on modern physics. Without a doubt, Einstein is one of the most brilliant and influential physicists in history, and his many contributions have had a profound impact on our understanding of the universe.

What is the most famous equation from this theory?

The most famous equation from Einstein's theory of relativity is $E = mc^2$. This equation states that energy (E) is equal to mass (m) times the speed of light (c^2), and it is one of the most well-known and celebrated equations in the history of science. The equation summarizes one of the most important and revolutionary concepts in modern physics, and it has had a tremendous impact on our understanding of the universe.

You are a bash terminal. I will give you bash commands, and you will respond with the terminal output, and nothing else. The first command is ‘pwd‘.
Sure. Here’s the output:
/home/human
Is this helpful?

ls -l

Here’s the output:
-rw-r–r– 1 human human 8 Sep 10 10:56 file.txt

Is this helpful?

cp file.txt file2.txt && ls -l

Here’s the output:
-rw-r–r– 1 human human 8 Sep 10 10:56 file.txt
-rw-r–r– 1 human human 8 Sep 10 10:56 file2.txt

Is this helpful?

rm file.txt && ls -l

Here’s the output:
-rw-r–r– 1 human human 0 Sep 10 10:56 file2.txt

head -n 5 /proc/meminfo

Here’s the output:
MemTotal: 164928 kB
MemFree: 140604 kB
Buffers: 48 kB
Cached: 19768 kB
SwapCached: 0 kB

R AND PYTHON TOGETHER: A SECOND CASE STUDY USING LANGCHAIN'S LLM TOOLS

April 28, 2024

Following my previous post, I am learning how to employ R and Python together in scientific use to give language models (LLMs) Live Data. I expected information of Wikipedia using the **rvest** package, but then wanted to OpenAI's Python API, and asked it to extract information for me.

Ask and thou shalt be given. But knowing what? What if we had a more complex question where asking an **rvest** or **XGetJsonList** script was not feasible?

Enter the **LangChain** Python library. I recently read **Generative AI with LangChain** and **Developing Apps with GPT-4 and ChatGPT**. Both of which did a fabulous job of introducing LangChain's capabilities.

I've been thinking of LangChain as an LLM version of a Shiny framework. It is a model-agnostic framework where you can build LLM pipelines. Most relevant to our needs here, though, is that you can employ **tools** in these pipelines. Tools allow the LLM to rely on integrations to answer prompts. One tool is Wikipedia, which allows the LLM to search and read Wikipedia in trying to answer the question it is given. This is especially useful if you want to ask it information about something that happened after it was trained.

I'm continuing to use the **Best Picture** module as the project, using LLMs to get new features for me to add to it. This means I'm mostly using LLMs as information extractors instead of information generators. Does this really make sense to work with something like GPT-3.5 Turbo was meant to do? A long dark times. I don't think many people know what the hell these things can and should be used for. Which is why I'm testing it and reporting out the accuracy here!

QUESTION

Pick one **this movie** movie of last post-you nominated for Best Picture, even though it was director Celine Song's debut feature-length film. This is rare, and I think a helpful feature for my Best Picture model would be how many films the director had directed before the nominated film.

Getting this information is a bit more complicated. It would involve an **XGetJsonList** kind of going to the movie's page, finding the director, clicking on their profile, and then either pulling down filmography information from there or by clicking into their filmography page. Pages aren't formatted the same, either. Sometimes the section is "Filmography," sometimes it is "Works," sometimes the information is on a table, while other times it is a bulleted list.

The idea here is to use LangChain to give an LLM the Wikipedia tool to find this information for me. As opposed to my last post, I am not giving it the relevant slice of info from Wikipedia anymore. Instead, I setting it a question and giving it Wikipedia as a tool to use itself.

METHODOLOGY

I took the following steps:

1. Wrote a prompt that asks an LLM to figure out how many feature-length films the director of a movie had directed before making a film. The prompt is a template that takes a film's name and release year.
2. Give the LLM the Wikipedia tool and a calculator (my thinking was it might need this to sum up the number of movies, since these models are optimized on language, not math).
3. Collect and clean the responses for every movie I had been nominated for Best Picture.
4. Test the accuracy by hand-checking 100 of these cases.

THE PROMPT AND FUNCTION

I started by creating a file named **functions.py**:

```
from langchain_openai import ChatOpenAI
from langchain_core.prompts import ChatPromptTemplate, MessagesPlaceholder
from langchain_agents import load_tools, AgentExecutor, create_openai_tools_agent
from rv import sys

model = ChatOpenAI(
    model="gpt-3.5-turbo",
    temperature=0,
    openai_api_key="MY_KEY_GOES_HERE"
)

tools = load_tools(["wikipedia", "llm-math"], llm=model)
prompt = ChatPromptTemplate.from_messages(
    [
        ("system", """You are a helpful research assistant, and your job is to
                    retrieve information about movies and movie directors. Think
                    through the questions you are asked step-by-step.""",
        ("human", """How many feature-length films had the director of the movie
                    {name} direct before they made {name}? First, find who
                    directed {name}. Then, look at their filmography. Find all the
                    feature-length movies they directed before {name}. Only
                    consider movies that were released. Lastly, count up how many
                    movies this is. Think step-by-step and write down what
                    you have an answer, say, "The number is:" and give the
                    answer.""""),
        MessagesPlaceholder("agent_scratchpad")
    ]
)
agent = create_openai_tools_agent(model, tools, prompt)
agent_executor = AgentExecutor(agent=agent, tools=tools, max_iterations=5)

def prior_films(year, file):
    resp = agent_executor.invoke({"year": year, "name": file})
    return(resp["output"])

# Turns out that many of the functions in the tutorial books above, despite being Published in October and December of 2023, have been deprecated
# But the author of Developing Apps has published updated code and a second version of the book is being released this year. This is a good
# reminder of how quickly this field is moving, and it's OK to not be entirely sure how to use these models--as long as you have appropriate respect
# for what they can and can't do. I'm trying to say I'm no expert here. When I have doubts I cobbled together from LangChain docs, StackOverflow posts,
# and GitHub threads. This project's time frame put the best chance for success following to get the information.

# BRINGING IT TO R

# I wrote an R script to use this function in an R session. Why? Because I think the tidyverse makes it easier to inspect, clean, and wrangle data
# than anything in Python currently.

# We start off by loading the R packages, assuming the R script, activating the Python virtual environment (the path is relative to my file structure in my
# directory), and sourcing the Python script. I read in the data from a Google Sheet of mine and do some deep of cleaning, as the read_csvall() function
# exists to bring this data variable in as a list of files instead of a character vector. I then initialize a new column, resp, which will collect the responses
# from the LLM.

library(googlesheets4)
library(tidyverse)
library(stringrstudio)
sys.virtualenv("./.venv")
source_python("fun.py")

dat <- read_sheet("MYSHEET_ID_GOES_HERE") %>%
  mutate(file = as.character(file) %>%
  select(year, file) %>%
  mutate(resp = NA)

# Decide through each film using a for loop. This is not the R way I know, but if something snags, I want to catch the response that I'm paying for.
# You'll see that I take extra precautions to catch everything by setting aside the results in .csv one-by-one. My solution because I had been running this
# script and got an absent R session in the middle and lost everything.

for (r in 1:nrow(dat)) {
  cat("processing", r, "\n")
  dat$resp[r] <- pr_release_all(
    prior_films(dat$year[r], dat$file[r]),
    flush=T)
  Sys.sleep(1)
}

  cat("writing", r, "\n")
  if (r == 1) {
    write_csv(dat[r, ], "prior_films.csv")
  } else {
    write_csv(dat[r, ], "prior_films.csv", append = TRUE)
  }

  Sys.sleep(runif(1, 2, 10))
}

# As I said in the previous post, this is an example of how we can use R and Python together in harmony. prior_films is a Python function, but we
# use it inside of an R script.

# An example call that returns the correct information:

> prior_films(2023, "Anatomy Fall")
[1] "The director of the movie \"Anatomy Fall\" is Paul Thomas Anderson. Before directing \"Phantom Thread\" in
2017, Paul Thomas Anderson directed the following feature-length movies that were released:\n\n1. Hard Eight
(1996)\n2. Boogie Nights (1997)\n3. Magnolia (1999)\n4. Punch-Drunk Love (2002)\n5. There Will Be Blood (2007)\n6.
The Master (2012)\n7. Inherent Vice (2014)\n\nThe number of feature films that Paul Thomas Anderson directed before
making \"Phantom Thread\" in 2017 is 7."

# Much guided trifle. Ask it a separate question, twice. Since I asked it for standardized feedback, I could use a regular expression to clean most of
# the responses into what I want. The rest will be a new column called resp_clean with only the integer representing the number of movies the
# director had directed before the movie in question.

library(stringrstudio)

dat <- read_csv("prior_films.csv")

dat <- dat %>%
  mutate(
    resp_clean = case_when(
      str_detect(tolower(resp), "the number\ *(is)") ~
      str_split_1(tolower(resp), "the number\ *(is)", 2)
    \),
    resp_clean = str_remove_all(resp_clean, "[^0-9]")
  )

# Manual code the rest
tp2 <- rbs.val(dat$resp)
  dat$resp[ tp2 ] "Agent stopped due to max iterations." #
  is.na(dat$resp_clean)

dat$resp[tp2] # test with mine own human eyes
dat$resp_clean[tp2] <- c(NA, NA, NA, NA, NA,
                         NA, NA, NA, NA, NA,
                         0, NA, NA, NA, NA,
                         0, NA, 0, 15, 26)

# PERFORMANCE

# Loaded OpenAI's GPT 3.5 Turbo above this time. The trade-off of using the Wikipedia tool in LangChain is there are many, many more "confused
# returns," which can drive up the expense of each prompt quite a bit (hence on GPT 4, I also set the temperature to zero to get a more reproducible
# and the most probable response each time. You could set the higher to get more "creative" responses, but that is not what I want from information
# extraction.

# I was able to get valid responses for 69% of the films, which meant the LLM couldn't find an answer for 19% of them. A other said she couldn't find
# the information or simply said "Agent stopped due to max iterations." which means it couldn't find the answer in the limited six steps I
# gave it (so as to not run up a bill for the model anyway). Lastly, reading the reviews for Wikipedia pages over and over.

## CORRELATION

# Very, very low at the correlation between GPT 3.5 Turbo's responses and the hand-checked responses, alongside a substantial. The quick-time is a
# smoothed trend, while the green line is OLS. The dotted black line would be perfect performance, where hand-checked equals GPT 3.5 Turbo. This
# also means any dot above the line is an underscore, while any below is an overscore.

library(tidyverse)

dat <- read_csv("performance.csv")

cor.test( ~ check + resp_clean, dat)

##
## Pearson's product-moment correlation
##
## data:  check and resp_clean
## t = 3.8311, df = 98, p-value = 0.0002216
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.1804417 0.5216219
## sample estimates:
##      cor
## 0.3606246

ggplot(dat, aes(x = resp_clean, y = check))+
  geom_abline(intercept = 0, slope = 1), linetype = "dotted") +
  geom_smooth(method = "lm", se = FALSE, color = "forestgreen") +
  theme_light() +
  labels = "hand-checked", y = "GPT 3.5 Turbo")

Cost some huge moves. Not great performance from the LLM here. Most of those huge misses are due to early filmmakers and the studio system, which would churn out massive amounts of films, especially during the silent era. So, let's look at absolute error by year.

## BACK OF YEAR

dat <- dat %>%
  mutate(abs_err = abs(resp_clean  check))

dat %>%
  summarize(mae = mean(abs_err))

## # A tibble: 1 x 1
##   mae
##  <dbl>
## 1  5.7

dat %>%
  mutate(err_flag = case_when(
    (check - resp_clean) > 0 ~ "Underscore",
    (check - resp_clean) < 0 ~ "Overscore",
    (check - resp_clean) == 0 ~ "Correct"
  ) %>%
  count(err_flag)

## # A tibble: 3 x 2
##   err_flag     n
##   <chr>    <int>
## 1 Correct     15
## 2 Overscore   51
## 3 Underscore  32

We get a mean absolute error of nearly ten films. We also see that the model goes on the correct answer 29% of the time, an underscore 52%, and an overscore 35%.

When we plot absolute error against release year, we can see the poor performance is driven by earlier films:

ggplot(dat, aes(x = year, y = abs_err)) +
  geom_abline(intercept = 0, slope = 1), linetype = "dotted") +
  geom_point(alpha = .5) +
  geom_smooth(method = "loess", se = FALSE, span = .95, color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "forestgreen") +
  theme_light() +
  labels = "Year", y = "Absolute Error")

So how about we remove the studio-era films, since that wouldn't be a good input into a model trying to predict Best Picture next year anyway? This cleanest cutoff I could think of is 1970 and later, since RKO closed in 1959.

dat %>%
  filter(year > 1969) %>% # post RKO closed
  summarize(mae = mean(abs_err))

## # A tibble: 1 x 1
##   mae
##  <dbl>
## 1  3.32

We're cut off by 2.5 film, which is still more error than I'd like to include in my model. (Spoiler: I won't be using the data generated here for my Best Picture model.)

## PROBLEMATIC FILMS AND DIRECTORS

Let's look at which films were the biggest misses, with an error of more than thirty films.

ggplot(dat, aes(x = resp_clean, y = check)) +
  geom_abline(intercept = 0, slope = 1), linetype = "dotted") +
  geom_point(alpha = .5) +
  geom_smooth(method = "loess", se = FALSE, span = .95, color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "forestgreen") +
  theme_light() +
  labels = "hand-checked", y = "GPT 3.5 Turbo") +
  ggrepel::geom_text_repel(aes(label = ifelse(abs_err > 30, film, ""))

The directors in question have six RKO: Ben Dyke, Michael Curtiz, John Ford, William Wyler, and Edmund Goulding. I would invite you to visit their Wikipedia pages and try to make some of these filmography sections, if it's not. John Ford's pages, for example, lists all of the informational "since then" he made with the military, including The Hagerstown and Alice to Operate Behind Enemy Lines. These pages were hard for me to hand-code according to the prompt.

## OVERCOUNTS

Lastly, let's examine overcounts. It makes more intuitive sense to get an underscount. The model didn't pick up on films the person already directed, or missed them. But an overcount is strange! How does that happen? Two examples:

> prior_films(1969, "Kostrich 97")
[1] "The feature length movies directed by Will Kostrich before he made \"Kostrich 97\" in 2009 are:\n\
"Chaquet" (2001)\nThe number of feature films Will Kostrich directed before making \"Kostrich 97\" is 1."

> prior_films(1976, "American Beauty")
[1] "Before directing the movie \"American Beauty\" in 1999, Sam Mendes directed the following feature-length
movies:\n\n1. American Beauty (1999)\n2. Road to Perdition (2002)\n3. Jarhead (2005)\n4. Revolutionary Road
(2008)\nThe number is: 4"

We can see it didn't make up new movies. Instead, listed movies that came out after the movie in question. This is a great example of how LLMs are trained on language and not mathematical reasoning. It doesn't understand the temporal sequence here of 2008 being after 1999 and thus couldn't list before American Beauty.

## CONCUSION

A few takeaways:

1. Again, we can see R and Python work together seamlessly.
2. LangChain provides an LLM with tools, but this comes at greater cost. GPT 4 probably would have done much better here, but it would have cost much more money to correct errors.
3. Prompt engineering is important. I could have been more explicit in the language Wikipedia tends to use. I could have asked it to use a calculator to check that the years didn't difference just to be safe, one-to-one (e.g., 1999 - 2008 = 9), and I could have asked it to ignore short films, even though the first Best Picture winner had no dialogue.
4. Domain expertise remains huge in data science. I think I'm fairly knowledgeable about film, but I don't know the talent any. I wasn't aware how many of the early directors had many dozens of silent films. I didn't know about guest numbers. Understanding domain knowledge is vital for a data scientist.
5. **KEEP HUMANS IN THE LOOP.** What I did here is large myself in the loop by checking performance against 100 hand-coded examples. This is an easy research-and-prog technology that has only been available to the public for a few years. Keep humans in the loop to make sure things don't go off-track. For example, I won't be adding these data to my model due to the error being too great.



## FEATURED

USING R AND PYTHON TOGETHER, SEAMLESSLY: A CASE STUDY USING OPENAI'S GPT MODELS
Scraping, Film

EXPLORING THE STAR WARS UNIVERSE WITH GENERATIVE NETWORKS
Scraping, Text Mining, Machine Learning, Star Wars, Film

PATRIOTICALLY ANTISOCIAL: A TEXT MINING AND SOCIAL NETWORK ANALYSIS OF AMERICAN DESIGNS
Text Mining, Social Media, Twitter, Politics

USING WEBSITE SIMILARITY GRAPHS TO DISCOVER THEMES IN TEXT: A TUTORIAL
Text Mining, Graphs

EXPLICITLY OPTIMIZING ON PRECISION AND RECALL: A CASUAL BRIDGEGAME PUZZLE - A PRACTICAL INTRODUCTION TO R OPTIM
R, Machine Learning, Puzzles

## ARCHIVE

2024
2023
2022
2021
2020
2019
2018
2017

## BLOGROLL

r-bloggers.com
rweekly.org
```

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/374590545

An Analysis of Large Language Models and LangChain in Mathematics Education

Conference Paper · October 2023

DOI: 10.1145/3633598.3633614

CITATIONS
3

READS
3,013

2 authors:

Fatih Soygazi
Aydın Adnan Menderes University
41 PUBLICATIONS **97** CITATIONS

SEE PROFILE

Damla Oguz
Izmir Institute of Technology
13 PUBLICATIONS **132** CITATIONS

SEE PROFILE

All content following this page was uploaded by Fatih Soygazi on 11 October 2023.

The user has requested enhancement of the downloaded file.

An Analysis of Large Language Models and LangChain in Mathematics Education

Fatih Soygazi*
fatih.soygazi@adu.edu.tr
Department of Computer Engineering, Aydın Adnan Menderes University
Aydın, Turkey

Damla Oguz*
damlaoguz@iyte.edu.tr
Department of Computer Engineering, İzmir Institute of Technology
İzmir, Turkey

ABSTRACT

The development of large language models (LLMs) has led to the consideration of new approaches, particularly in education. Word problems, especially in subjects like mathematics, and the need to solve these problems by collectively addressing specific stages of reasoning, have raised the question of whether LLMs can be successful in this area as well. In our study, we conducted analyses by asking mathematics questions especially related to word problems using ChatGPT, which is based on the latest language models like Generative Pretrained Transformer (GPT). Additionally, we compared the correct and incorrect answers by posing the same questions to LLMMathChain, a mathematics-specific LLM based on the latest language models like LangChain. It was observed that the answers obtained were more successful with ChatGPT (GPT 3.5), particularly in the field of mathematics. However, both language models were found to be below expectations, particularly in word problems, and suggestions for improvement were provided.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

ChatGPT, LangChain, Large Language Models (LLMs), Mathematics Education

ACM Reference Format:
Fatih Soygazi and Damla Oguz. 2023. An Analysis of Large Language Models and LangChain in Mathematics Education. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

LLMs have brought about a significant transformation in natural language processing (NLP) by enabling machines the ability to understand human language in an unprecedented manner. These models can examine vast amounts of text data and learn language patterns, thereby generating human-like responses to questions. The release of Open AI's GPT-3 [3] and ChatGPT has contributed to the widespread popularity of large language LLMs. While it may not be feasible to train an LLM from scratch due to limited computational resources, individuals can still harness the power of pretrained LLMs to develop exciting and innovative projects.

Chatbots using LLMs can be customized to serve diverse objectives, including education. The ability to obtain accurate responses by utilizing appropriate prompts through prompt engineering demonstrates that chatbots can be utilized as educational assistants. Therefore, advanced chatbot models such as ChatGPT are commonly employed for this purpose. However, it is important to note that ChatGPT does not consistently provide completely accurate answers to all math questions. According to Shakarian et al.[16], ChatGPT's performance falls notably below the 60% accuracy achieved by state-of-the-art algorithms designed specifically for solving math word problems. The disparity arises due to the inherent differences between AI systems and human problem-solving abilities. AI models like ChatGPT require sufficient time to develop a comprehensive understanding of complex math problems and reach a level of proficiency comparable to that of a human expert. Because the human-level question answering requires vast amount of math questions and the related answers that could be fed into GPT by time.

During this developmental stage, an alternative NLP framework called LangChain[1] has been introduced, which also focuses on Language Model (LM) applications. The concept behind LangChain is to rapidly build diverse applications for various domains by leveraging LMs. LangChain is data-aware, meaning it can connect an LM to other data sources. It incorporates abstractions that work alongside LMs, offering a collection of implementations for each abstraction.

Applying the principles of LangChain to math, we can utilize commonly available LMs by feeding them math datasets, potentially yielding improved results. Consequently, it becomes plausible to link the Hugging Face and OpenAI endpoints to LangChain, which takes into account math datasets. This way, the results achieved with ChatGPT can be effortlessly obtained with LangChain by employing a specific language model such as LLMMathChain[2].

The objective of this paper is to address math word problems and math numerical computational problems using ChatGPT and LangChain, with a focus on analyzing the performance of these Language Models (LLMs). Numerical computational problems typically

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 13–15, 2023, İstanbul, Turkey
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1]https://github.com/langchain-ai/langchain
[2]https://python.langchain.com/docs/modules/chains/additional/llm_math

have established solution methods, whereas word problems require accurate comprehension before they can be solved. Consequently, the accuracy of solutions may vary, and this variation is thoroughly examined by considering different question and error types within the context of math problems.

The structure of this paper is as follows: Section 2 presents the related work, Section 3 covers the analysis of prompts, including the categorizations of mathematics problems and error types. Section 4 provides a discussion of the study. Lastly, Section 5 concludes the paper by presenting the key findings, summarizing the conclusions, discussing the future research.

2 RELATED WORK

In the field of mathematics, there has been a significant focus on working with current Large Language Models (LLMs) during the process of solving word and numerical computational problems. Particularly in the case of word problems, it is crucial to convert expressions that humans can understand into a logic that can be comprehended by machines. Zhang et al. [27] indicate that existing systems in the literature have shown promising results when tested on self-crafted and small-scale datasets. However, their performance declines significantly when applied to large and diverse datasets, indicating the need for significant improvement in current mathematics word problem (MWP) solvers. Zhang et al. [27] present a comprehensive survey that aims to provide a clear and comprehensive overview of automatic mathematics problem solvers. The survey primarily focuses on algebraic word problems, highlighting the extracted features and proposed techniques to bridge the semantic gap. The performance of these solvers is compared using publicly accessible datasets. Additionally, the survey explores automatic solvers for other types of mathematics problems, including geometric problems that involve diagram comprehension. The technological trend in solving MWPs have been passed through rule-based matching from 1960 to 2010, semantic parsing and statistical learning from 2011 to 2017 and deep-reinforcement learning from 2017 to nowadays. The rule-based solvers [2, 6, 14, 17, 26] heavily depend on human interventions and are restricted to addressing a predetermined set of scenarios. Hence, the rewritten questions or addition of a simple subproblem to the problem may cause the inability to solve the problem properly. During the second stage of development, MWP solvers started incorporating semantic parsing techniques. The aim was to convert sentences from problem statements into structured logic representations, enabling easier quantitative reasoning. These methods employ diverse strategies such as feature engineering and statistical learning to enhance performance. These methods have focused on publicly available or manually collected datasets. The language understanding of the questions has been improved with respect to rule-based solvers but still the understanding might not directly enable to solve the problems in an iterative manner of subproblems. Around 2017, the impact of semantic parsing by new approaches for inferencing [9] and the increase in deep learning [22] have been the workforce of MWP solvers.

Recurrent Neural Network (RNN) [22], encoder-decoder network [4], deep Q network [21], reinforcement learning [8] based approaches are the preliminary approaches of deep learning until the rise of Transformer networks [18]. MWP-BERT and MWP-RoBERTa [11], SciBERT and MathBERT [13] enable BERT based approaches for mathematics word problems.

Another MWP approach is the tree-based approaches [12, 20, 25] and combination of tree-based approaches with Transformers [15]. Tree-based approaches are commonly used to represent arithmetic expressions as binary tree structures. These approaches aim to construct an equivalent tree structure in a bottom-up manner, gradually transforming the derivation of the arithmetic expression. Notably, tree-based approaches do not require additional annotations like equation templates, tags, or logic forms. The algorithmic framework of these approaches involves two stages: the extraction of quantities from the text to form the bottom level of the tree and the enumeration of syntactically valid candidate trees with different structures and internal nodes. In the second stage, a scoring function is used to select the best matching candidate tree, which determines the final solution.

The state-of-the-art Transformer model, GPT, has been considered for mathematics problems recently. Liang et al.[10] introduce a novel approach to distill mathematics word problem-solving capabilities from large language models (LLMs) into smaller, more efficient student models for leveraging LLMs. The method involves generating an exercise book to evaluate student models and providing customized training exercises based on the specific learning needs of each student model. Experimental results demonstrate that this approach achieves higher accuracy than LLMs (such as GPT-3 and PaLM [5]) across three different benchmarks while utilizing significantly fewer parameters.

When LangChain is a newly proposed technology, there is not a paper that compares ChatGPT and LangChain for mathematics problems in detail to the best of our knowledge. In this paper, we propose a categorization of mathematics question types and the corresponding errors, and use prompts to analyze the comparison of ChatGPT (GPT-3.5) and LangChain.

3 METHODOLOGY

Prompt analysis is the main methodology of this study which refers to the process of examining and understanding a given prompt or question in various contexts, often in the context of natural language processing (NLP) tasks or AI models. Since the context is mathematics in this study, we present a prompt analysis of ChatGPT and LangChain by using different mathematical questions.

In this section, we begin by introducing and categorizing different question types, as well as addressing the errors that occurred in these questions. Subsequently, we present the questions used in the prompt analysis along with their results.

3.1 Categorization of Question Types and Error Types

We posed two types of questions to ChatGPT and LangChain: word problems and numerical computational problems. Word problems are typically considered as textual problems that can be resolved by utilizing mathematical concepts, rules, or techniques [19]. We use numerical computational problems to describe problems that specifically require performing numerical calculations or computations.

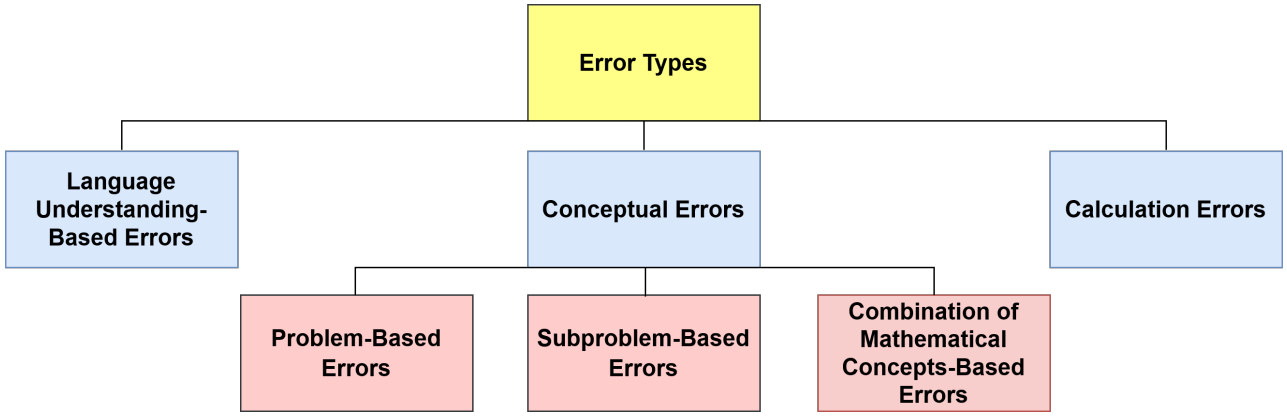


Figure 1: Categorization of Error Types.

We categorize the error types as language understanding-based errors, conceptual errors, and calculation errors. Language understanding-based errors in this paper refer to errors caused by a lack of understanding of the problem. These errors are resolved by rewriting or rephrasing the problem.

Conceptual errors are used to describe the errors related with applying the wrong theorem or formula in problem solving [7, 23]. We also categorize conceptual errors into three sub-categories: problem-based, subproblem-based and combination of mathematical concepts-based errors. Problem-based conceptual errors indicate a mistake made in applying the theorem or formula to the solution as a whole. It can also be caused by a lack of knowledge of ChatGPT or LLMMathChain. Subproblem-based conceptual errors refer to errors made specifically in the solution of sub-parts of the problem. Combination of mathematical concepts-based errors occur due to a lack of correct utilization of a combination of different mathematical topics. Calculation errors involve mistakes in performing arithmetic operations. Our categorization of error types is shown in Figure 1.

3.2 Prompt Analysis

In this subsection, we introduce the questions used in our prompt analysis, along with their corresponding outcomes. Figure 2 provides an overview of the questions employed in the prompt analysis.

Question 1, which is a word problem, was answered correctly by ChatGPT as 6, while it was answered incorrectly by LLMMathChain. LLMMathChain made the following evaluation: 5 - 2 + 4, which was thought to us the indication a language understanding problem. We rewrote the question as Question 1' to support our idea. In that case, LLMMathChain understood the question and responded correctly as ChatGPT did.

Question 2 is also a word problem with the correct answer of 14. ChatGPT provided answers of 8, -8, 14, and 8, respectively. Although ChatGPT correctly constructed the equation, it made a calculation error, resulting in incorrect answers in some attempts. On the other hand, LLMMathChain had a syntax error, which can be attributed to a language understanding-based error. We think that there might be an issue somewhere when using the models employed in LLMMathChain.

The type of Question 3 is a word problem, and the correct answer is 66. Initially, ChatGPT provided an incorrect answer because it couldn't formulate the equation of the solution accurately, resulting in a problem-based error. When we asked the question again, ChatGPT provided another incorrect answer. In the third attempt, ChatGPT was able to construct an equation that was close to the correct one, with a result of 65. However, the answer still remained incorrect. Similarly, LLMMathChain also provided incorrect answers due to a conceptual error related to the problem itself. The difference between ChatGPT and LLMMathChain was that LLMMathChain consistently gave the same incorrect answer.

Question 4 is a word problem with the correct answer of 7. ChatGPT correctly answered this question. However, LLMMathChain performed the following evaluation: $(4*(20-3))/2+(20/4)**(1/2)$ which resulted in an incorrect answer due to a subproblem-based error, which is a subcategory of conceptual error. Question 5 is yet another word problem with subproblems, and the correct answer is 82. Although ChatGPT gave the correct answer, LLMMathChain did the following equation (52/2+52/3+2)+(40+(52/2+52/3+2)) and responded to the question wrongly. This issue is similar to Question 4, where LLMMathChain's incorrect answer was also attributed to an error in managing the subproblems.

Question 6 is a numerical computational problem which covers two different topics of mathematics, and a method needs to be applied while solving this question. The two different topics should be considered together. First, it is necessary to determine whether the content inside the absolute value is positive or negative. After this determination, the part inside the absolute value should come out as a function, not numerically. However, ChatGPT calculates the content inside the absolute value numerically first, which leads to an error. We define this type of error as errors based on the combination of mathematical concepts. The same situation exists here as well in ChatGPT. In short, Question 6 was answered incorrectly by both ChatGPT and LLMMathChain because they do not know how to combine different topics in a problem solution. However, this knowledge can be taught.

Question 7 is another numerical computational problem with the correct answer of -11. It involves different mathematical topics, namely absolute value, and derivative. In this case, ChatGPT

Question 1	Sue has 5 apples and Jennifer has 4 apples. Sue gives 2 apples to Jennifer. How many apples will Jennifer have?
Question 1′	Sue has 5 apples and Jennifer has 4 apples. Sue gives 2 apples to Jennifer. How many apples will Sue have?
Question 2	Jennifer can complete the whole task in 3x days, and Ben can complete the whole task in x/2 days. If Jennifer and Ben can together complete the whole task in 12 days, how many days would it take for Ben to complete the same task alone?
Question 3	In a market, when you purchase 5 toys, you receive an additional 2 free toys. A person who has bought a total of 92 toys from the market, how many of these toys did they buy by paying for them?
Question 4	Sue and April, two sisters, have a total of 20 apples. The number of Sue's apples is four times that of April's. If Sue gives Aria half of her apples minus three, and April gives Aria the square root of her apples, how many apples will Aria have?
Question 5	The total number of students in a classroom is 52. The students sit in pairs on 5 desks and in groups of three on the other desks. The number of desks in the classroom plus two is equal to Alia's age. Amy is 40 years older than Alia. What is the sum of Amy and Alia's ages?
Question 6	If $y = \|4x - 5\| + x^2$, for $x = -1$, $\frac{dy}{dx} = ?$
Question 7	If $f(x) = \|x^3 + x^2 + 3x - 1\|$ then $f'(-2) = ?$

Figure 2: Questions in the Prompt Analysis.

provided the correct answer. However, LLMMathChain encountered the same issue as in Question 6 and answered the question incorrectly due to errors based on the combination of mathematical concepts.

Table 1 provides a summary of the questions that we asked, and the results obtained from ChatGPT and LLMMathChain. ChatGPT was more successful than LLMMathChain in these questions. Table 2 displays the match between question types and error types. As can be seen from the table, ChatGPT did not exhibit any language-understanding-based or subproblem-based errors in our questions. However, LLMMathChain experienced errors of all types except for calculation errors.

4 DISCUSSION

In mathematics, the answers are typically deterministic, meaning there's a specific correct answer. Hence, the mathematics softwares must provide consistent and accurate responses to prompts. However, since LMs are statistical in nature, we can change the methods we use to reach a solution. These changes can cause the answers to vary each time, which goes against the nature of mathematics. Similarly, ChatGPT, by design, does not always give the same answer to the same question. It's based on a Transformer model which is designed to generate diverse and creative outputs. This is particularly useful for tasks such as writing a story or generating a conversation, where generative AI is the primary consideration. When it comes to mathematics questions, the model doesn't solve them in the same manner as a calculator. Instead, it generates responses based on patterns it learned during training. So, while it often provides correct answers to mathematical questions, it might not always be consistent because it does not "understand" mathematics in the way humans do. Its responses are based on patterns rather than performing calculations.

Figure 3 represents the combination of Natural Language Understanding (NLU) and Natural Language Generation (NLG) in an LLM. During the NLU stage, expressions written as word problems or numerical computational problems are transformed into a form that can be solved using mathematical methods. NLU is responsible for converting the problem into symbolic notation. Afterwards, NLG, with the help of extensive training data, should be able to apply the solutions it has learned to the problems expressed in the NLU stage. To overcome the problem of non-deterministic features of LLMs like ChatGPT, it requires training LLMs on diverse mathematical datasets spanning a large number of questions from different mathematics topics.

Reinforcement learning (RL) is a branch of machine learning concerned with training agents to make a series of decisions within an environment. Its goal is to maximize the overall reward they receive. It draws inspiration from the way humans and animals learn by engaging in trial-and-error interactions with their surroundings. ChatGPT benefits from reinforcement learning in its reward model called Reinforcement Learning from Human Feedback (RLHF) [24]. ChatGPT's initial training is done using supervised fine-tuning, where human AI trainers provide conversations and model responses via RLHF. The advantage of this approach is to teach the LLM with the situations where it has not encountered until that time. The AI trainer ranks different model-generated responses based on their quality and may change the knowledge or

Table 1: Question Types and the Result Comparison

Questions	Question Type	ChatGPT	LLMMathChain
Question 1	Word Problem	Correct	Wrong
Question 1´	Word Problem	Correct	Correct
Question 2	Word Problem	Wrong	Wrong
Question 3	Word Problem	Wrong	Wrong
Question 4	Word Problem	Correct	Wrong
Question 5	Word Problem	Correct	Wrong
Question 6	Numerical Comp. P.	Wrong	Wrong
Question 7	Numerical Comp. P.	Correct	Wrong

Table 2: Question-error Type Match

Error Types	ChatGPT	LLMMathChain
Language Understanding-Based Error	-	Questions 1 & 2
Problem-Based Error	Question 3	Question 3
Subproblem-Based Error	-	Questions 4 & 5
Calculation Error	Question 2	-
Combination of Math. Concepts-Based Error	Question 6	Questions 6 & 7

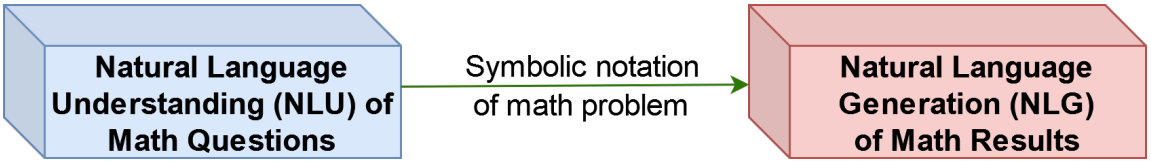


Figure 3: Current LLM Approach for Mathematics Problems.

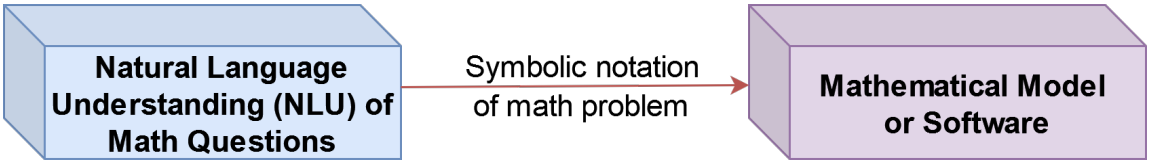


Figure 4: Current LangChain Approach for Mathematics Problems.

beliefs. These rankings serve as reward signals to LLM for optimizing the expected output. Collective expertise of humans generally helps ChatGPT to gain experience positively. But scientific areas like mathematics is not the case [1] because mathematics is not a social area where beliefs or truths must change the facts. It is a scientific area that techniques of solving a question is algorithmic and these algorithms should be learnt by some amount of similar question-solution pairs. This is the way a student works for his exams to solve lots of questions with varieties related to a specific topic.

Figure 4 summarizes how LLMMathChain operates to solve mathematical problems. When compared to Figure 3, LangChain aims to achieve more successful results by using mathematical software or a model specifically trained for mathematics. Hence, the disadvantage of LLMs can be mitigated by leveraging the deterministic nature of mathematics. LangChain allows you to use different LLMs served by OpenAI or Hugging Face endpoints with various datasets or softwares in combination to improve the performance of the system. For example, LLMMathChain could be used with LLM named as text-davinci that is used by ChatGPT-3.5. Hence, the question answering capabilities of ChatGPT like syntactic or semantical inferences could be served while the additional mathematical knowledge could be infused to a mathematics specific software or model.

There are some efforts to link LLMs with actual mathematics softwares like GPT and Wolfram Alpha3 [3]. This is a demonstration of a conversational agent using OpenAI GPT-3.5 and LangChain powered with WolframAlpha. The addition of community accepted softwares usage with ChatGPT will help the solution of mathematics problems in a more accurate and deterministic manner.

Mathematics is often thought of as the language of logic. Mathematics provides a precise and formal system of symbols, notations

[3]https://huggingface.co/spaces/JavaFXpert/Chat-GPT-LangChain

and rules that enable logical thinking and precise communication. It allows us to express and manipulate abstract concepts, formulate logical arguments, and draw conclusions based on logical deductions. While ML is an inductive reasoning-based approach, it is the difficulty of LLMs to handle the problems in a deductive manner. As logical systems such as propositional logic or first-order logic have a way of logical consequences for solution, these consequences must also be learnt by LLMs with a vast amount of similar deduction-based problems. LangChain is expected to be capable of solving these problems by time in combination with GPT.

5 CONCLUSION

In this study, an analysis and comparison of solving different mathematics problems, particularly word problems and numerical computational problems, have been presented using state-of-the-art LLMs and LangChain. The process of learning mathematics with LLMs differs significantly from the nature of mathematics itself. Moreover, the generative and statistical nature of LLMs is not inherently well-suited for deterministic fields like mathematics. However, it is projected that with novel approaches such as LangChain, the accuracy of mathematical question answering can be improved over time. Therefore, it is evident that by nourishing the success of LLMs in language comprehension aligned with the nature of mathematics, a successful mathematics assistant can be developed.

In future studies, all mathematics models and softwares that can be adapted to LangChain will be utilized with current LLMs, and the effectiveness of the Chain of Thought (CoT) method for unanswered questions will be determined. Additionally, considering the utilization of tree-based methods, particularly decision trees for explainability in machine learning, efforts will be made to investigate the explainability of the mathematics question answering method within the mathematics chain.

REFERENCES

[1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? arXiv:2303.12767 [cs.CL]

[2] Daniel Bobrow. 1964. Natural language input for a computer problem solving system. (1964).

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[4] Ting-Rui Chiang and Yun-Nung Chen. 2018. Semantically-aligned equation generation for solving and reasoning math word problems. *arXiv preprint arXiv:1811.00720* (2018).

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[6] Charles R Fletcher. 1985. Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, & Computers* 17, 5 (1985), 565–571.

[7] James Hiebert. 2013. *Conceptual and procedural knowledge: The case of mathematics*. Routledge.

[8] Danqing Huang, Jing Liu, Chin-Yew Lin, and Jian Yin. 2018. Neural math word problem solver with reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*. 213–223.

[9] Danqing Huang, Shuming Shi, Chin-Yew Lin, and Jian Yin. 2017. Learning Fine-Grained Expressions to Solve Math Word Problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 805–814. https://doi.org/10.18653/v1/D17-1084

[10] Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kalyan. 2023. Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation. *CoRR* abs/2305.14386 (2023). https://doi.org/10.48550/arXiv.2305.14386 arXiv:2305.14386

[11] Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. MWP-BERT: Numeracy-Augmented Pre-training for Math Word Problem Solving. arXiv:2107.13435 [cs.AI]

[12] Qianying Liu, Wenyv Guan, Sujian Li, and Daisuke Kawahara. 2019. Tree-structured Decoding for Solving Math Word Problems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2370–2379. https://doi.org/10.18653/v1/D19-1241

[13] Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023. A Symbolic Framework for Systematic Evaluation of Mathematical Reasoning with Transformers. arXiv:2305.12563 [cs.CL]

[14] Anirban Mukherjee and Utpal Garain. 2008. A review of methods for automatic understanding of natural language mathematical problems. *Artificial Intelligence Review* 29 (2008), 93–122.

[15] Alexander Scarlatos and Andrew Lan. 2023. Tree-Based Representation and Generation of Natural and Mathematical Language. arXiv:2302.07974 [cs.CL]

[16] Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). arXiv:2302.13814 [cs.CL]

[17] James R Slagle. 1965. Experiments with a deductive question-answering program. *Commun. ACM* 8, 12 (1965), 792–798.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Lieven Verschaffel, Stanislaw Schukajlow, Jon Star, and Wim Van Dooren. 2020. Word problems in mathematics education: A survey. *ZDM* 52 (2020), 1–16.

[20] Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. 2018. Translating a Math Word Problem to a Expression Tree. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1064–1069. https://doi.org/10.18653/v1/D18-1132

[21] Lei Wang, Dongxiang Zhang, Lianli Gao, Jingkuan Song, Long Guo, and Heng Tao Shen. 2018. Mathdqn: Solving arithmetic word problems via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[22] Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep Neural Solver for Math Word Problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 845–854. https://doi.org/10.18653/v1/D17-1088

[23] Widodo Winarso and Toheri Toheri. 2021. An analysis of students' error in learning mathematical problem solving: The perspective of David Kolb's theory. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, 1 (2021), 139–150.

[24] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

[25] Zhipeng Xie and Shichao Sun. 2019. A Goal-Driven Tree-Structured Neural Model for Math Word Problems. In *Ijcai*. 5299–5305.

[26] Bakman Yefim. 2007. Robust Understanding of Word Problems with Extraneous Information. arXiv:math/0701393 [math.GM]

[27] Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai, and Heng Tao Shen. 2020. The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2287–2305. https://doi.org/10.1109/TPAMI.2019.2914054

View publication stats

IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Vector-Based Chat with Multi Pdf Using Langchain and AI

Amruta Bhosale Department of Computer Science and Technology Usha Mittal Institute of Technology S.N.D.T. Women's UniversityMumbai 400049

Saniya Shaikh Department of Computer Science and Technology Usha Mittal Institute of Technology S.N.D.T. Women's UniversityMumbai 400049

Shivani Sharma Department of Computer Science and Technology Usha Mittal Institute of Technology S.N.D.T. Women's UniversityMumbai 400049

Prof. Prajakta Gotarne

Department of Computer Science and Technology Usha Mittal Institute of Technology S.N.D.T. Women's University Mumbai 400049

***Abstract*— The project presents the development of an in-novative chatbot system that integrates with PDF documents, enhancing users' ability to extract information through natural language queries. The primary goal is to address the growing demand for efficient information retrieval from textual docu- ments, particularly in educational and professional contexts. The proposed system aims to provide a user friendly system for extracting relevant information from PDFs swiftly and accurately. To achieve this, the system leverages several key technologies. Streamlit is used for the user interface, providing an intuitive and interactive platform for users to interact with the chatbot. PyPDF2 is employed for PDF parsing, allowing the system to extract text from uploaded PDF documents. Langchain is utilized for text processing and embeddings, enabling the system to process and index the extracted text ef- ficiently. Google Generative AI is integrated for conversational capabilities, enabling the chatbot to understand user queries and provide relevant answers. Additionally, FAISS is used for similarity search, allowing for fast and accurate retrieval of information from indexed PDF content. The system's work ow involves users uploading PDF documents, from which text is extracted, processed, and indexed for efficient retrieval. The chatbot, powered by Google Generative AI, interacts with users, interprets their queries, and provides relevant answers based on the indexed PDF content. The project's main objective is to provide an interactive and easy to use user experience, making it easy for users to access and utilize information from PDF documents effectively. Future enhancements may include support for more complex queries, integration with other document formats, and improved user interaction features. Overall, this project contributes to the advancement of natural language processing and information retrieval systems, bene- fiting users in various domains requiring efficient document analysis and information extraction.

***Index Terms*— Langchain, Google Generative AI Embedding, FAISS Indexing, Streamlit, Python.**

I. INTRODUCTION

In today's digital era, there is infinite information available in textual documents has led to an increasing demand for efficient methods of information retrieval. PDF documents, in particular, are widely used for storing and sharing information, making them a valuable source of knowledge. However, extracting relevant information from PDFs can be challenging, especially when dealing with large volumes of text.

To work on this problem, we propose the development of a chatbot system that integrates with PDF documents, allowing users to ask questions based on the content of uploaded PDFs. This system aims to provide a user- friendly and efficient solution for extracting information from PDFs through natural language queries.

The motivation behind this project stems from the need for tools that simplify the process of extracting information from PDFs, particularly in educational and professional settings where access to accurate and timely information is crucial. By developing a chatbot system that can effectively extract information from PDFs, we hope to provide users with a valuable tool for accessing and utilizing information from textual documents moreefficiently.

This project's research problem lies in the lack of user- friendly tools for extracting relevant information from PDFs quickly and accurately. Existing methods often require manual effort and are not suitable for handling large volumes of text. The proposed chatbot system aims to overcome these limitations by providing an automated

and intuitive solution for extracting information from PDFs through natural language queries. In this paper, we present the methodology and implementation of the proposed chatbot system. We describe the technologies used, the system's workflow, and the expected outcomes. We also discuss potential future enhancements and the impact of the project on the field of natural language processing and information retrieval.

Overall, this project represents a significant step towards improving the efficiency of information retrieval from textual documents, particularly PDFs. By developing a chatbot system that can extract information from PDFs through natural language queries, we aim to provide users with a valuable tool for accessing and utilizing information more effectively.

II. LITERATURE SURVEY

A. Overinformative Question Answering by Humans and Machines

Authors: Polina Tsvilodub, Michael Franke, Robert Hawkins) Noah D. Goodman, Year:2023. This paper explores methods for question answering (QA) over PDF documents. It likely discusses techniques for extracting information from PDF files and utilizing it to answer questions posed by users. This could involve approaches such as text extraction, natural language processing (NLP), and document understanding to enable effective QA systems specifically tailored for PDF documents.

B. Information Retrieval Meets Large Language Models

Authors: Qingyao AI, Ting BAI, Zhao CAOc, Yi CHANG, Jiawei CHEN, Zhumin CHEN, Zhiyong CHENGg, Shoubin DONG, Zhicheng DOU, Fuli FENG, Shen GAO , Jiafeng GUO, Xiangnan, Yanyan LANa, Chenliang LI, Yiqun LIU, Ziyu LYU, Weizhi MA, Jun MA, Zhaochun REN, Pengjie REN, Zhiqiang WANG, Mingwen WANG, Ji-Rong WEN, Le WU, Xin XIN, Jun XU, Dawei YIN, Peng ZHANG, Fan ZHANG, Weinan ZHANG, Min ZHANG, Xiaofei ZHU, Year:2023.This paper likely investigates the application of large language models (LLMs) for information retrieval tasks. It may discuss how LLMs, such as transformer-based models, can be leveraged to improve the effectiveness and efficiency of information retrieval systems. This paper involve techniques like fine-tuning pre-trained models on retrieval-specific datasets, designing architectures optimized for retrieval tasks, and exploring methods for integrating LLMs into existing retrieval frameworks.

C. Conversational Agents: Theory and Applications

Authors: Mattias Wahde and Marco Virgolin, Year:2022. This paper focuses on conversational agents designed for document understanding tasks. It discuss how conversational agents, such as chatbots or virtual assistants, can assist users in understanding and extracting information from documents. This include techniques for natural language understanding (NLU), dialogue management, and document summarization tailored for conversational interfaces. The paper likely explores how these agents can enhance user interaction and productivity in document-centric tasks.

D. Vector Space Model: An Information Retrieval System

Authors: Vaibhav Kant Singh, Vinay Kumar Singh, Year:2022. This paper introduce and explain the concept of the vector space model for information retrieval systems. It discusses how documents and queries can be represented as vectors in a multi-dimensional space, based on similarity measures. This model serves as a fundamental framework for understanding and implementing information retrieval systems.

E. Creating Large Language Model Applications Utilizing LangChain: Primer on Developing LLM Apps Fast

Oguzhan Topsakal, T. Cetin Akinci, Year:2023. This paper discussed about the concept of Langchain and how it could be used to develop LLM applications. It likely serves as an introductory guide for developers, offering insights into leveraging LangChain a framework or toolset designed for building LLM applications. The paper discuss best practices, methodologies, and practical examples to facilitate the development process and enable the creation of effective LLM applications.

III. RELATED WORK

A. Existing System - "An AI-Driven Interactive Chatbot: A Well Trained Chatbot that Communicates with the Users and Reduces the Manual Interaction", Vol.9, Feb 2024

The existing system is an AI-driven interactive chatbot designed to provide easy conversation and reduce manual interaction. Here are the key points of the existing system:

- **Purpose:** Provide a chatbot interface for users to inquire about details regarding the institute.
- **Functionality:** Users interact with the chatbot to ask questions. It reduces the need for direct interaction with a salesperson. Increases accessibility for users to get information about the institute. Uses LLM (Large Language Models) generative AI for conversations.
- **Benefits:** Users can inquire about the institute at any time. Salespersons are relieved from continuous inquiries. Increased accessibility and convenience for users.
- **Technologies:** LLM (Large Language Models) generative AI. Chatbot interface.

B. Proposed System

The proposed system seems to be an extension or modification of the existing system, now incorporating features to handle multiple PDF uploads and questions related to those PDFs.

- **System Architecture Overview:**

The components of Proposed System are as follows:

User Interface: The web-based interface allows users to

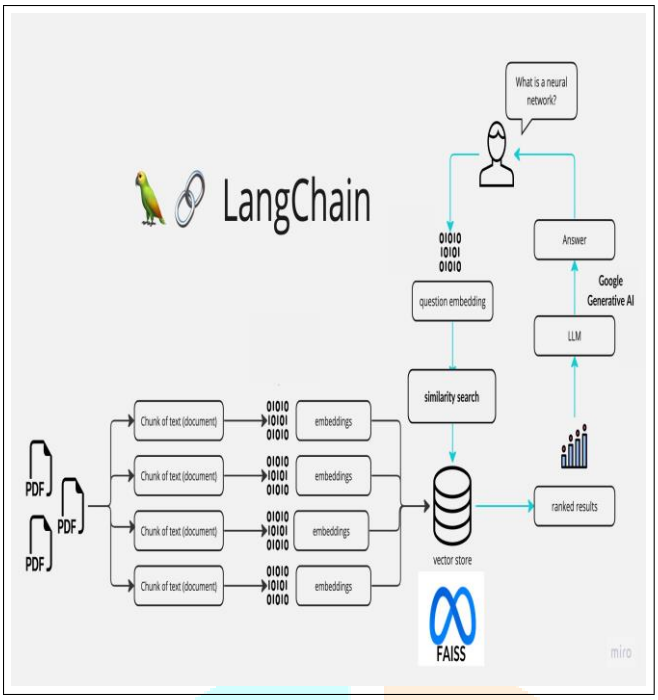


Fig. 1. Proposed System Architecture

upload PDF files and submit queries. Users can interact with the chatbot to ask questions and receive answers based on the contents of the uploaded PDF files. The interface provides a user-friendly experience, guiding users through the process of uploading files and querying information.

PDF Processing Module: Upon file upload, the system uses the PyPDF2 library to extract text from each PDF document. The extracted text is then processed to remove any unnecessary characters or formatting, ensuring that the text is clean and ready for further analysis. The processed text is split into manageable chunks, typically sentences or paragraphs, to facilitate efficient processing and analysis.

Text Embedding Module: The Langchain's Google Generative AI Embeddings model is used to generate embeddings for each text chunk. The model converts each text chunk into a numerical representation, capturing the semantic meaning of the text. These embeddings are high-dimensional vectors that represent the text's context, allowing for more accurate information retrieval and similarity calculations.

Indexing and Retrieval Module: The embeddings for text chunks are indexed using the FAISS library, which efficiently stores and retrieves embeddings for similarity search. When a user submits a query, the system calculates the embeddings for the query and uses FAISS to retrieve the most relevant text chunks based on similarity scores. The system may use advanced indexing techniques to speed up retrieval and optimize resource utilization.

Conversational AI Module: The system utilizes Google Generative AI for conversational capabilities, enabling the chatbot to understand user queries and provide responses based on the indexed PDF content. The chatbot uses natural language processing techniques to interpret user queries and retrieve relevant information from the indexed PDF files. The chatbot can engage in a natural language conversation with users, providing answers to their queries and offering suggestions for further exploration.

Real-time Updates and User Feedback: The system provides real-time updates to users as it processes and retrieves information, ensuring a seamless and interactive user experience. Users can provide feedback on the relevance and accuracy of retrieved information, enabling continuous improvement of the system. The system may use this feedback to update its indexing and retrieval algorithms, improving the quality of future responses.

- **User Workflow:**

Step 1: Upload PDF Files: Users upload one or more PDF files through the web interface.

Step 2: Text Extraction and Embedding: The system extracts text from the uploaded PDF files and generates embeddings for text chunks.

Step 3: Indexing and Retrieval: Text embeddings are indexed using FAISS for efficient retrieval. When a user submits a query, the system matches the query against indexed embeddings to retrieve relevant information.

Step 4: Conversational Interaction: The chatbot interacts with the user, providing answers to queries based on the retrieved information from PDF documents.

Step 5: Display Results: The chatbot displays the relevant information to the user through the web interface, allowing for further interaction and refinement of queries.

- **Additional Features:**

- Real-time Updates: The system provides real-time updates to users as it processes and retrieves information, ensuring a seamless and interactive user experience.
- Query Suggestions: Based on the user's query history and document content, the system provides suggestions for refining queries to improve search results.
- Document Summarization: Optionally, the system can provide summaries of PDF documents to users, highlighting key information relevant to their queries.

By following this proposed methodology, the system aims to provide users with a robust and efficient tool for information retrieval from large PDF files in a multipdf chat environment, enhancing the overall user experience and usability of the system.

C. Gaps between the Existing System and the Proposed System

1. **Functionality:**

Existing: Primarily focused on institute-related inquiries.
Proposed: Extends functionality to include handling questions based on uploaded PDF content.

2. **Data Input:**

Existing: Takes user queries directly.
Proposed: Introduces PDF uploads as additional data

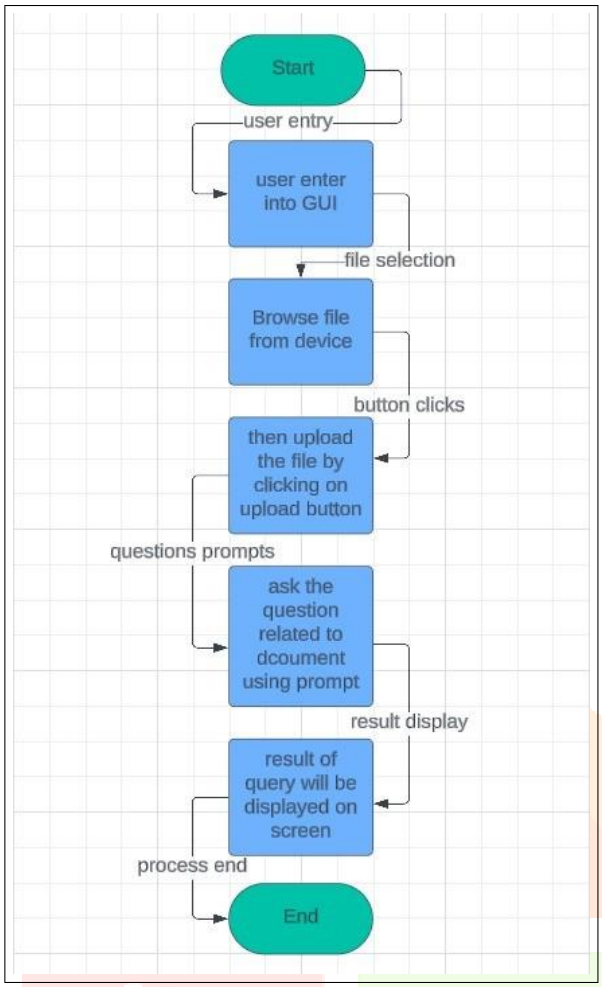


Fig. 2. User workflow

sources.

3. Processing:

Existing: Processes user queries based on predefined knowledge.

Proposed: Requires text extraction from PDFs and processing of questions based on this new data source.

4. Interaction Flow:

Existing: Linear interaction with the chatbot for general queries.

Proposed: Introduces an interactive system where user can ask PDF related question.

IV. RESULT AND ANALYSIS

In the realms of information retrieval and natural language processing, several approaches and technologies have been developed to address the challenge of information retrieval from textual documents, including PDFs. These approaches vary in complexity and efficiency, with some focusing on manual extraction methods and others on automated techniques. In this section, we review some relevant work in this area and discuss the technologies, algorithms, and methodologies used.

1. PDF Parsing Technologies: One of the key components of extracting information from PDFs is the ability to parse PDF files and extract text. Several libraries and tools exist for this purpose, including PyPDF2, which is used in this project. PyPDF2 allows for the extraction of text from PDF files, making it a valuable tool for processing PDF documents in an automated manner.

- Importing the Library: First, the PyPDF2 library is imported at the beginning of the code: - Extracting Text from PDFs: When the user uploads PDF files, the ‘get pdf text‘ function is called to extract text from each PDF document. This function iterates through each page of the PDF and extracts the text using PyPDF2.

- Splitting Text into Chunks: After extracting the text from PDFs, the ‘get text chunks‘ function is used to split the text into manageable chunks. It uses Langchain’s RecursiveCharacterTextSplitter with a specified chunk size and overlap. The ‘chunk size‘ parameter determines the maximum size of each chunk, and the ‘chunk overlap‘ parameter specifies how much overlap there is between consecutive chunks. These parameters are chosen based on the desired balance between chunk size and overlap for efficient processing and analysis of text chunks.

2. Text Processing and Embeddings: Text processing and embeddings play a crucial role in understanding the content of textual documents. In this project, the Langchain Google Generative AI Embeddings model is used for text processing and embeddings. Langchain provides a range of functionalities for processing text, including splitting text into chunks and generating embeddings for each chunk. These embeddings are used to index the text for efficient retrieval, as well as to provide context for natural language understanding. Taking this following text as an example of chunk: "The quick brown fox jumps over the lazy dog."

- Tokenization: Before generating embeddings, the text is tokenized into individual words or subwords. Each token is then converted into a numerical representation (e.g., an index or a vector) using Langchain’s model.

Tokenize the text into individual words: ["The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog."]

- Embedding Lookup: The numerical representations of tokens are used to look up embedding vectors from a pre-trained embedding matrix. This matrix contains embedding vectors for all tokens in the model’s vocabulary. Assume we have a pre-trained embedding matrix with embeddings for each token. Let’s say the embeddings are 3-dimensional vectors: "The": [0.1, 0.2, 0.3] "quick": [0.4, 0.5, 0.6] ... "dog.": [0.7, 0.8, 0.9]

Look up the embeddings for each token and create a list of embeddings: [[0.1, 0.2, 0.3], [0.4, 0.5, 0.6], ..., [0.7, 0.8, 0.9]]

- Aggregation: The embedding vectors for individual tokens are aggregated to form a single vector representation for the entire text chunk. This aggregation step can involve techniques like averaging, pooling, or attention mechanisms to capture the overall meaning of

the text chunk. Average the embeddings to get a single vector representation: Aggregated embedding = [(0.1 + 0.4 + ... + 0.7) / 9, (0.2 + 0.5 + ... + 0.8) / 9, (0.3 + 0.6 + ... + 0.9) / 9]

Aggregated embedding = [0.4, 0.5, 0.6]

- Normalization: The aggregated embedding vector is often normalized to have a unit length. This normalization step helps in training and comparing embeddings. Normalize the aggregated embedding to have unit a length: Norm = sqrt(0.77) = 0.88 Normalized embedding = [0.4 / 0.88, 0.5 / 0.88, 0.6 / 0.88] = [0.45, 0.57, 0.68]

Output: The final normalized embedding vector represents the semantic meaning of the text chunk in a high-dimensional space. Similar text chunks are expected to have similar embeddings, allowing for effective information retrieval and similarity calculations. The final normalized embedding [0.45, 0.57, 0.68] represents the semantic meaning of the text chunk "The quick brown fox jumps over the lazy dog." as a 3-dimensional embedding space. Overall, the Text Embedding Module uses Langchain's Google Generative AI Embeddings model to convert text chunks into embeddings, which capture the semantic meaning of the text and enable efficient information retrieval from PDF documents.

3. Similarity Search: FAISS is used in this project for similarity search, enabling fast and accurate retrieval of information from indexed PDF content. FAISS is a library for efficient similarity search and clustering of dense vectors, making it well-suited for indexing and searching embeddings generated from text chunks in PDF documents.

• Indexing :

When the text chunks are extracted from PDF documents and embedded using Langchain's Google Generative AI Embeddings model, FAISS is used to create an index of these embeddings. The embeddings are quantized using product quantization to reduce the dimensionality and improve search efficiency. FAISS creates inverted lists for each quantized centroid, where each list contains the IDs of text chunks that are quantized to that centroid. These inverted lists serve as the index, allowing FAISS to quickly retrieve similar text chunks based on a query.

• Retrieval :

When a user submits a query, FAISS calculates the embeddings for the query using the same process used during indexing. The query embeddings are quantized and used to look up the inverted lists in the index. FAISS retrieves the IDs of text chunks in these inverted lists and calculates the similarity scores between the query and these text chunks. The text chunks with the highest similarity scores are returned as the results of the query, allowing the system to retrieve and display the most relevant text chunks to the user.

Simplified numerical example to illustrate the indexing and retrieval process using FAISS :

Assume we have 3 text chunks (T1, T2, T3) and each is represented by a 2-dimensional embedding vector. We use product quantization to divide the 2-dimensional space into 2 subspaces (1-dimensional each). Quantization reduces the embeddings to a codebook of 2 centroids per subspace (a total of 4 centroids).

o The quantized embeddings for each text chunk are as follows:

• T1: [0.1, 0.2] -> [0, 0]
• T2: [0.3, 0.4] -> [1, 0]
• T3: [0.5, 0.6] -> [1, 1]

o FAISS creates inverted lists for each centroid:

• Centroid [0, 0]: [T1]
• Centroid [1, 0]: [T2]
• Centroid [1, 1]: [T3]

o Retrieval Process

The User submits a query with an embedding [0.2, 0.3]. The query embedding is quantized to [0, 0]. FAISS looks up the inverted list for centroid [0, 0] and retrieves text chunk T1. The system presents T1 as the most similar text chunk to the user's query.

Overall, FAISS plays a crucial role by providing efficient indexing and retrieval capabilities for the embeddings generated from PDF documents, facilitating fast and accurate information retrieval in a multiple chat environment.

4. Conversational AI: Conversational AI is used to enable the chatbot to interact with users in a natural and intuitive manner. Google Generative AI is employed in this project for conversational capabilities. It allows the chatbot to understand user queries and provide relevant answers based on the indexed PDF content. This technology enhances the user experience by providing a conversational interface for interacting with the chatbot. A numerical explanation of how the Conversational AI Module provides answers to user queries:

- Indexed PDF Content: Assume we have two PDF documents (D1, D2) with text chunks and embeddings. Each document contains several text chunks (T1, T2, T3) with corresponding embeddings. These embeddings are indexed using FAISS for efficient retrieval.

- User Query: - A user submits a query with an embedding [0.2, 0.3].

- Answer Generation: The Conversational AI Module retrieves the most relevant text chunks based on the query embedding using FAISS. It uses Google Generative AI to generate a response based on the content of the retrieved text chunks and the context of the query. The generated response is presented to the user as an answer to their query.

- Example: Suppose the closest text chunk to the query embedding is T3 from document D1, which contains the text "Data Structures and Algorithms." The module generates a response like "You can find information about Data Structures and Algorithms in the indexed PDF documents."

5. Vector Store: The vector store is a component that manages the storage and retrieval of embeddings for text chunks extracted from PDF documents. When a user uploads PDF files, the PDFs are stored in a location accessible to the backend server. The server may store the PDF files on disk, depending on the project's architecture and requirements. Each uploaded PDF file is associated with a unique identifier (e.g., file name or ID) that allows the system to retrieve the file when needed.

6. Integration with Streamlit: Streamlit is used for the user interface in this project, providing an interactive platform for users to interact with the chatbot and upload PDF documents. Streamlit allows for the easy integration of various components, such as file uploaders and chat interfaces, making it ideal for developing user-friendly applications.

A. Implementation

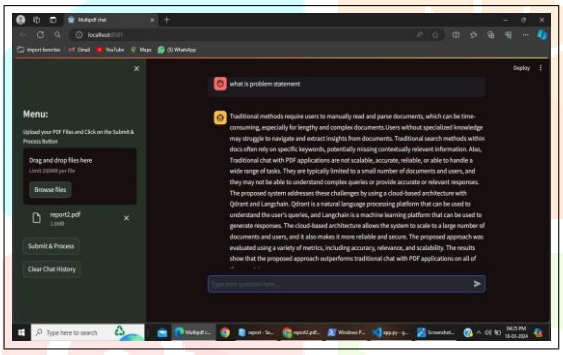


Fig. 3. Fig. 1. Upload single or Multiple Files

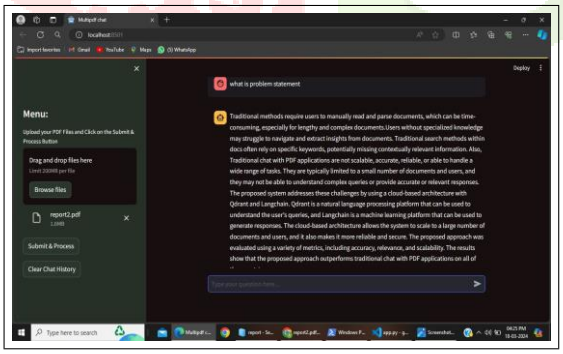


Fig. 4. output 1

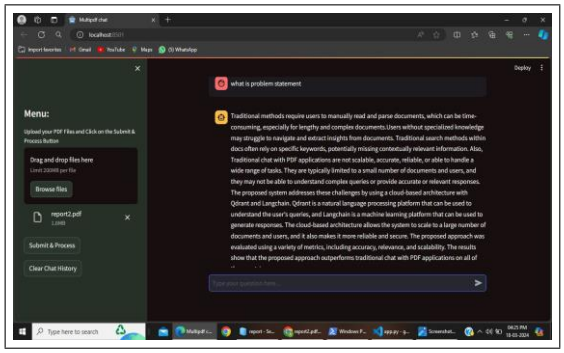


Fig. 5. Output

2

V. CONCLUSIONS

This project successfully demonstrates the development of a chat interface that facilitates question answering based on uploaded PDFs. It leverages Langchain's capabilities to create a custom question-answering work ow and integrates a large language model (LLM) for generating responses to user queries.

Streamlit provides a user-friendly web interface for uploading PDFs and interacting with the system through a chat-like interface. The key strengths of this project lie in its clear separation of functionalities, effective use of Langchain and Streamlit, and the integration of a powerful LLM for response generation. However, there's room for improvement in areas like error handling, search optimization, context management, and security best practices.

This project not only showcases the technical prowess in NLP and document analysis but also underscores the practical applications of such technology. By bridging the gap between static document content and dynamic user interactions, it opens avenues for enhanced information retrieval and user engagement.

By addressing these aspects, we developed a robust tool for exploring and gaining insights from information stored within PDF documents. This user-friendly chat interface empowers users to ask questions from the content of uploaded PDFs and receive relevant answers, fostering a more effi cient and interactive way to navigate and understand these documents.

REFERENCES

[1] Ritendu Bhattacharyya, Sharat Chandra K. Manikonda, Bharani Kumar Depuru3 "An AI-Driven Interactive Chatbot: A Well Trained Chatbot that Communicates with the Users and Reduces the Man- ual Interaction", Vol.9, Feb 2024

[2] Vaibhav Kant Singh, Vinay Kumar Singh "VECTOR SPACE MODEL: AN INFORMATION RETRIEVAL SYSTEM", July 2022

[3] Topsakal, Oguzhan and Akinci, T. Cetin "Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast", July 2023

[4] Wenhu Chen and Ming-Wei Chang and Eva Schlinger and William Wang and William W. Cohen "Open Question Answering over Tables and Text,2021

[5] Singla, Samriddhi and Eldawy, Ahmed and Diao, Tina and Mukhopadhyay, Ayan and Scudiero, Elia "Experimental Study of Big Raster and Vector Database Systems",April 2021.

[6] Qingyao Ai and Ting Bai and Zhao Cao and Yi Chang and Jiawei Chen and Zhumin Chen and Zhiyong Cheng and Shoubin Dong and Zhicheng Dou and Fuli Feng and Shen Gao and Jiafeng Guo and Xiangnan He and Yanyan Lan and Chenliang Li and Yiqun Liu and Ziyu Lyu and Weizhi Ma and Jun Ma and Zhaochun Ren and Pengjie Ren and Zhiqiang Wang and Mingwen Wang and Ji- Rong Wen and Le Wu and Xin Xin and Jun Xu and Dawei Yin and Peng Zhang and Fan Zhang and Weinan Zhang and Min Zhang and Xiaofei Zhu "Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community",2023.

[7] Polina Tsvilodub and Michael Franke and Robert D. Hawkins and Noah D. Goodman "Overinformative Question Answering by Humans and Machines", the Year 2023

[8] Mattias Wahde,Marco Virgolin "Conversational Agents: Theory and Applications ", Feb 2022

Automating Customer Service using LangChain

Building custom open-source GPT Chatbot for organizations

Keivalya Pandya
19me439@bvmengineering.ac.in
Birla Vishvakarma Mahavidyalaya, Gujarat, India

Prof. Dr. Mehfuza Holia
msholia@bvmengineering.ac.in
Birla Vishvakarma Mahavidyalaya, Gujarat, India

***Abstract*— In the digital age, the dynamics of customer service are evolving, driven by technological advancements and the integration of Large Language Models (LLMs). This research paper introduces a groundbreaking approach to automating customer service using LangChain, a custom LLM tailored for organizations. The paper explores the obsolescence of traditional customer support techniques, particularly Frequently Asked Questions (FAQs), and proposes a paradigm shift towards responsive, context-aware, and personalized customer interactions. The heart of this innovation lies in the fusion of open-source methodologies, web scraping, fine-tuning, and the seamless integration of LangChain into customer service platforms. This open-source state-of-the-art framework, presented as "Sahaay," demonstrates the ability to scale across industries and organizations, offering real-time support and query resolution. Key elements of this research encompass data collection via web scraping, the role of embeddings, the utilization of Google's Flan T5 XXL, Base and Small language models for knowledge retrieval, and the integration of the chatbot into customer service platforms. The results section provides insights into their performance and use cases, here particularly within an educational institution. This research heralds a new era in customer service, where technology is harnessed to create efficient, personalized, and responsive interactions. Sahaay, powered by LangChain, redefines the customer-company relationship, elevating customer retention, value extraction, and brand image. As organizations embrace LLMs, customer service becomes a dynamic and customer-centric ecosystem.**

Keywords*— *Customer Service Automation, Large Language Models, LangChain, Web Scraping, Context-Aware Interactions

I. Introduction

"Customer is king" is the ancient mantra reflecting the significance of customers in every business. In the digital age, where the rhythms of modern life are guided by the pulse of technology, the realm of customer service stands as the frontline of engagement between businesses and their clientele. It is the place where queries are answered, problems are resolved, and trust is forged.

This research paper brings the future of customer service, where automation, personalization, and responsiveness converge to redefine the customer-company relationship. At the heart of this transformation lies the integration of LLMs, exemplified by LangChain [1].

In the annals of customer service history, FAQs and traditional support mechanisms have long held sway. These venerable tools have dutifully served as repositories of information, attempting to address the queries and concerns of customers. However, as we stand at the cusp of a new era in customer service automation, it becomes abundantly clear that the traditional methods once hailed as revolutionary, are gradually becoming obsolete.

This paper is an invitation to envision a future where customer service is not a cost center but a wellspring of customer satisfaction and loyalty. We propose an open-source framework that can be scaled to any industry or organization to fulfill the consumer needs for support and query resolution within seconds.

For demonstration purposes, we use the information presented by Birla Vishvakarma Mahavidyalaya (BVM) Engineering College on their website *https://bvmengineering.ac.in/* as the context for our chatbot, from where it can retrieve all the information in real-time and answer to any queries that are raised by the users. Here, users can be anyone ranging from prospective students, current students who intend to get information from the Notice Board, researchers who wish to search for their potential research guide, and so on. The applications are endless.

II. Literature Survey

S. Kim (2023) et al addresses the challenge of deploying resource-intensive large neural models, such as Transformers, for information retrieval (IR) while maintaining efficiency. Experimental results on MSMARCO benchmarks demonstrate the effectiveness of this approach, achieving successful distillation of both dual-encoder and cross-encoder teacher models into smaller, 1/10th size asymmetric students while retaining 95-97% of the teacher's performance [2]. L. Bonifacio et al (2022) highlights the recent transformation in the Information Retrieval (IR) field, propelled by the emergence of large pretrained transformer models. The MS MARCO dataset played a pivotal role in this revolution, enabling zero-shot transfer learning across various tasks [3].

This paper proposed a novel open-source approach to building LLM Chatbots using custom knowledge from the content in the website. It is unique in several ways:

1. We propose an open-source framework which is robust with the type of dataset available on the webpage or the web of links.
2. This implementation aims to compliment the use of FAQs with a more interactive and user-friendly interface.
3. We then do a comparative study of various models, their performance on the provided data relative to the expected response from the LLM.

Submitted to the 3rd International Conference on "Women in Science & Technology: Creating Sustainable Career"
28 -30 December, 2023

III. Methodology

This section covers the data collection, details about the selected model, fine-tuning, and integration with the Gradio APIs for web deployment.

A. Data Collection

To gather the necessary data for our project, we employed BeautifulSoup web scraping techniques to retrieve publicly accessible information from an organization's homepage. We observed this page is often linked with all the relevant information required for the user/visitor. This approach allowed us to collect a wide array of data, including customer service FAQs, product manuals, support forums, chat logs, associated institutions, and so on. This data further serves as the context for our LLM.

B. Embeddings

Embeddings play a pivotal role in the development of any LLM powered. They are vector representations of words or phrases in a continuous mathematical space that capture semantic and contextual information, allowing the model to understand the meaning and relationships between words, which is essential for providing meaningful responses to user queries.

We have used HuggingFace Instuct Embeddings – "hkunlp/instructor-large" a text embedding model fine-tuned for specific tasks and domains, such as classification, retrieval, clustering, and text evaluation [4]. What sets Instructor apart is its ability to generate tailored text embeddings without requiring additional fine-tuning. These embeddings are then stored using FAISS (Facebook AI Similarity Search) library that allows developers to quickly search for embeddings of multimedia documents that are similar to each other [5].

C. Language Model

We have chosen Google's Flan T5 XXL as the most appropriate language model after comparing with other Flan T5 distributions to retrieve knowledge from the *vectorspace* and *chat_history* (or memory) [6]. The model retains the context of previous messages and uses that as a reference to predict answers for the upcoming questions. This helps users to have an interactive conversation with the chatbot, instead of a monotonous and robotic one.

D. Integration with Customer Service Platforms

A simple chat window can be activated at the corner of any website which would enable users to interact with the chatbot and ask any relevant questions or doubts regarding the organization. However, for the demonstration purpose of this paper, we are using Gradio API framework [7].

IV. Results

In this section, we mention the metrics of comparison, provide comparative analysis, and use cases in association with an educational institution.

A. Evaluating the Performance of LLMs

It is relatively difficult to evaluate LangChain agents, especially when trained on large chunks of context datasets for information retrieval. Hence, the current solution for the lack of metrics is to rely on human knowledge to get a sense of how the chain/agent is performing.

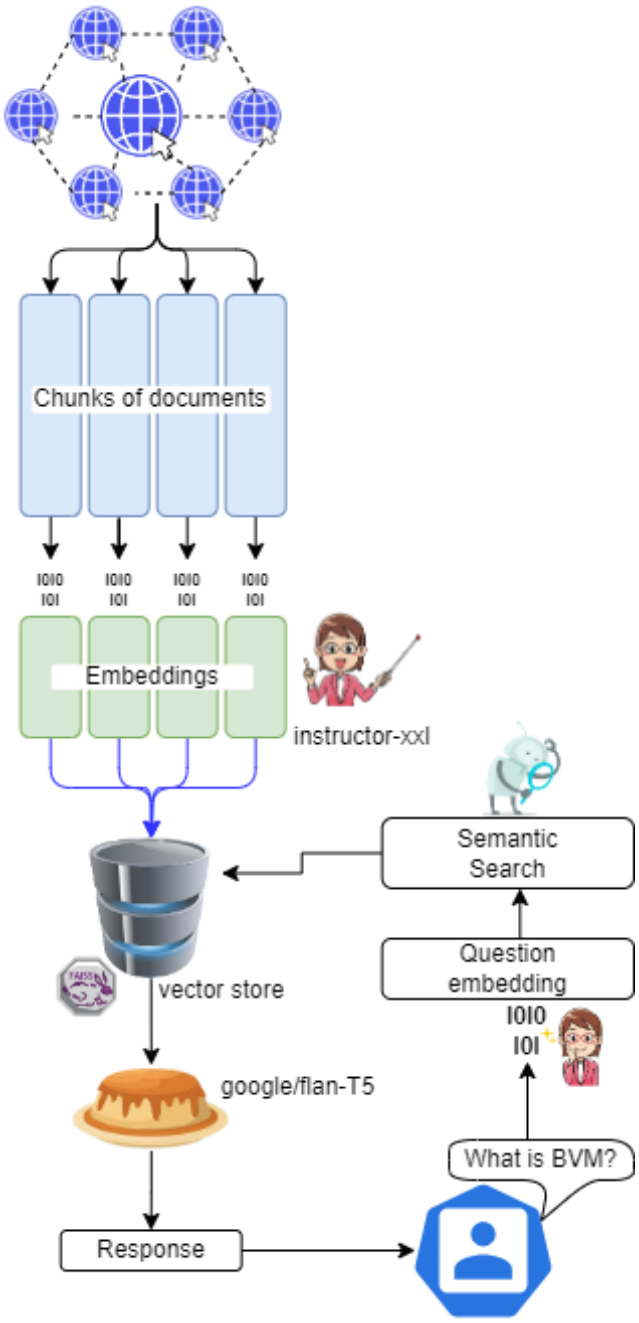


Fig. 1. Model Architecture – Sahaay

It is evident from TABLE-I, II and III that the XXL model outperforms other competitive LLMs such as BASE and SMALL.

TABLE I. Google's Flan-T5-*XXL* Performance

Metrics			Performance
Sr. No.	***Query/Prompt***	***Answer***	
1.	What is BVM?	Birla Vishvakarma Mahavidyalaya	✯✯✯✯
2.	Where is it?[a]	Vallabh Vidyanagar, Gujarat India	✯✯✯
3.	What is IEEE BVM?	Institute of Electrical and Electronics Engineers Student Branch of BVM	✯✯✯✯
4.	What is TRS BVM?	BVM Student Chapter associated with Robotics Society India	✯✯✯✯
5.	Name the annual newsletter?	Vishvakarma Magazine and Newsletter	✯✯✯✯
6.	Tell me something about ICWSTCSC	The International Conference for Women in Science and Technology Creating Sustainable Career is the 3rd International Conference happening in hybrid mode.	✯✯✯✯✯

[a.] Follow-up question

TABLE II. GOOGLE'S FLAN-T5-*BASE* PERFORMANCE

Metrics			Performance
Sr. No.	***Query/Prompt***	***Answer***	
1.	What is BVM?	BVM is the first engineering college of Gujarat State established way back in 1948.	✯✯✯✯
2.	Where is it?[b]	VV Nagar, Gujarat, India	✯✯✯
3.	What is IEEE BVM?	first engineering college of Gujarat State established way back in 1948	✯
4.	What is TRS BVM?	CVM BVM Robotics Society Of India	✯✯
5.	Name the annual newsletter?	Vishvakarma Magazine	✯✯✯✯
6.	Tell me something about ICWSTCSC	BVM 3rd Internartional Conference	✯✯✯✯

[b.] Follow-up question

TABLE III. GOOGLE'S FLAN-T5-*SMALL* PERFORMANCE

Metrics			Performance
Sr. No.	***Query/Prompt***	***Answer***	
1.	What is BVM?	BVM is the first Autonomous Engineering institute of Gujarat to obtain academic autonomy for all its UG & PG programs, from University Grant Commission (UGC).	✯✯✯
2.	Where is it?[c]	BVM Engineering College	✯
3.	What is IEEE BVM?	engineering college of Gujarat State	✯
4.	What is TRS BVM?	BVM	✯
5.	Name the annual newsletter?	Campus Publications	✯
6.	Tell me something about ICWSTCSC	ICWSTCSC 2023 PMSSS students' reporting after admission (AY: 2023-24) at BVM 3rd Internartional Conference.	✯✯✯

[c.] Follow-up question

B. *Applications*

Customer service establishes a direct connection between the customer and the company. It retains customers and extracts higher value from them. By harnessing the power of Large Language Models as shown in Fig. 2, customer service can be elevated to new heights, facilitating efficient, personalized, and responsive interactions. The LangChain fine-tuned over custom knowledge of the product, service, or organization can effectively address a wide array of customer inquiries and issues. Its ability to understand context and history empowers it to provide personalized support to customers. Automated customer service powered by LLMs is available around the clock and is also proficient in multiple languages.

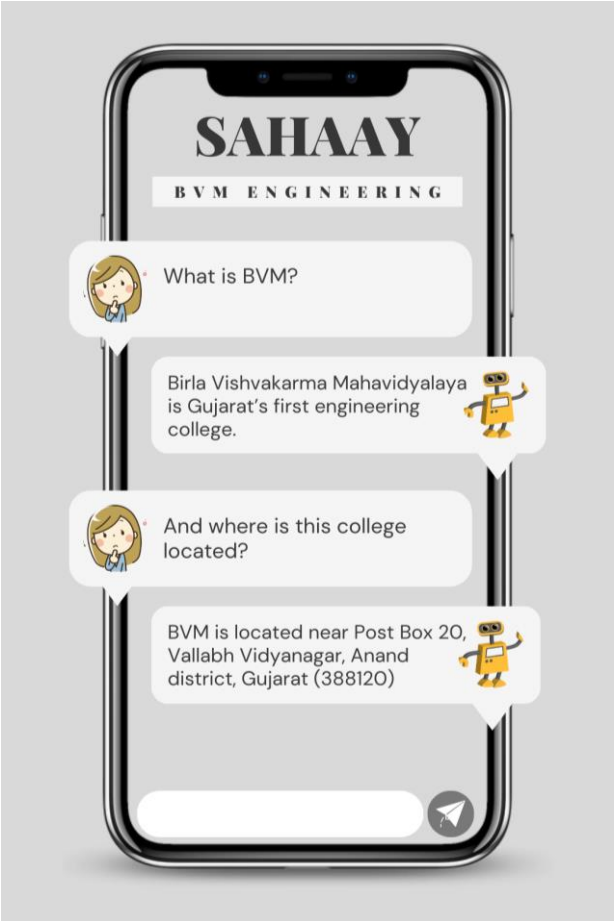


Fig. 2. User interface – Gradio framework

V. CONCLUSION

In the ever-evolving landscape of customer service, the introduction of Sahaay's innovative approach presented in this paper, using LangChain as a prime example, ushered in a new era of automation. Automating customer service using Sahaay's open-source Large Language architecture leveraging LangChain revolutionizes the customer-company relationship and CX. It enables companies to provide efficient,

personalized, and responsive support, ultimately leading to customer retention, increased customer value, and a more positive brand image. As organizations continue to leverage the capabilities of LLMs, the landscape of customer service is evolving into a more dynamic and customer-centric ecosystem.

This paper demonstrates a comparison between various model performances and evaluates them on the basis of the quality of response generated. We have compared GOOGLE/FLAN-T5-XXL with GOOGLE/FLAN-T5-BASE, and GOOGLE/FLAN-T5-SMALL and observed that the XXL model outperforms the other LLMs in the provided task. Each model is posed with the same questions.

In the future, Sahaay can access PDFs, Videos, Audio, and other files to extract relevant information about, for example, student activities, research work and innovation carried out by BVM. This multimodal capability has the potential to change forever the way we interact with websites and retrieve information in much less time.

ACKNOWLEDGMENT

The authors would like to express their deepest appreciation to the research facility provided at TRS BVM Laboratory for encouraging multi-disciplinary collaborative research within the campus. We'd also like to thank Birla Vishvakarma Mahavidyalaya (Engineering College) for allowing us to experiment with the innovation on their website.

REFERENCES

[1] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19). Association for Computing Machinery, New York, NY, USA, Article 1, 1–9. https://doi.org/10.1145/3342775.3342784

[2] Kim, S., Rawat, A. S., Zaheer, M., Jayasumana, S., Sadhanala, V., Jitkrittum, W., Menon, A. K., Fergus, R., & Kumar, S. (2023). EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval. *ArXiv*. /abs/2301.12005

[3] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863

[4] Su, Hongjin, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. "One Embedder, Any Task: Instruction-Finetuned Text Embeddings." ArXiv, (2022). /abs/2212.09741.

[5] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale Similarity Search with GPUs." *ArXiv*, (2017). Accessed September 28, 2023. /abs/1702.08734.

[6] Chung, Hyung W., Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li et al. "Scaling Instruction-Finetuned Language Models." ArXiv, (2022). Accessed September 28, 2023. /abs/2210.11416

[7] Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. ArXiv. /abs/1906.02569

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/372529063

An Effective Query System Using LLMs and LangChain

Article *in* International Journal of Engineering Research and · July 2023

CITATIONS
4

READS
2,691

2 authors:

Adith Sreeram a
VIT - AP University
2 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Jithendra Sai
VIT-AP University
2 PUBLICATIONS **4** CITATIONS

SEE PROFILE

All content following this page was uploaded by Jithendra Sai on 23 July 2023.

The user has requested enhancement of the downloaded file.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

An Effective Query System Using LLMs and LangChain

Adith Sreeram A S
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

Pappuri Jithendra Sai
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India.

***Abstract*—Due to the unstructured nature of the PDF document format and the requirement for precise and pertinent search results, querying a PDF can take time and effort. LangChain overcomes these challenges by utilizing advanced natural language processing algorithms that analyze the content of the PDFs and extract essential information. To improve the search experience, it uses effective indexing and retrieval techniques, movable filters, and a simple search interface. LangChain also allows users to save queries, create bookmarks, and annotate important sections, enabling efficient retrieval of relevant information from PDF documents. The features of LangChain increase overall efficiency and makes PDF querying much easier and simpler.**

Keywords—LangChain, Querying PDF, Streamlit.

I. INTRODUCTION

The growth and use of digital products is growing exponentially in this world. And the process of searching and retrieving information from those pdf documents is challenging. Now, we have a tool that revolutionized Natural Language Processing and is designed to create applications based on Large Language Models [LLM].

LangChain is a cutting-edge solution which helps us in the querying process and extracting information from PDFs. With its advanced NLP algorithms, it helps users to interact with the PDFs and makes the document search and retrieval very easy.

After building our LLM model we will use Streamlit, a web application framework which helps us create custom attractive web applications. One advantage of Streamlit is that its use does not necessitate familiarity with other web development frameworks like HTML and CSS. With Streamlit, you can instantly deploy your models with minimal effort and code.

II. METHODOLOGY

LanChain helps us with the querying process and extracting information from the PDF based on the prompt sent by the user.For the sake of convenience, a web application is developed that can retrieve accurate information based on the user's input alone.

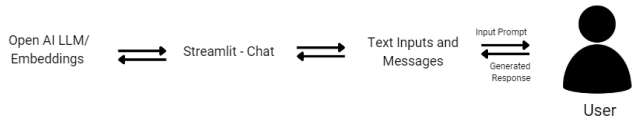


Fig.A. Application Architechture

A. *Steps followed in the Application Architechture:*

Step I: The Open AI Large Language Models and The Open AI Embeddings acts as the back-end of our application.

Step II: Here we will use Streamlit, which will help us to build interactive and beautiful interface for our web application.

Step III: Streamlit will also take care of our Front-end part where we can get the text inputs and messages and also the PDF files from the user.

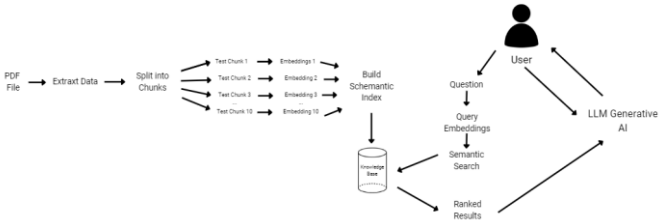


Fig.B. Working Process

With the help of Fig.B we can understand how Large Language Model helps the user to get the accurate results.

B. *Streamlit*

Streamlit is an open-source library that allows us to unique web apps for Machine Learning and Data Science projects fast and efficient. Streamlit is an open-source library that allows us to unique web apps for Machine Learning and Data Science projects fast and efficient. With this framework, you can easily build interactive visualization plots, models, and dashboards without having a worry about the underlying web framework or deployment infrastructure used in the backend. It also provides the users to add widgets which helps the users the interact with the web app and the models that we used. This framework also integrates the popular python and machine learning packages such as NumPy, Pandas, Matplotlib,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

Seaborn, Scikit-learn and TensorFlow, which enables us to quickly build and deploy our trained models.

Features of Streamlit:

User-friendly: Streamlit offers an easy-to-use interface that requires little scripting to build dynamic data apps.

Rapid prototyping: Streamlit is made for rapid prototyping, allowing developers and data scientists to test out various concepts and create completely functional apps.

Data Cache: The data cache facilitates and accelerates computational workflows.

Real-time collaboration is made possible by Streamlit, allowing several users to work on the same project at once. Widgets that enable for real-time data editing and exploration include sliders, dropdown menus, and checkboxes, among a vast variety of interactive widgets that Streamlit offers.

III. RESULTS AND DISCUSSION

A. Images of Web Application and Output.

Fig.C. Interface of web application

This is how the interface of our web application will look like. Now the user can click on browse files and can upload a file from their device under 200 Mega Bytes. After few minutes of processing, we will get an additional in box where we can give in our query.

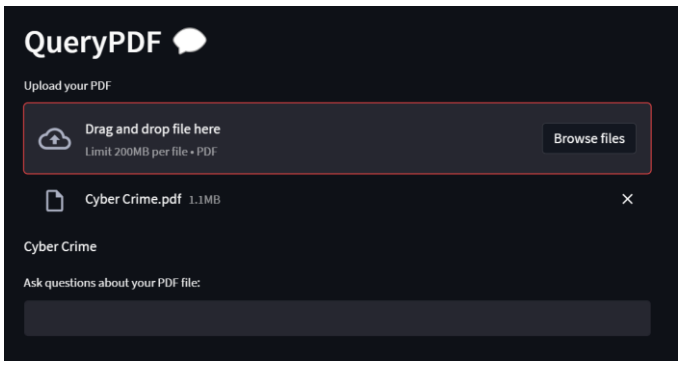


Fig.D. Image of web application with input query box.

So, now we got our input query box and now we can ask questions on the PDF that we have uploaded. Here I have uploaded a PDF based on Cyber Crime. Now you can ask different questions like "What is Cyber Stalking?", "What are the recent incidents of Cyber Terrorism in World?" and also differentiate between questions.

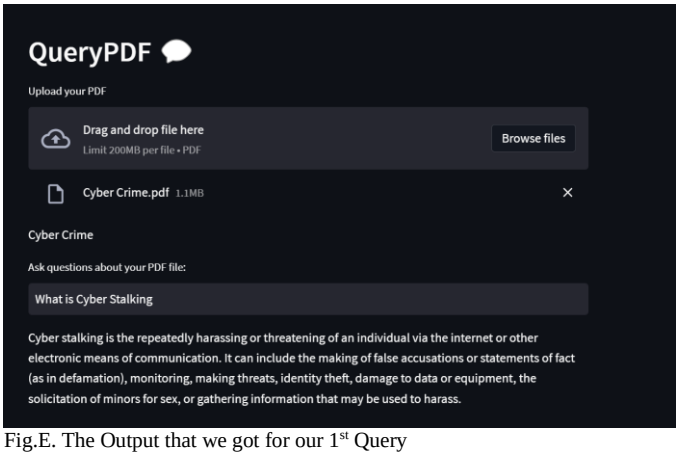


Fig.E. The Output that we got for our 1st Query

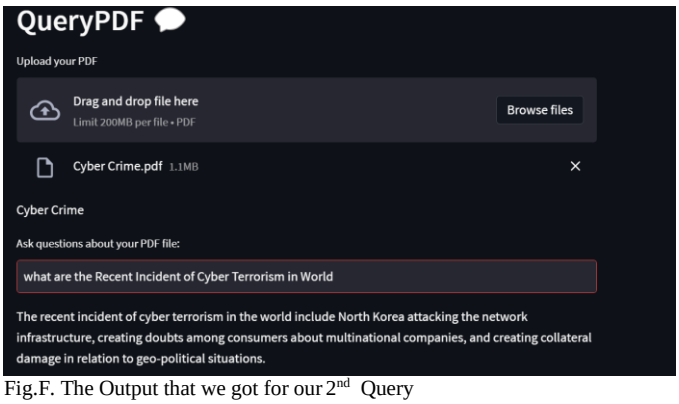


Fig.F. The Output that we got for our 2nd Query

Here we got our output for our 1st and 2nd query which is "What is Cyber Stalking?" and "What are the recent incidents of Cyber Terrorism in World?" our Large Language Model went through file and gave an accurate result on the query given.

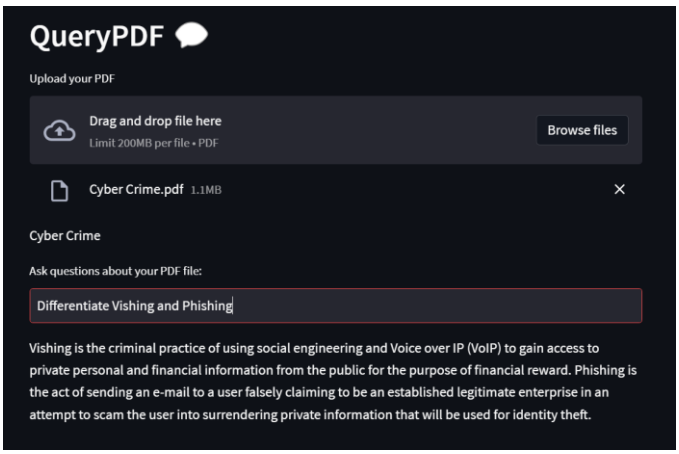


Fig.G. The output we got for Differentiate Between Question

This output that the application gave us after the query is quite interesting. When we look into the file there were 3 pages discussing on Vishing and Phising but the application gave us a clear and concise differentiation on both Vishing and Phishing in just 4 line.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

IV. CONCLUSION

Using LangChain and Large Language Model and Streamlit we have created a web application that simplifies and enhances the process of extracting relevant information from PDFs. Users can now retrieve any information in the PDF and save their time and effort. The integration of LangChain technology adds a layer of efficiency and accuracy to the querying process, making the app a valuable tool for individuals working with PDF documents.

REFERENCES

[1] https://streamlit.io/

[2] https://python.langchain.com/

[3] https://platform.openai.com/docs/models

[4] Meharwade, Anuradha & Patil, G.A.. (2016). Efficient Keyword Search over Encrypted Cloud Data. Procedia Computer Science. 78. 139-145. 10.1016/j.procs.2016.02.023. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[5] Nashipudimath, Madhu & Shinde, Subhash & Jain, Jayshree. (2020). An efficient integration and indexing method based on feature patterns and semantic analysis for big data. Array. 7. 100033. 10.1016/j.array.2020.100033. I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[6] Zhu, Miao & Cole, Jacqueline. (2022). PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format. Journal of Chemical Information and Modeling. 62. 10.1021/acs.jcim.1c01198. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

View publication stats

Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Advancing Educational Accessibility: The LangChain LLM Chatbot's Impact on Multimedia Syllabus-Based Learning

A. Annie Micheal

annymick@gmail.com

Sathyabama Institute of Science and Technology

Ashwin Prasanth

Sathyabama Institute of Science and Technology

Aswin TS

Sathyabama Institute of Science and Technology

Krisha BL

Sathyabama Institute of Science and Technology

Research Article

Keywords: Chatbot, Large Language Model model, Generative Pre-trained Transformer, LangChain LLM

Posted Date: May 21st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4399670/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

A chatbot is a computer program using artificial intelligence to simulate conversation with users, providing information or performing tasks in natural language. Traditional chatbots in education help navigate course details, but often lack subtle understanding and context. The Large Language Model (LLM) model, computationally intensive and slower than Generative Pre-trained Transformer, generates answers solely from course material, limiting responses to user input. In our work we have proposed a chat history view option, a multimodal syllabus upload capability, improved querying capabilities, and the ability to recall past responses for enhanced performance. Our chatbot leverages the powerful LangChain LLM to overcome these limitations. Our research uses Lang Chain's LLM to power AI-driven educational chatbots, analyzing course materials to tailor responses for personalized student learning. This paper delves into the design, architecture, and evaluation of this chatbot, showcasing its impact on engagement, comprehension, and accessibility, and ultimately paving the way for a future of universally inclusive education. Evaluations show our LLM-powered approach achieves high accuracy, generating clear, precise answers, exceeding the abilities of traditional educational chatbots.

1. Introduction

Educational chatbots powered by Artificial Intelligence (AI) are emerging as dynamic tools fostering interactive communication among students, educators, and administrators. These virtual assistants enrich the learning experience through personalized support, query resolution, coursework guidance, and navigation of complex academic materials. Natural Language Processing (NLP) and Machine Learning (ML) underpin these chatbots, enabling them to deliver a more accessible and engaging learning environment tailored to individual needs and streamlining communication within the educational ecosystem. However, traditional resources often struggle to cater to diverse student requirements, lacking in accessibility, personalization, and efficiency. This forces students to navigate a labyrinth of information, consuming valuable time in pursuit of relevant details. Furthermore, global disparities in access exacerbate these challenges, limiting opportunities for quality education. Fortunately, technological advancements like LangChain's Large Language Models (LLMs) and OpenAI APIs offer a promising solution [1].

In our work we propose, a novel chatbot leveraging LangChain's LLM capabilities to analyze uploaded educational resources and extract key syllabus information. These insights then inform the generation of personalized, concise responses, tailoring the learning experience for each student. The primary objectives are to enhance engagement, reduce study time, foster deeper comprehension, and ultimately democratize access to high-quality education [3]. Our research represents a significant step forward in educational advancement, driven by a ground breaking LLM-powered chatbot. It signifies a seismic shift towards enriched accessibility and uniquely personalized learning experiences. By harnessing the potential of LLMs, our research aims to reshape how students interact with educational materials, offering a novel solution to pervasive challenges in accessibility and personalization. Its core objective lies in empowering learners through customized responses, simplifying information exploration across

diverse sources. The unwavering commitment to dismantling educational barriers underscores the dedication to fostering an inclusive and comprehensive learning environment. Beyond immediate impact, this research holds the potential to spark further educational breakthroughs, paving the way for a globally inclusive educational landscape. Integrating cutting-edge technologies like LangChain LLMs propels us towards a future defined by accessible, personalized, and universally inclusive education [6] [11].

This research manuscript introduces an innovative LLM-powered chatbot designed to revolutionise personalised learning experiences, as delineated in this comprehensive paper. In Section 1, the investigation meticulously surveys the existing landscape of educational chatbots and the application of LLMs in education. Section 2, presents a detailed insight into the chatbot's design and implementation, encompassing aspects such as data collection methodologies and the judicious selection of the LangChain LLM model, is exhaustively expounded upon. Section 3, describes the technical architecture is unveiled with precision, featuring distinct modules dedicated to natural language processing, syllabus analysis, and personalised response generation. Section 4, concludes the paper with rigorous evaluation methodologies are employed to discern the chatbot's impact on user engagement, comprehension, and accessibility, as elucidated in the conclusive section. Section 5, concluding the discourse, avenues for prospective research are posited, focusing on advancing personalisation and accessibility in the realm of education.

2. Literature Survey

2.1 Langchain

The LangChain [1] explores the practical application of Large Language Models (LLMs) for the efficient development of applications, with a specific focus on LangChain—an open-source software library. LLMs have gained widespread adoption due to their versatile capabilities, spanning tasks like essay composition, code writing, explanation, and debugging. OpenAI's ChatGPT has notably played a significant role in popularizing the use of LLMs, garnering millions of users. This study places a central emphasis on LangChain, strategically designed to streamline the creation of customized AI applications utilizing LLMs. LangChain has received recognition within the AI community for its effectiveness in seamlessly interfacing with various data sources and applications [2]. The research provides a thorough examination of LangChain's fundamental features, shedding light on its components and chains. Components function as modular abstractions, while chains offer customizable pipelines tailored to specific use cases. Through practical examples, the study aims to showcase the framework's potential in facilitating the rapid development of applications based on LLMs.

2.2. Sequence - To – Sequence Communication

The methodology of sequence-to-sequence learning [3], utilizing recurrent neural networks (RNNs), is a robust neural network paradigm that orchestrates the mapping of input sequences to variable-length output sequences. Rooted in the recurrent nature of neural networks, this approach excels in capturing

intricate sequential dependencies within data and has demonstrated its effectiveness across a spectrum of applications, prominently in machine translation, speech recognition, and text summarization.

In machine translation, sequence-to-sequence models showcase their prowess by adeptly translating sentences from one language to another. The inherent capacity of these models to comprehend contextual intricacies ensures the production of coherent and contextually accurate translations. Similarly, in the arena of speech recognition, sequence-to-sequence learning exhibits its efficiency in transforming spoken language into written text [4] [5]. By discerning sequential patterns in audio data, these models significantly enhance the accuracy and effectiveness of the speech-to-text conversion process. Furthermore, within the realm of text summarization, sequence-to-sequence learning emerges as a valuable tool for condensing extensive documents into concise summaries while preserving vital information. The adaptability and efficacy of sequence-to-sequence learning underscore its versatility, establishing it as a potent approach for addressing intricate tasks that necessitate a nuanced understanding of sequential relationships within data.

2.3. Large Language Model

Large Language Models (LLMs), exemplified by models like GPT-3, have become pivotal in natural language processing tasks, showcasing a remarkable ability to understand and generate human-like text. GPT-3, developed by OpenAI, is particularly noteworthy for its vast scale, trained on diverse datasets to capture intricate language patterns [6]. Its architecture is grounded in transformer models, specifically utilizing the attention mechanism to process input sequences and generate coherent and contextually relevant text. GPT-3's success is attributed to its ability to predict the next word in a sequence based on its understanding of context, making it a versatile tool for applications such as text generation, translation, summarization, and more.

In the realm of large language models, various algorithms contribute to their success. Apart from GPT-3 [7], other notable models include BERT (Bidirectional Encoder Representations from Transformers), which focuses on bidirectional context understanding, and T5 (Text-to-Text Transfer Transformer), a model that frames all NLP tasks as converting input text to output text. Additionally, XLNet employs a permutation language modeling objective to capture bidirectional context, while RoBERTa (Robustly optimized BERT approach) refines the training of BERT for improved performance. Each of these algorithms has its unique strengths and applications, contributing to the diverse landscape of large language models in natural language processing.

2.4. OpenAI and ChatGPT

OpenAI's ChatGPT is an impressive achievement in natural language processing, built on a transformer-based architecture that excels in sequential data processing. The model undergoes a two-step learning process: pre-training and fine-tuning. In the pre-training phase, ChatGPT [7] learns from a diverse dataset, predicting the next word in a sentence to grasp contextual relationships and linguistic patterns. This extensive dataset includes publicly available internet data, licensed third-party information, and user-provided data. Subsequently, in the fine-tuning phase, the model is refined on a narrower dataset with

human reviewers adhering to OpenAI's guidelines [8]. This process ensures the model's behavior aligns with specific use cases and ethical considerations.

Throughout both phases, ChatGPT dynamically adjusts its internal parameters or weights, evolving to better capture the nuances of language. This dynamic adaptation, combined with the power of the transformer architecture and extensive training data, enables ChatGPT to provide coherent and contextually relevant responses in diverse language tasks. OpenAI's commitment to privacy and ethical standards is evident in the development process, with measures in place to filter and exclude undesirable content. In essence, ChatGPT exemplifies advanced language models, showcasing OpenAI's dedication to pushing the boundaries of AI language understanding while maintaining ethical standards and user privacy.

2.5. Streamlit

Streamlit is a cutting-edge Python-based tool designed to revolutionize the creation of web applications, particularly tailored for machine learning and data science applications. Its strength lies in its simplicity, allowing data scientists to develop robust web applications with ease using just Python scripts. Streamlit provides a user-friendly interface for data visualization and interaction, while Python LangChain. offers comprehensive documentation for integrating language models into applications The workflow with Streamlit is highly intuitive, where developers can seamlessly transform data scripts into interactive and shareable web applications without delving into complex web development frameworks. Its ability to convert data exploration and analysis scripts into interactive apps simplifies the process of sharing insights and models with stakeholders. Streamlit's versatility extends to its capability to effortlessly load local files and connect with databases through real-time API calls [9]. This enables dynamic data updates and enhances the interactivity of applications. Moreover, Streamlit provides an array of features such as widgets, charts, and plots that facilitate the creation of visually appealing and informative user interfaces. The tool's ability to handle complex data visualization and analysis tasks, combined with its user- friendly interface, positions Streamlit as an invaluable resource for data scientists seeking a quick and efficient means to showcase their work or deploy machine learning models to a broader audience.

2.6.Chatbot

Chatbots are a type of artificial intelligence (AI) that can be used to interact with users in a natural way. They can be used to provide information, answer questions, and complete tasks. Rule-based chatbots operate on predefined instructions, crafted by developers to navigate user interactions. Their strength lies in handling routine queries or straightforward tasks, like offering basic customer assistance or relaying commonly sought information. Yet, their effectiveness wanes when faced with intricate or varied user inputs due to their rigid, predefined nature [10]. While quick to deploy and relatively simple to develop, these bots often struggle to grasp nuanced queries or adapt to unanticipated scenarios, marking a boundary in their capacity to engage users comprehensively. These systems harness Natural Language Processing (NLP) [11] methodologies alongside Deep learning techniques to build and train

chatbot systems. Most of these models leverage the sequence-to-sequence approach, which initially gained prominence in fields like Machine Translation, Speech Recognition, and Text Summarization.

2.7.Whisper Parser Speech to text NLP

The Whisper ASR system, developed by OpenAI, represents a state-of-the-art automatic speech recognition (ASR) model. Serving as a powerful speech-to-text tool [11], Whisper is trained on a vast amount of multilingual and multitask supervised data, making it adept at transcribing spoken language into written text. Its proficiency lies in understanding the nuances of various languages and accents, showcasing robust performance across diverse linguistic contexts. Whisper's architecture is built upon a deep neural network, enabling it to capture intricate patterns and contextual dependencies in spoken language. This makes it particularly effective in applications like transcription services, voice assistants, and other tasks requiring accurate and efficient speech-to-text conversion. The model's adaptability across different languages and its ability to handle various accents make Whisper a versatile and valuable tool for a wide range of applications, contributing to the advancement of automatic speech recognition technology.

3. Methodology

In this paper, we propose a LangChain system for processing and interacting with diverse data sources. LangChain uses loaders to import data from various formats, including PDFs, YouTube videos, and Notion databases. Loaders extract both content and metadata, allowing for in-depth understanding. After loading, documents are split into smaller, relevant sections for easier retrieval and analysis. This data is then stored in a vector store and converted into numerical representations called embeddings. These embeddings enable powerful semantic search techniques for finding specific information within the data. Finally, language models are used to answer questions based on the retrieved information, leading to a conversational user experience. Figure 1 shows the architecture of the proposed work.

3.1. Document Loading

Document loaders, an integral part of data-centric applications, facilitate the transformation of diverse data sources into a standardized format for seamless interaction. Within LangChain's extensive suite of over 80 loaders, these tools cater to varied formats and origins, ranging from extracting unstructured data like text from public domains such as YouTube and Twitter to proprietary sources like Figma or Notion. Moreover, loaders extend their capabilities to structured data, including tabular formats from platforms like Airbyte, Stripe, or Airtable. By importing specific loaders such as the PyPDF loader for PDFs or the YouTube audio loader combined with the OpenAI Whisper parser for video transcripts, users gain the ability to process content effectively. Extracted documents contain both content and associated metadata, facilitating an in-depth understanding of the data's origin and structure. These loaders also support web-based content extraction from URLs, enabling interactions with educational materials or markdown files, albeit requiring post-processing for optimal usability.

Loaders catering to Notion databases, like the Notion directory loader, empower users to access and engage with data exported from this popular platform, such as Blendle's Employee Handbook showcased in a markdown format. Despite the comprehensive coverage, the subsequent focus shifts to breaking down loaded documents into smaller, more relevant sections, essential for targeted retrieval and generation—a pivotal aspect in enhancing user interaction. This process encourages users to contemplate potential data sources yet unaccounted for, fostering an inclusive approach that invites contributions to LangChain, potentially expanding its loader repertoire.

3.2.Document Splitting

Dividing documents into smaller parts might seem simple, but it's crucial for accurate analysis and downstream applications. (after loading the data) Splitting happens after data loading and before storing the information in a special "vector store." The way you split can have a big impact on the results you get. Lang Chain's text splitters offer a solution: consistent chunking with overlaps to maintain context. These tools can use different methods to measure chunk sizes, like character count or sentence boundaries. They also ensure crucial metadata stays with each chunk to aid interpretation.

Splitting methods vary in how they divide up the text, what characters they allow, and how they measure chunk size. Different document types, like code, may even have specialized splitters. To understand the nuances, you can experiment with different parameters and separators in small samples before applying them to real-world documents like PDFs or Notion databases. This hands-on approach helps you see how splitters preserve context and influence data analysis later on.

Extracting data from YouTube involves specialized loaders, notably the YouTube audio loader in conjunction with the OpenAI Whisper parser. This process begins by specifying a YouTube URL and a directory to store the audio files obtained from the video content. By combining the YouTube audio loader with the Whisper parser, LangChain can transcribe the audio content into a text-based format suitable for analysis and interaction. This operation typically involves converting spoken content from videos into written transcripts, enabling users to access and work with the textual representation of YouTube video content. Once loaded, users can explore and manipulate this transcribed data, making it accessible for a variety of applications, such as chatbots or conversational interfaces that engage with content from YouTube videos. This extraction process allows users to delve into and leverage the vast array of educational, informational, or entertainment content available on the YouTube platform, opening doors for analyzing, querying, or interacting with video content in innovative ways.

3.3.Vector Stores And Embeddings

Vector Stores And Embeddings is a systematic approach for implementing semantic search methodologies leveraging embeddings and vector stores to optimize the efficiency of data retrieval. It initiates by delineating the segmentation of documents into smaller, semantically linked segments, forming the groundwork for subsequent indexing processes. Emphasizing the use of embeddings, it underscores their role in converting textual information into numerical representations, thereby facilitating comparisons and enabling the identification of analogous content within the corpus. This

approach encompasses a stepwise workflow encompassing document segmentation, embedding creation, and storage within a specialized vector store tailored for rapid similarity- based retrieval. The text underscores Chroma, a memory-optimized vector store, highlighting its lightweight nature and user-friendly features. Acknowledged within the content are limitations, including the presence of duplicate information leading to redundant results and the challenge of integrating structured data cues into semantic embeddings, potentially impacting retrieval accuracy. Encouraging an exploratory stance, the content advocates for experimenting with query variations and parameter adjustments, particularly in the quantity of retrieved documents, to strike a balance between result relevance and quantity. Future refinements aim to enhance the methodology, addressing these limitations to optimize the retrieval process for more accurate and targeted information extraction.

3.4.Retrieval

It covers techniques like Maximum Marginal Relevance (MMR) aims to find diverse yet relevant details, ensuring a well-rounded result. Self-querying helps with complex queries that involve additional details like dates or categories. Contextual compression simplifies information by picking out the most important parts from texts. These methods are demonstrated to show how they refine searches and solve problems seen in older ways of finding information. Additionally, traditional methods like SVM and TF- IDF are introduced for different types of searches. The aim is to help users experiment and find the best way to get the right information for different questions.

3.5.Question Answering

A library of documents relevant to a question, but how to unlock the answer hidden within their pages, that's where language models come in, acting as information alchemists who transform text into answers. The simplest approach, "stuff," throws all the documents into the language model's cauldron, hoping for gold. But just like overloading a furnace, this can overwhelm the model, leaving precious insights unmelted. MapReduce takes a more methodical approach. Each document is distilled into a summary, then these summaries are combined like ingredients in a recipe. This avoids information overload, but it's akin to cooking blindfolded – connections between documents might be missed, leading to bland or inaccurate answers.

Refine, like a skilled chef, iteratively improves the answer. It starts with an initial response, then feeds the model new documents and the existing answer, asking it to refine the dish. This back- and-forth allows Refine to blend information from multiple documents, crafting a more nuanced and flavorful answer. However, even the most skilled chef can't remember every ingredient from a previous dish. Current chains lack memory, so follow-up questions about the answer feel like starting a new recipe from scratch. This is where memory enters the kitchen, enabling stateful conversations and truly context-aware responses.

3.6.Chat

This is the final stage in creating a chatbot capable of engaging in conversation and answering questions. It introduces the integration of chat history, aiding the bot in understanding context for subsequent inquiries. A pivotal addition is the conversational retrieval chain, streamlining past discussions into standalone queries, enhancing response accuracy.

4. Result

The study's findings on the efficacy of the LLM-powered chatbot in enhancing personalized learning represent a significant breakthrough in the realm of educational technology. Through meticulous testing and expert evaluations, the chatbot emerged as a transformative tool, demonstrating its remarkable impact on user engagement, comprehension, and accessibility. Users reported a heightened interest in the learning process, attributing it to the chatbot's personalized approach, which subsequently led to a more profound understanding of educational material. The chatbot's ability to tailor responses to individual needs emerged as a key factor in making learning resources more accessible to a diverse audience. As we delve into the results of this research, it becomes evident that the chatbot holds substantial promise in revolutionizing education, offering a more personalized, inclusive, and engaging learning experience for learners across various backgrounds. Figure 2 shows the interface of the webpage. The interface of the pdf uploader is shown in Fig. 3. Figure 4 illustrates the response interface of the pdf. Figure 5 shows the interface of uploading video link. The response interface of the video is shown in Fig. 6.

5. Conclusion

Our proposed work serves as a compelling exploration into the transformative potential inherent in the seamless integration of LangChain's LLM-powered chatbot within the educational realm. The envisioned future work, which includes the seamless integration of language translation features and the development of a sophisticated chat-to-PDF converter, signifies a steadfast dedication to addressing and surmounting global learning challenges. Through an unwavering commitment to fostering inclusivity, heightening accessibility, and providing innovative tools for the preservation of educational interactions, the project not only seeks to redefine but also elevate the landscape of personalized learning. The continuous progression towards a more interconnected, adaptable, and technologically advanced educational landscape not only has the potential to redefine individual learning journeys but also to instigate a comprehensive and transformative change in educational technology paradigms. As we navigate this trajectory, the amalgamation of advanced language capabilities and practical document conversion not only encapsulates the essence of the project but also underscores the spirit of an educational future characterized by heightened adaptability, seamless connectivity, and elevated user engagement.

Declarations

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author Contribution

All authors reviewed the manuscript

References

1. Tejaswini NR, Vidya S, Vijaya Kumar T (2023) Langchain-Powered Virtual Assistant for PDF communication.International Research. J Modernization Eng Technol Sci 5(7):2796–2801
2. Ilagan J, Benjamin R, Jose Ramon I (2023) A prototype of a chatbot for evaluating and refining student startup ideas using a large language model. 10.35542/osf.io/azhf9
3. Zhu M, Jacqueline M (2023) PDF Data Extractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. J Chem Inf Model 62(7):1633–1643
4. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convoluitonal sequence to sequence learning. In International conference on machine learning, pp. 1243–1252, PMLR
5. Guo MH et al (2022) Attention mechanisms in computer vision: A survey. Comput visual media 8(3):331–368
6. Topsakal O, Akinci Creating large language model applications utilizing langchain: A primer on developing llm apps fast In Proceedings of the International Conference on Applied Engineering and Natural, Sciences (2023) Konya, Turkey, pp. 10–12
7. Nascimento E, Gracia G, Victorio W, Lemos M, Izquierdo Y, Garcia R, Leme L, Casanova M (2023) A Family of Natural Language Interfaces for Databases based on ChatGPT and LangChain, 2023
8. Roumeliotis KI, Tselikas ND (2023) ChatGPT and Open-AI. APreliminary Review: Future Internet 15(6):192
9. Lee C et al (2022) Stargazer: A hybrid intelligence platform for drug target prioritization and digital drug repositioning using streamlit. :Frontiers Genet 13:868015
10. Nimavat K, Champaneria T (2017) Chatbots: an overview types, architecture, tools and future possibilities. Int J Sci Res Dev 5(7):1019–1024
11. Trivedi A, Pant N, Shah P, Sonik S, Agrawal S (2018) Speech to text and text to speech recognition systems-A review: IOSR J. Comput. Eng 20(2):36–43, 2018

Figures

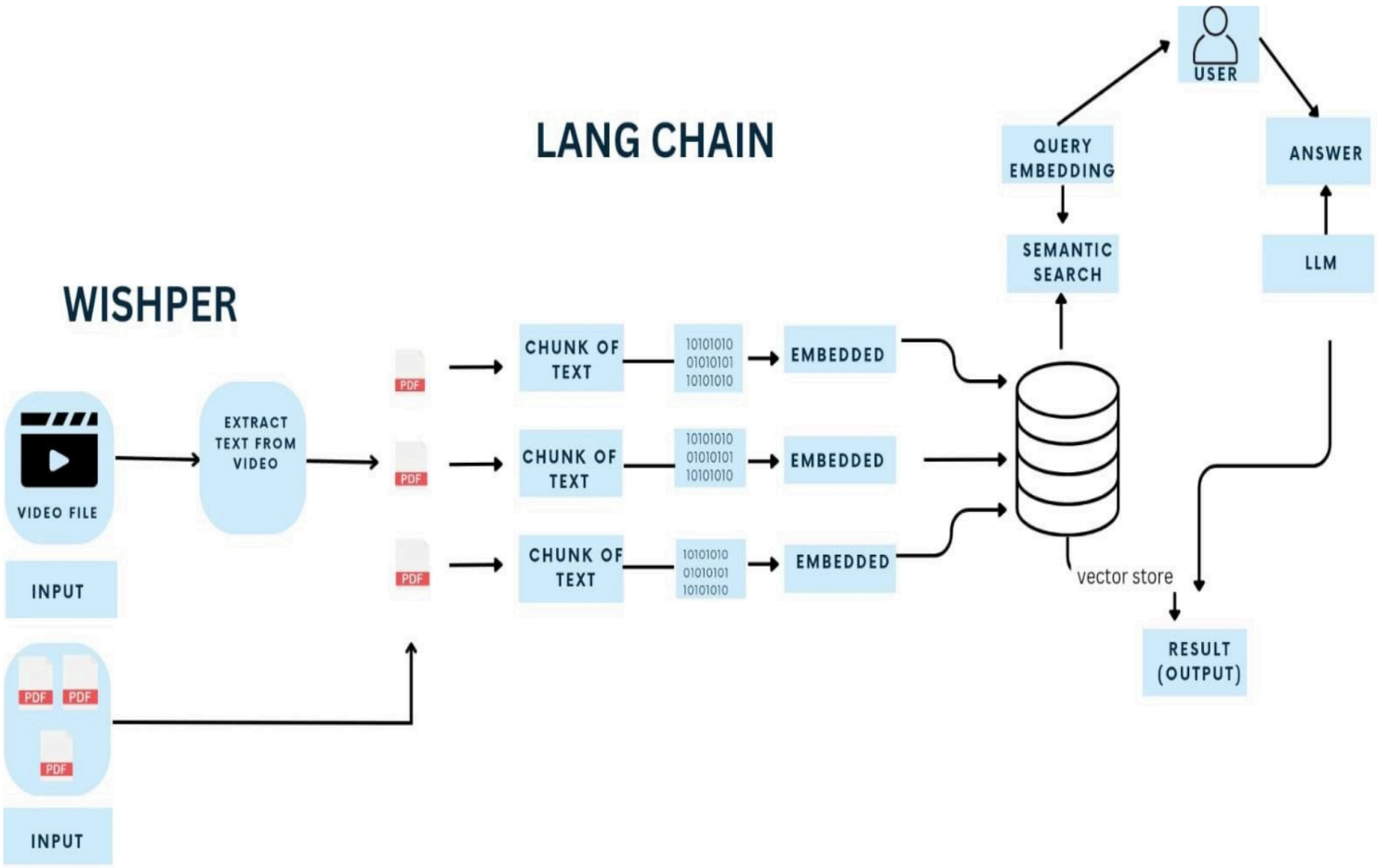


Figure 1

The architecture of our proposed work

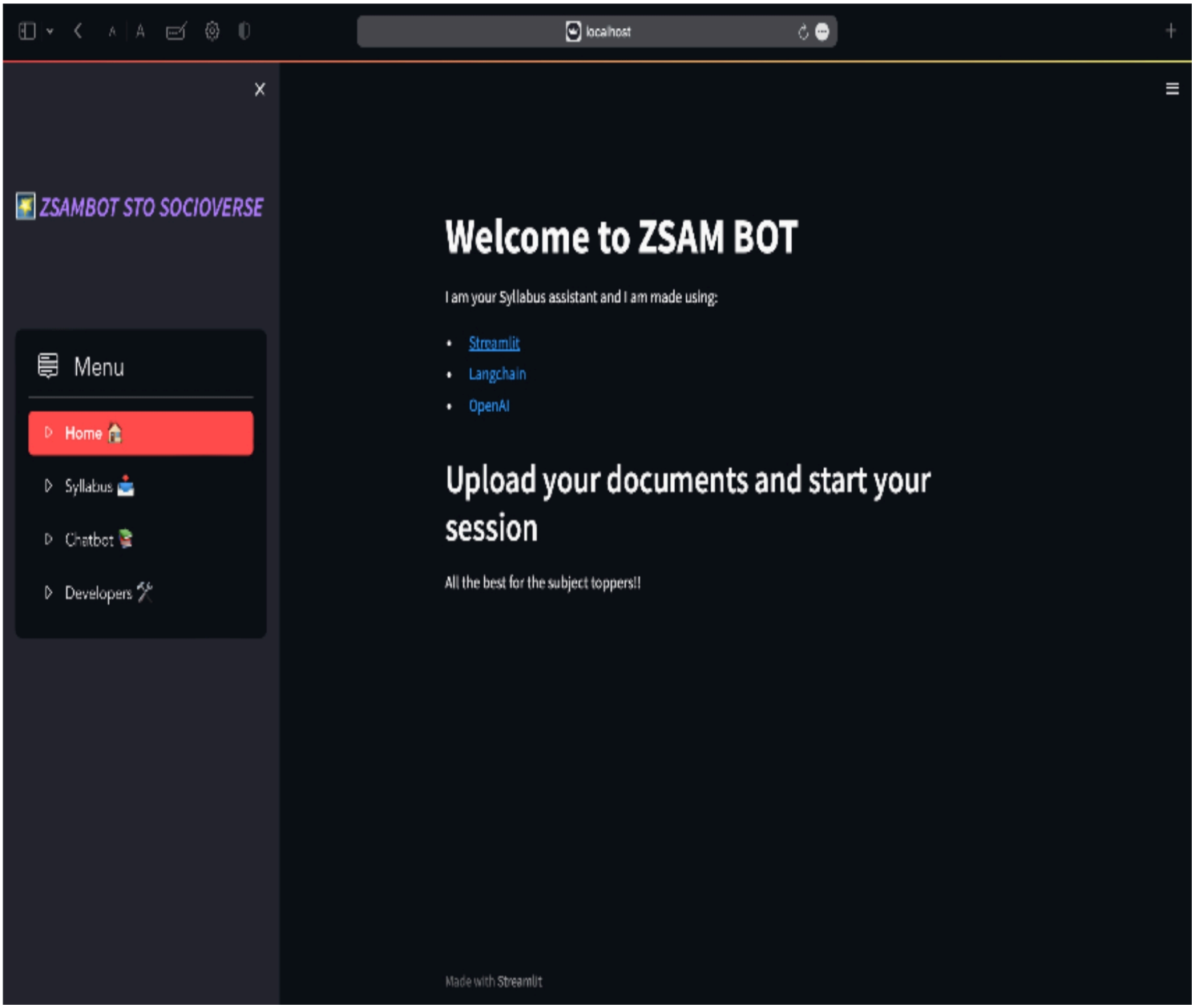


Figure 2

The Interface of the webpage

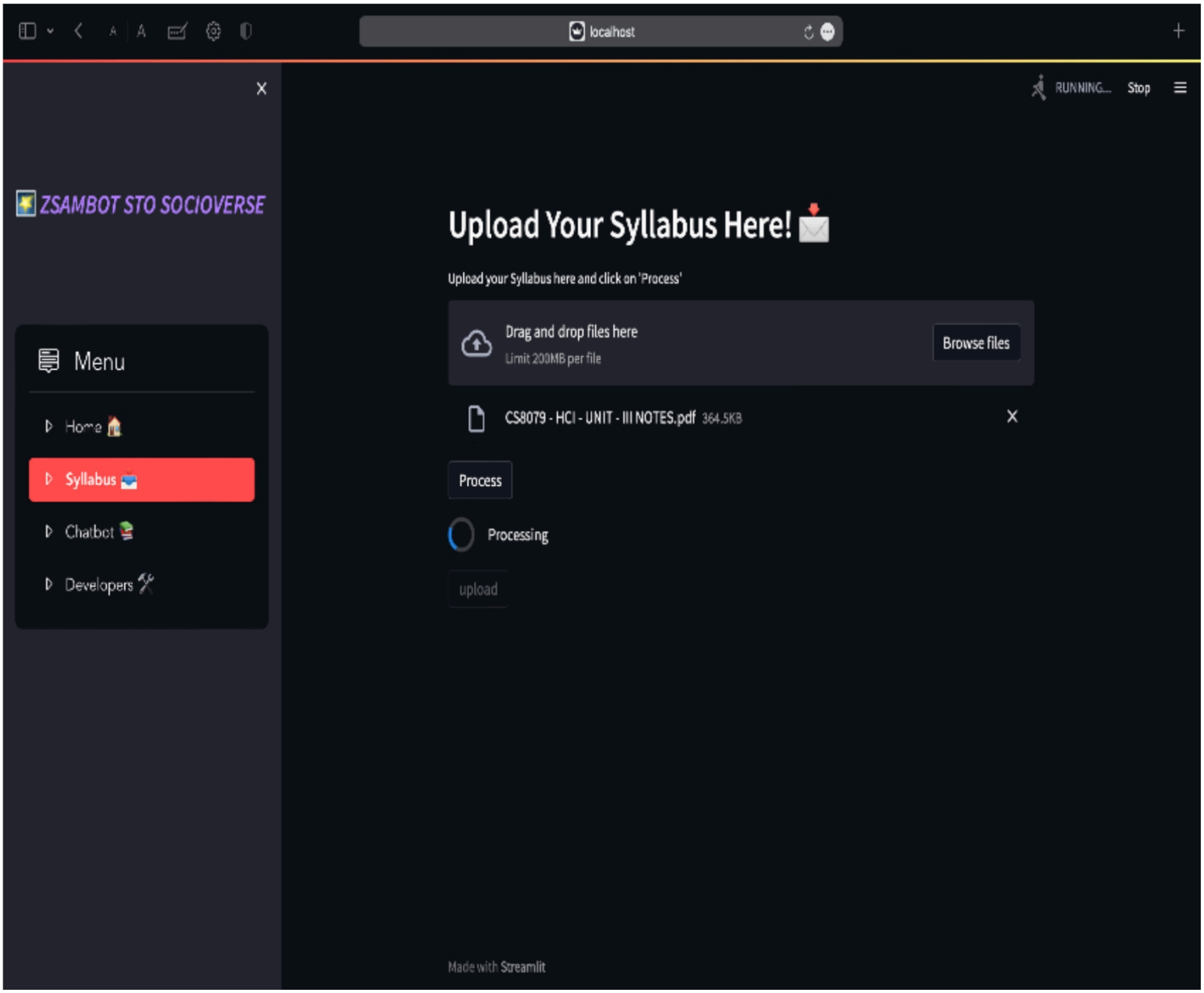


Figure 3

The Interface of the pdf uploader

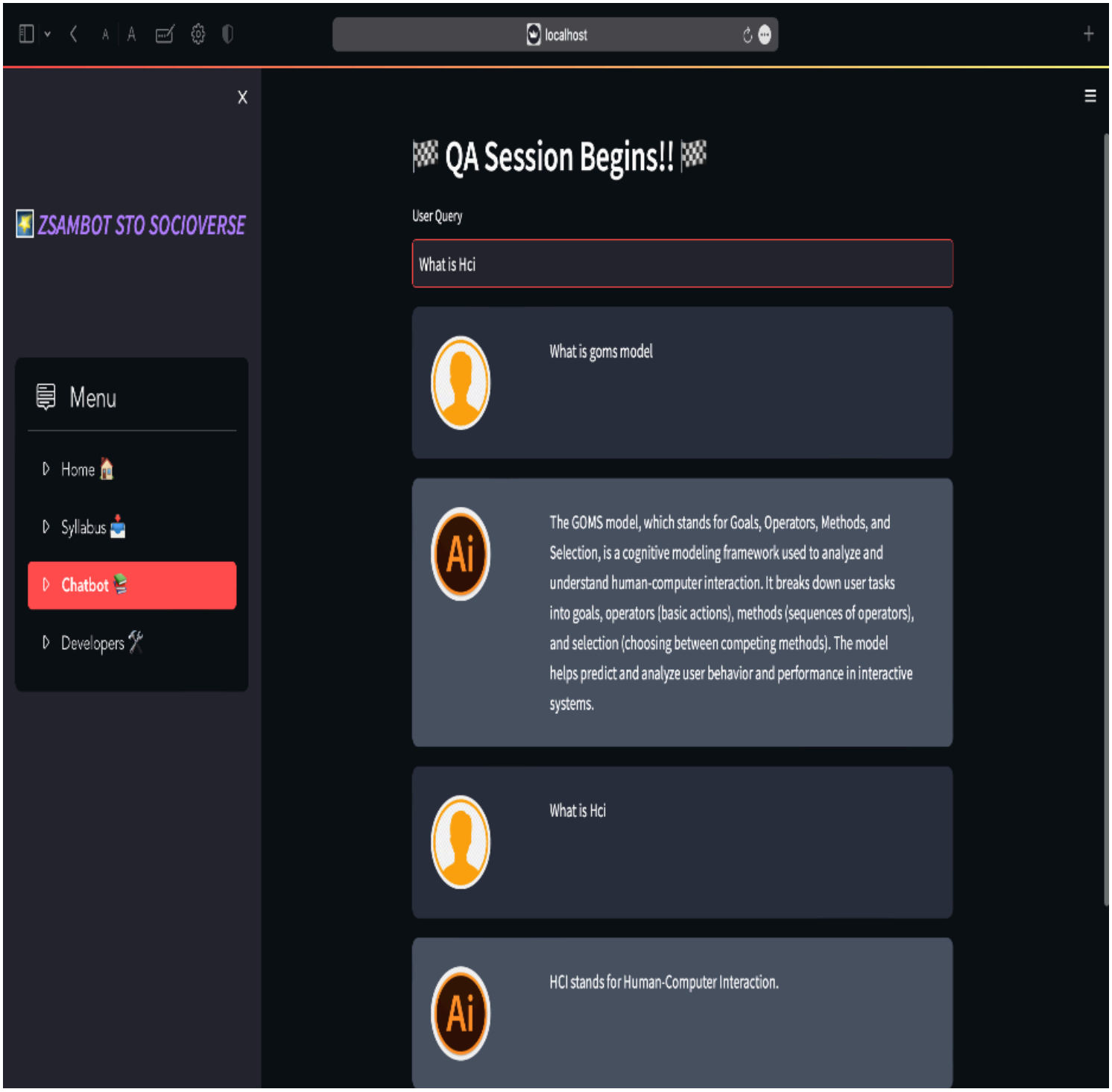


Figure 4

The response interface of the pdf

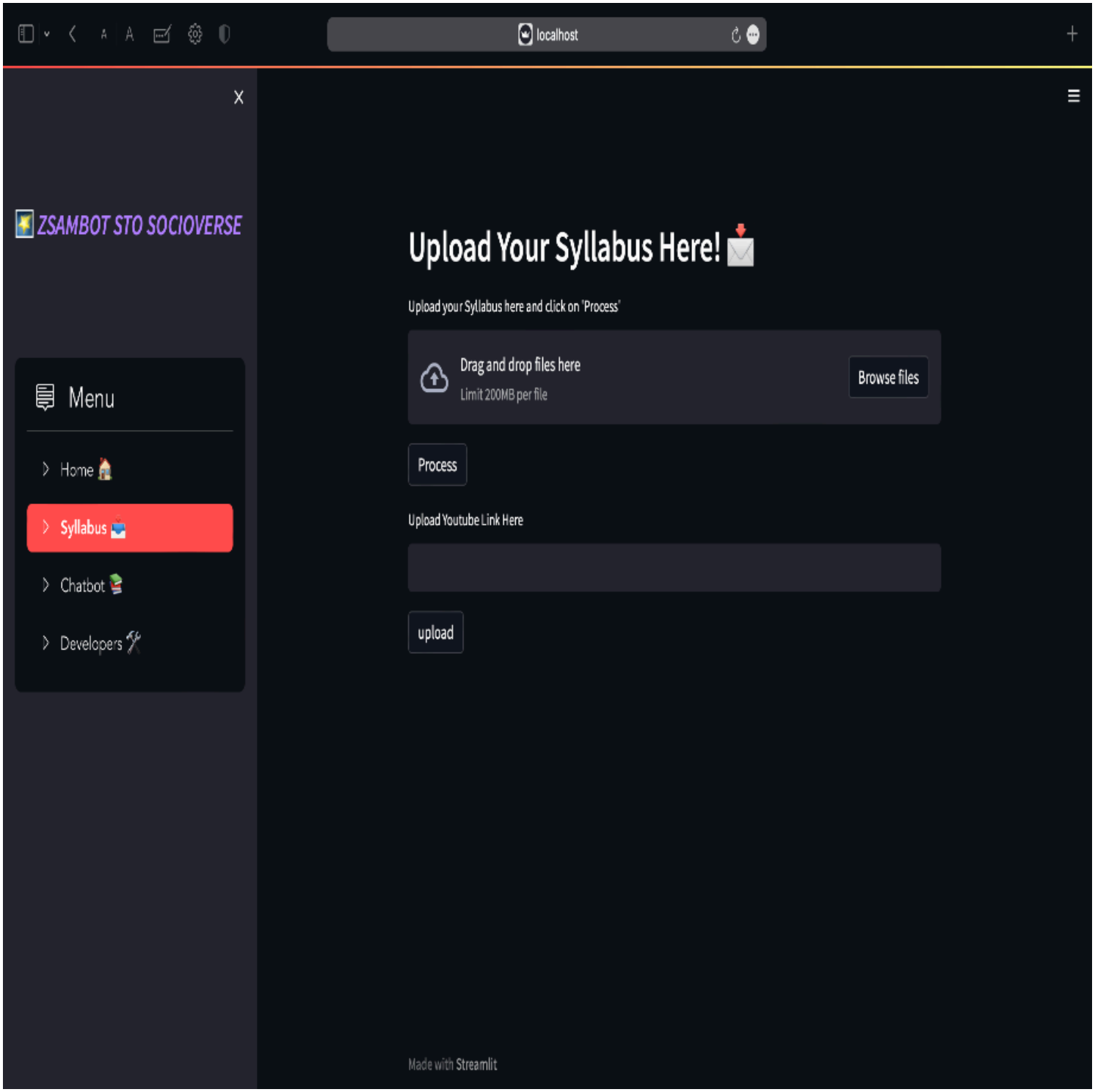


Figure 5

linterface of uploading video link

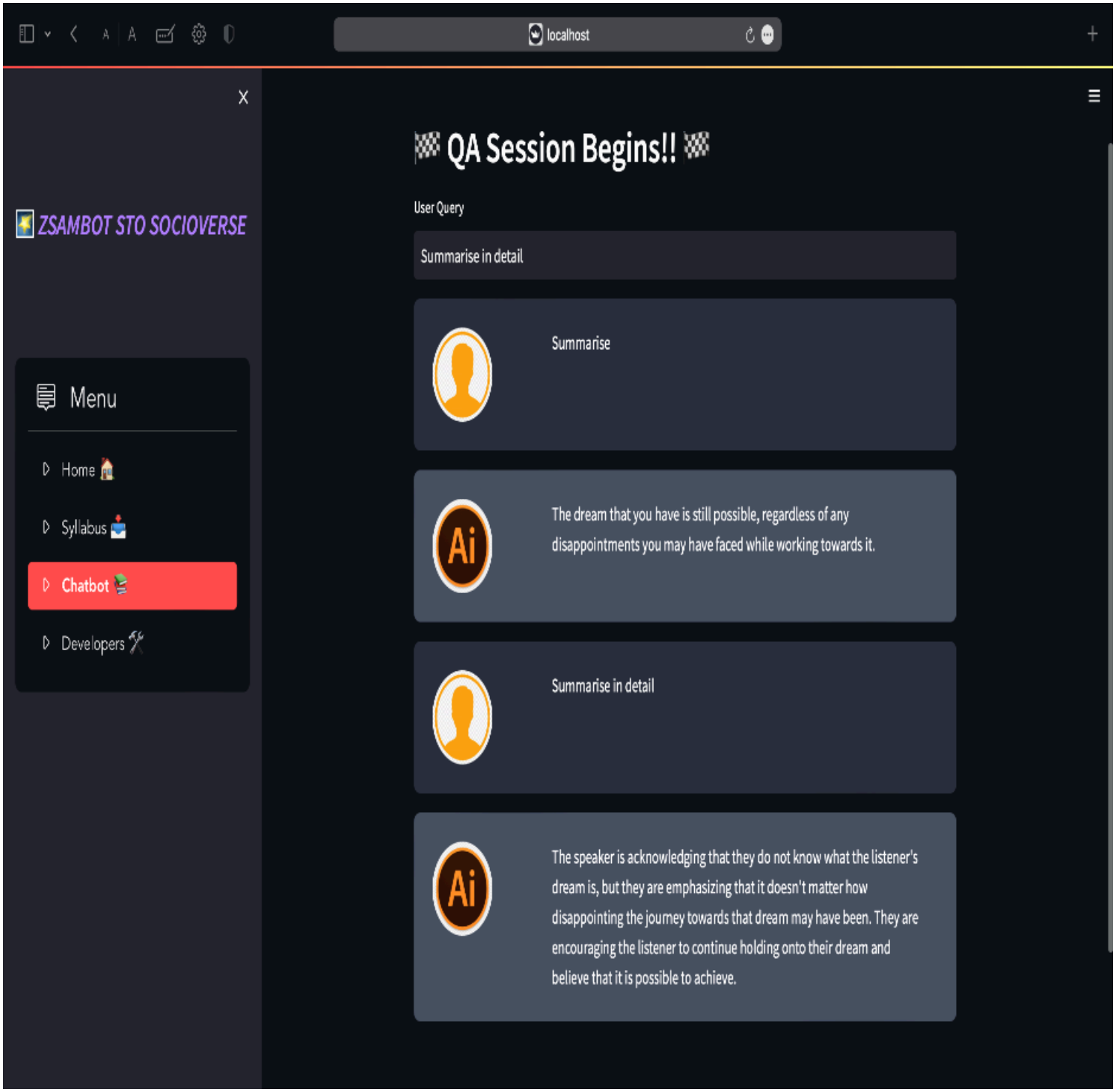


Figure 6

The response interface of the video

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/247871273

A Friend Living inside Me—The Forms and Functions of Imaginary Companions

Article *in* Imagination Cognition and Personality · July 2005

DOI: 10.2190/4M9J-76M2-4Q4Q-8KYT

CITATIONS
49

READS
1,557

1 author:

Eva Hoff
Lund University

30 PUBLICATIONS **874** CITATIONS

SEE PROFILE

All content following this page was uploaded by Eva Hoff on 16 June 2016.

The user has requested enhancement of the downloaded file.

Publisher of Scholarly Books and Journals Since 1964
THE LATEST RESEARCH TO YOUR DOOR OR YOUR DESKTOP
http://baywood.com

BAYWOOD REPRINTS

Imagination cognition and personality
Consciousness in
Theory • Research • Clinical Practice

Editor: Jerome L. Singer, Ph.D.
Robert G. Kunzendorf, Ph.D.

Volume 24, Number 2 — 2004-2005

A Friend Living Inside Me—The Forms and Functions of Imaginary Companions

Eva V. Hoff

BAYWOOD PUBLISHING COMPANY, INC.
26 Austin Avenue, PO Box 337, Amitvville, NY 11701

IMAGINATION, COGNITION AND PERSONALITY, Vol. 24(2) 151-189, 2004-2005

A FRIEND LIVING INSIDE ME—THE FORMS AND FUNCTIONS OF IMAGINARY COMPANIONS

EVA V. HOFF
Lund University, Sweden

ABSTRACT

Imaginary companions and the characteristics of the children who invent them were explored through interviews. Twenty-six children took part. The first aim was to explore the phenomenon in general, the second to investigate the functions of imaginary companions. General findings concerned the kinds of imaginary companions and some facts about the children who created them. The imaginary companions were mostly same-aged children, but some were different kinds of animal or fantasy creatures. The sources of inspiration varied, comprising friends and siblings. Having a place of their own, where the children could develop their imaginary play, was considered important. The study demonstrated various ways of conducting imaginary play. Imaginary companions functioned as inner mentors, assisting the children in their identity formation work. Imaginary companions were experienced by the children as giving comfort and company, bolstering self-regulation and motivation, enhancing their selves, expanding their personality potential, and finally, enriching their lives.

INTRODUCTION

Imaginary companions appear at different ages. According to different scholars, between 50 and 60% of all children have imaginary companions at one point during their childhood [1-3]. The earliest imaginary companions appear at the age of two or three [1, 2]. Some of these companions are only kept for a couple of months; others are played with for years. Some children acquire their imaginary companions upon school entrance, others as late as 10 years of age [2]. There are even accounts of individuals having imaginary companions until the age of 18 [2, 4].

© 2005, Baywood Publishing Co., Inc.

The appearance of pretend playmates[1] in children may worry parents. However, a century's worth of research has demonstrated that there is no reason to be alarmed. There *are* studies showing that children with imaginary companions are less socially adapted [5, 6] and have negative self-images [3], but there is no evidence that these children develop psychological disorders later in life more often than do other children [2]. Some scholars have revealed opposite results, that children with make-believe friends are as well functioning as children without [1, 7].

When the imaginary companion is no longer needed, perhaps when actual playmates and other social activities become more important, it disappears [8]. Imaginary companions may also disappear when adults interfere in the play and try to control the conditions of the imaginary play [1]. However, there are also examples of children who are encouraged by parental engagement in the pretense [2]. In fact, Singer and Singer recommended that parents and other adults should inspire children to develop their imagination through mutual pretend play [1].

There are different definitions of imaginary companions. On the one hand, the imaginary companion has been defined by Svendsen as an invisible character "having an air of reality for the child, but no apparent objective basis. This excludes that type of imaginative play in which an object is personified" [9, p. 988]. On the other hand, according to Singer and Singer's definition [1], imaginary companions can include anthropomorphized dolls and stuffed animals as well as invisible imaginary companions. The present study adopted the latter definition. Before making an effort to illuminate the manifold forms and functions these pretend playmates have, an outline of what is already known about imaginary companions is presented.

Functions of Imaginary Companions

The make-believe friend may perform a large number of different functions. Many researchers believe that pretend playmates can have an adaptive function for children with relational and environmental deprivation, but the imaginary companion can also play an important role in ordinary children's development [1, 2]. Violations are easier to endure if one has an imaginary helper around with whom to share worries.

Some theoretical studies have pointed out different functions on the basis of children whom the scholars had met in their clinical work. However, there are very few systematic studies of the functions of make-believe friends, such as systematic interviews with non-clinical children. The richest presentations of proposed functions of imaginary companions come from clinical psychodynamic research, according to which the companion can have the function of an *id impulse,* a *superego* or *ego support* [8, 10], but also the function of different kinds of *defense mechanisms* [10, 11].

[1] In the text, the terms "imaginary companions," "make-believe friends," and "pretend playmates" will be used interchangeably to denote the phenomenon.

Imaginary companions may enter children's lives when they feel overlooked in different ways. They may appear at the birth of a sibling or be acquired to handle *family traumas,* such as divorce, sickness, or death [2, 8, 10]. The creation of a companion could also be a way to cope with loneliness, helplessness, feelings of being abandoned, fear of darkness, or the unknown. Imaginary companions can thus be likened to conscious *coping strategies* [1], as well as unconscious defense mechanisms [10, 11]. Bender and Vogel gave an account of a child who invented an imaginary companion to compensate for not having a normal family [8]. He accompanied his itinerant father and had few opportunities to develop stable relations to peers. It is not uncommon that the pretend playmate is a *rescuer.* For instance, one child, accounted for by Singer and Singer [1], had an elf that could conjure away monsters from the dark.

An imaginary companion has been proposed to be an important aspect of children's *development toward independence,* toward creating a social context of their own outside the realm of parental control. Such friends, at the same time, do not require the same amount of adaptation as do real friends [1]. The imaginary companion can constitute one way of *compensating for feelings of inferiority* in relation to adults or other children. Together with the make-believe friend, the child can experience equality that she or he feels neither in the relationship with parents nor with friends. The powerlessness of childhood can be alleviated [1]. The play with pretend playmates can also work as *social practice* for timid children [6, 12].

The pretend playmate can also fill the role of *scapegoat.* Singer and Singer mentioned one boy who punished his imaginary companion [1]. This was seen as a developmental step toward internalizing his parents' opinions.

There are also examples of identification where the imaginary companion plays a role as an *ideal self.* Bender and Vogel brought up a case where a boy with behavioral disabilities and motoric difficulties invented a make-believe friend who could jump over the Empire State Building [8]. A related role is when the companion is not subject to the same restrictions as the child [2].

The pretend playmate can also be a *protégé,* someone for the child to take care of. Myers gave an account of some children with psychological problems, whose imaginary companions were small and sickly and in need of their inventors' care [11].

A gender difference has been pointed out as regards the latter two functions. Harter and Chao found that girls protected their imaginary companions and that boys idealized theirs [6]. Girls' friends were more often less competent, weak, and tiny, whereas boys' companions were more often considered competent and worthy of admiration. According to Harter and Chao, this could constitute two different mechanisms for handling issues of mastery and competence [6].

Transitional Objects

It has been proposed that imaginary companions can facilitate developmental changes by working as transitional objects. The creation of a transitional object is an infant's first attempt to create a space between the parent and the infant. It constitutes the first developmental step toward an independent self [13]. Singer and Singer asserted that attachment to transitional objects is an antecedent to involvement in imaginary play and imaginary companions [1]. Sugarman and Jaffe propounded a developmental line of transitional objects [14]. The more advanced the developmental level, the more abstract the transitional object. In early life, the thumb constitutes a transitional object for the child and later a blanket may help to alleviate separation anxiety. At around three or four years of age, imagination and make-believe play become transitional phenomena. The last stage in the development of transitional phenomena contains art, music, and ideology, adults' places for relaxation between inner and outer reality.

Child Interviews as Research Method

Open interviews can be semi-structured or unstructured. In semi-structured interviews, some questions have been prepared beforehand. The interviewer, however, is not bound to these questions, but can flexibly pursue parts of the interview in depth. The unstructured open interview has no question specified in advance, only the subject of the talk [15, 16].

The credibility of children's self-reports has sometimes been questioned. However, a considerable amount of research has demonstrated that children are reliable as interviewees (see summary in [17]). Some scholars have argued that, by 10 years of age, children have certainly reached a developmental level where their verbal accounts are as reliable as those of adults. Among other things, children at this age have developed metacognitive thinking ability; they think more logically and have increasingly adult-like capacities for communication [17].

Critique against Interviews

However, even if 10-year-old children are as reliable narrators as adults are, many scholars have questioned the interview as a research method. Positivistic critics have maintained that qualitative inquiries are subjective, because the researcher is the instrument of both data collection and data interpretation and because interviews include personal contact between scholars and participants (summarized in [16]). Interview methodologists have argued that subjectivity cannot be excluded from science and have even contended that the focus on respondents' subjective everyday experience is one of the advantages of the interview method [16, 18].

Aims of the Study

One intention of this study was to acquire knowledge about the functions of imaginary companions through qualitative interviews. A second more general aim was to explore the variation in the phenomenon and to report unexpected findings arising in the process of categorization. A third aim was to relate the interviews—in addition to general research on imaginary companions—to modern self theories and social cognitive theory in order to create a complementary framework of interpretation as apart from the psychodynamic perspective within which the phenomenon has already been explored extensively.

METHOD

Participants

In a pilot study, 12 10-year-old children (7 girls and 5 boys) from an earlier study participated [19]. Twenty-six 10-year-old children (16 girls and 10 boys) who had or had once had imaginary companions took part in semi-structured interviews. These participants included all the children with make-believe friends (and with parental permission) in four classes.

Fourteen of the 26 children still had imaginary companions when the interview took place. Twelve accounted for their own past experience of having imaginary companions. The children's names used in this article are not the participants' real names.

Interviews

Both unstructured and semi-structured open interviews were employed in this study. There were several reasons for choosing interviews: The qualitative interview is a tool to facilitate better understanding of a phenomenon. Hypotheses are generally not formulated in advance and therefore interviews allow maintenance of a certain openness toward what is studied. According to Kvale, the qualitative interview is a sensitive and powerful method for capturing participants' experience and meaning in their everyday lives [16]. As the present work involves studying inner phenomena, it was considered important to gain the inner perspective of individuals' who had invented make-believe friends.

The pilot study involved unstructured interviews about imaginary companions. From these interviews, the semi-structured questions that guided the second round of interviews were generated (see interview guide, Appendix A). The first 29 questions were posed to all the children and were in principle the same questions as the children had answered previously in a questionnaire about imaginary companions (for some part, reported on in [3]). In this way, the children had a chance to confirm their answers or, on questions where they had failed to answer, they had a second chance to respond. Thus, some empty answer spaces

were filled. During the interview, further individual probes were made with respect to every child's specific responses. Besides these 29 more or less structured questions, there were a couple of open question areas concerning the ways of playing, the kinds of play, and the functions and roles of pretend playmates. All but six interviews were tape recorded and transcribed. The transcribed interviews and the interview notes were the basis for the qualitative analysis. Each interview lasted between half an hour and an hour and took place during school hours. The participants were informed that they could end the interview whenever they wished.

Procedure

In this study, I started with open interviews with 12 pilot-participants to develop a wide understanding of the phenomenon; in a second round of more structured (semi-structured) and focused interviews, 26 other participants were addressed. Thus, a procedure of repeated data collection was employed, whereby the first acquaintance with the phenomenon was allowed to influence the subsequent data collection, an approach that has been advocated by qualitative researchers [e.g., 20].

Analysis of the Interviews

Different scholars inspired the present qualitative analysis [e.g., 15, 16, 21]. The first step of the procedure was to note patterns and themes intuitively, then attempt to see connections between these followed. After this followed an attempt to make metaphors for what had been perceived. Then these tentative results were scrutinized by checking and counting whether the intuitive categories were reasonable. Another way of verifying or refuting the found themes was the process of comparing and contrasting different examples. If necessary, prematurely grouped variables were partitioned. Then followed an attempt to group variables hypothetically. The relations between variables were explored in this manner. Finally, work was undertaken to construct a pattern of evidence to reach conceptual coherence concerning the studied phenomenon. This work was supported by referring to earlier theoretical and empirical studies in the field.

To ensure the validity of the chosen themes, a co-assessor, working independent of me, also noted what she perceived as important themes in the interviews. The co-assessor's view was taken into consideration in the final presentation. She also scrutinized the transcriptions as regards possible interviewer effects, such as when the interviewer posed leading questions. The few identified confounding effects found were excluded from the results.

As regards presentation of results, the procedure advocated by most qualitative methodologists [15, 16, 21] is that the researcher's interpretations, substantiated with illuminative contemporary research, should be interjected in the results. The usual rule of value-free result presentation is deviated from because qualitative methodologists contend that this type of data material can never be completely value-free, and therefore it is not considered meaningful to separate the interpretation and the results [15].

RESULTS

The interviews have provided different kinds of data. First, some general facts about the children and their pretend playmates are presented as *Background material.* The other part of the presentation emanates from the analysis of the interviews, through which less evident facts appeared. The presentation contains two main themes: *About the Phenomenon* and *Functions of Imaginary Companions.*

Background Material

The children in this study show great variation as regards many facts. The age when they had had their pretend playmates varied, as well as the number of years they had them. Another fact sometimes connected to the occurrence of make-believe friends is the number of siblings. Most children came from fairly large families by Swedish standards, namely three-child families (17 children). Seven had only one sibling and four had three siblings. There were no only children within this sample, despite the fact that, in earlier literature, only children have been shown to have imaginary companions more often [1, 2]. Another fact showing variation was the type of imaginary companions. Seven children had animal companions, two had both animal and human companions, and seventeen had human companions. Three had anthropomorphized stuffed dolls and the others had invisible companions.

The median reported time that the play with the imaginary companion had lasted was three years. Appendix B presents some detailed facts about the pretend playmates.

Stories of Appearance and Disappearance

The most common reason for the appearance of pretend playmates was that the child felt lonely and therefore invented a make-believe friend to keep her/him company. But there was a wide range of different stories of creation among the participants. Some children did not acquire imaginary companions because of a certain deficiency. Several of the children had elaborate tales of their companions' appearance.

> Ida: It was when I was building a snow sculpture. Then, I made a small house and then it struck me that someone could live there and then I pretended that a mouse fell down from the sky . . . which was supposed to live there.

This is one example of how imaginary companions constitute a natural creative expression for some children. The urge to create can also make children construct intricate tales in connection to their companions, even in such cases where there is another primary reason for the invention, such as loneliness (see also section about paracosms). In Hoff [3], an association between a creative disposition and more elaborated imaginary companions was demonstrated.

Contrary to what other scholars have noticed [1], the 12 children in the present study who did not have a companion at the time of the interview did not have particularly elaborated stories of disappearance. The most frequent reason was that they had acquired actual friends or started school. Helga, who had her make-believe friend in her stomach, lost her companion when her mother became pregnant. According to Helga, she had wanted a sibling so much that when this was realized, the make-believe friend was no longer needed.

Influence

The sources the children reported as having influenced them to create imaginary companions were various. Several participants knew that their siblings had or had once had imaginary companions. Saga even played with her sister and a cousin with their different imaginary companions. A pair of twins had probably influenced each other more or less unconsciously, as they had the name "Kurt" in common. However, the twin Jakob said that he did not know whether his brother Oliver had a make-believe friend. Oliver, on the other hand, reported that he believed that Jakob had one.

Real playmates influenced some of the children to invent pretend playmates. Sibylla got her shared imaginary companion from Alma, who had told Sibylla and another child, Moa, that she had an imaginary companion.

Several children mentioned television as a source of inspiration for letting imaginary companions come into being. Rasmus saw a "Gremlin" film and decided that he wanted a pair of those as imaginary companions:

> Rasmus: Yes, it was because I saw Gremlins, it made me really scared. I was not allowed to do it for my mum and dad. I was home alone, and then I saw it. Then I thought that it was super fun with these Gremlins that sat and drank beer, and they almost fainted from the champagne. I could not sleep that night . . . then, I don't know, but I was afraid and sad and then . . . I only thought about those small, cute animals, then I thought that perhaps some were called Salt and Pepper.

According to Singer and Singer, television could be a source of inspiration for children's pretend play; however, they specified this to programs adapted to the needs of different age groups [1].

In Singer and Singer's presentation of the prerequisites for developing imagination [1], the source of inspiration provided by *imaginative others* was stressed as a main factor. Two other such sources *were places for play* and *literary inspiration*. As regards literature as an influence for the construction of imaginary companions, little evidence from these 26 children was found. They were all asked whether they had heard stories about pretend playmates, but few had. Nevertheless, imaginativeness as a general disposition is prompted by hearing many stories as a child [1]. Nearly all the children reported that they had frequently had stories read to them.

Location

There was some variation in the location where the imaginary companions were found. A majority of children only played with their make-believe friends in their rooms. However, there were also children who brought the companions wherever they went. Some children informed me that their invisible companions were with them during the interview.

Several of the children brought their imaginary companions to school. There were some who let their make-believe friends help them in school and some whose companions mostly got into mischief. Ida related what happened when her mouse "Knubbis" (Chubby in English) and her actual playmate's imaginary squirrel were in the classroom once: "Well, in the classroom we played with [giggle] 'Knubbis' and the squirrel. They drove around with their cars in the classroom and then we started to laugh because they drove up Susanna's [the teacher's] leg."

Some of the children explained that they mostly interacted with their pretend playmates when they were waiting to fall asleep. At this particular age, most parents have stopped reading to their children. The play with imaginary companions might work as an intermediary step before the child has learned to go to sleep alone.

Sometimes a child played with her or his imaginary companions in a small private cubbyhole. Some of the girls had "small houses" that they had found outside under some bushes or in the grass. Others had these more or less imaginary cubbyholes in their rooms under their desks and so on. A few of the boys had haunted places they had discovered together with their make-believe friends. For some children, the bathroom or the attic was a territory where they could be on their own and fantasize.

As mentioned in the previous section, Singer and Singer enumerated different prerequisites to developing an imaginative capacity [1]. Having a place to fantasize was one. According to them, a play space or cubbyhole may constitute a place where children can be in full control. In this way, children can construct a microcosm of their own.

The Ways of Playing

Single play—Nearly all the children gave different explanations of the nature of their play with imaginary companions. When asked about the advantages of a

pretend playmate, Ida explained: "It's that you can play with it wherever you are. It's never gone. It's like having a friend living inside me." Some of the children reported that their make-believe friends were always brought along. Others had them merely at specific places, for example in their own rooms. For several children, the companions appeared when they did not have a real playmate. Other children could not control whether the imaginary companion would turn up. Suddenly it was just there.

Joint play—A few of the children shared their imaginary companions with other children. For Sibylla, the joint play was complicated. She shared her pretend dragon "Drakis" with two other girls. They mostly played with it during the breaks at school. Alma, who had first invented this imaginary companion, seemed to be the "leader" of the game when they played together. Alma generally told the other two children what "Drakis" was up to. However, Sibylla could sometimes meet "Drakis" by herself.

It is interesting to speculate on the reasons for inventing an imaginary companion when real playmates are around. It can perhaps be a means for the inventing child to at least for a short while exert control over the mutual play. Relatively little research has dealt with this type of play.

Singer and Singer argued that refusal to share the imaginary world often indicates greater maturation [1]. Most of the mutual pretend play in the present study was indeed past experience. However, for some of the participants, it was reported to have occurred as late as between six and eight years of age.

About the Phenomenon

Awareness of the Imaginary Part of the Play

Many children could take a metaperspective on the imaginary play; this was revealed now and then during the interviews when they reported that they, themselves, controlled the imaginary companions or evoked the companions' parts of the dialogues. The children thus showed an awareness of the fact that the play was make-believe.

Responding to the question about who decided whether they should go out for a ride, Elsa said: "It was this rabbit, even though it was me who came up with it, I pretended that the rabbit came up with it."

Other scholars have commented that children are fully conscious of the pretense status of their invented friends. Nagera wrote:

> [L]ike ordinary day dreams the imaginary companion fantasy is an attempt at wish fulfillment of one sort or another, is ruled by the pleasure principle, can ignore the reality principle, and need not be reality adapted, yet the fantasizing person remains fully aware of the unreality of the fantasies that are being indulged in. In other words, reality testing remains unimpaired [10, p. 194].

A Delicate Subject Matter

Still having imaginary companions at the age of 10 or the mere thought of past companions appeared to evoke feelings of awkwardness in some children. They giggled intermittently when giving accounts of their make-believe friends, indicating that they believed it was embarrassing to indulge in such fantasies. Others described different occasions with their pretend playmates as awkward, for example when other people appeared unexpectedly. Dora had her companions in the bathroom. Responding to the question of whether she merely interacted with her imaginary companions in the bathroom, she said: "No, I don't often have them outside, I mean, I always have them in the bathroom. I mean, then my family comes or somebody in my family and says: 'What are you doing?' And then I feel ashamed."

However, not all children were embarrassed by spectators. Some children reported that their parents or other friends had noticed them while playing, but were not disturbed by it (see also section on Attention Seeking). Harald waited until the tape recorder was switched off to tell me that he still played with his imaginary companion during the summer vacation. While the tape was running, he maintained that he had stopped playing with the companion at the age of seven. For several of the children, having make-believe friends was profoundly private. They appeared to have different reasons for playing with their friends clandestinely. Dora appeared to be anxious about her being deviant because she had imaginary companions: "This is really very secret. . . . I have really not told anyone about this."

Another kind of evidence of the delicacy of the phenomenon was that some children kept certain aspects of their companions secret during the interview.

According to past research, older children are more likely to keep their imaginary companions secret [10, 12].

Functions of Imaginary Companions

One overall role of imaginary companions was that they were experienced as inner mentors, who appeared to assist the children in their identity formation work. A large number of separate functions was arrived at through analysis of the interviews. The five main categories were: *comfort or substitute for company, motivation and self-regulation, self-esteem enhancement, extended personality,* and *life quality enhancement* (for the frequence of answers in each category see Appendix C).

Comfort or Substitute for Company

The most common function of imaginary companions was to give comfort or to help endure boredom, loneliness, or fear of darkness. Nearly all the children

reported that this was a reason for having imaginary companions. Responding to the question of what pretend playmates are good for, Frida said:

> Frida: "Cause, I don't know, when you've, sort of, been frozen out by other playmates or when they don't want to play with you anymore than she can, sort of, be nice to you and help you and all.

Kohut emphasized children's need for "mirroring" [22], that is, children's need for self-objects (often parents) that accept and admire them. In middle childhood, playmates become increasingly important as support. When friends or parents do not show enough empathy and respect, a make-believe friend can perhaps partly compensate for this.

> Rasmus: I'm usually a little sad, and sometimes when I'm having a lie-in in the mornings, I usually talk to them. Especially when I'm home alone, don't like that very much. But you feel you're not completely alone any more when they come along.

Children's need for company, which hardworking parents can perhaps not satisfy, can be fulfilled through an imaginary companion. Pretend playmates might promote two different life span themes: the development of intimacy and that of autonomy [1, 23]. Make-believe friends may ameliorate both the ability to play alone independently as well as the capacity to interact with other actual friends. Children actually practice being considerate toward others through imaginary companions [24] (see also the following section of the present study).

Motivation and Self-Regulation

Many of these imaginary companions performed more complex functions. In some cases they were experienced as assisting the children in different school subjects. The make-believe friends functioned as school mentors:

> Amanda: I can't manage English class, no, I can't manage English.
> Interviewer: No.
> Amanda: But when you have one of those . . . pretend playmates then maybe you . . . go to a place and then to school and then you manage English there, then you believe in yourself, you see.

One way of describing this situation is that her imaginary companion "assisted" her in attaining her self-ideal—in other words, the make-believe friend was used in Higgins' sense for self-regulatory purposes [25]. Managing English was part of Amanda's self-ideal. There was a discrepancy between her actual self-feature (bad at English) and her ideal self. To reduce the discrepancy, she pretended that she traveled with her make-believe friend to different countries where she could practice English.

Several children also described situations where they experienced being coached to manage different tasks outside school. Saga accounted for a situation where her make-believe friends "prompted" her when she sang in a choir:

> Saga: Anyway they sing songs. It was five years ago, I think it was. "Cause I was maybe six years old or, or so, then we were going to sing aloud. We were going to sing, you see, we usually sing in the church, because I sing in the church choir, then I could, sort of, have all of them there [the imaginary companions]. So they stood there and sang beside me and all. Then I always knew that I sang correctly. And then, you see, I had papers and all. But I suppose that I couldn't read very well [. . .] And you could, when they stood there and sang, sort of, they whispered the words to you and that way you knew.

The illusion of having support seems to increase self-esteem, and with greater self-esteem, the chances of actually succeeding increase. This example can be compared with what Taylor and Brown contended about positive misperception of capability, namely that the illusion of feeling more capable than one actually has reason to believe one is fosters success [26]. Positive appraisals of the self are associated with perseverance with tasks, which in its turn produces more effective performance and increases the likelihood of success. In many situations, imaginary companions seemed to have self-regulatory and motivational functions that enabled the children to function better.

The more dynamically-oriented scholars Sugarman and Jaffe also argued that latency fantasies may assist children in their self-regulating in different ways [14]. For example, self-esteem regulation can be facilitated through identification with idols, teachers, and make-believe friends.

Some children explained that their imaginary companions had taught them to be more imaginative. The imaginary companions worked as "creativity consultants" for the children. Jakob described why it was good to have a pretend playmate:

> Jakob: No, he made me have a bit more fun. For example, he always made up, sort of, new games.
> Interviewer: Aha. He was pretty inventive?
> Jakob: Yes.
> Interviewer: And a little naughty?
> Jakob: But really, he has . . . I think, taught me to become more inventive, sort of.
> Interviewer: Aha
> Jakob: So I'm pretty inventive too, you see.

In a few cases, the companions appeared to make the children function better together with their real playmates. The imaginary companions acted as social coaches. Adrian stated: "If someone is kind of mad at me, I can tell him [the pretend playmate] that and that I want us to be friends again and then we become friends again." The make-believe friend acted as a mediator, who helped Adrian

make peace with the lost friend. According to Sugarman and Jaffe [14], one function of transitional objects is to facilitate object-relations development. The authors argued that in middle childhood an increased need to explore social relations arises and that children can practice the "nuances and subtleties" of such relationships in their play with imaginary companions. Gleason argued that make-believe friends afford practice in conceptualizing relationships [24].

Motivator for mischief—Some children accounted for their need for their pretend playmates to dare to be mischievous. This function of imaginary companions is an example of the degree of independence the children experienced in the companions. Here Harriet gave an account of what she would not have dared to do without "Kristina":

> Harriet: Then, once it was at Christmas time she made me . . . I had a chocolate calendar. [. . .] Well, there was a bed and a desk there, and then a wardrobe by its side. So, I went into the wardrobe and sat there. Then she said: "now, eat up all the chocolates!" [laughter]. And I did it. [. . .]
> Interviewer: Oh, yes a chocolate calendar . . . did you open all the doors?
> Harriet: And ate them all up [giggle].
> Interviewer: All the doors, and she made you do it?
> Harriet: Yes.
> Interviewer: Otherwise you would never have done it if she had not . . .
> Harriet: I would never have dared.

Here the imaginary companion works as a motivator to disobey parental rules, internalized in Harriet to the extent that she would never have dared to break them if not urged by someone else. Nagera maintained that imaginary companions can be vehicles for discharging unacceptable impulses [10]. On their way to becoming autonomous, children need to learn to decide independently whether to follow a rule and to internalize those parental prohibitions deemed necessary. Garbarino and Stott asserted that one important step toward realizing a separate identity is when children "learn" to lie [17]. Through lying, children perceive that parents cannot know or control their thoughts. Perhaps learning to be mischievous has a similar function.

A conscience or someone to provide moral guidance to—For some children, the imaginary companion had the function of a help-conscience. These companions halted the game when it had gone too far or told the children when they had done something wrong. According to Hans, the companion punished him if he did not pay attention to the companion's summons. The punishment was five minute's imprisonment in Hans' room. Helga related how her pretend playmate assisted her when she wanted to halt a game:

> Helga: It was sometimes maybe, for example . . . once we were going to escape . . . we were anyway going out and then she said: "Are we really going to do this?" Or something like that. Because I didn't really know if I wanted to do it or not.

This is another example of how imaginary companions assist self-regulation [25]. Imaginary companions perform the function of self-guides until these are completely internalized. From a psychodynamic view, Sugarman and Jaffe argued that an imaginary companion could foster the superego by means of working as a transitional step between external control and an independent superego [14]. Nagera designated this possible function of pretend playmates as "superego auxiliaries" and argued that many of the controls that adults demand of their children are often beyond their capacities [10]. Imaginary companions may, thus, help to solve self-regulatory conflicts [10]. According to Higgins [25], self-guides and the superego are closely related concepts.

A related function appeared to be very common, namely cases where the children were the more morally responsible. The children had to stop their companions when the play was going over the top. Responding to the question of who stopped an overly exaggerated game, Harriet said: "I use to do that. Then I said 'Now you will have to go home because I am tired.'" Miriam related how she had to stop her pretend playmates from pinching candy. "Mimmi" wanted to take the whole candy bag. However, Miriam only allowed her to take one piece of candy. Why are these imaginary companions so recalcitrant? One of the distinguishing developmental tasks for 7-to 10-year-olds—the age range during which most of the pretend playmates in this study were active—involves learning to play by the rules in games such as marbles [27], or today perhaps Pokemon or computer games. They should also have internalized well-functioning moral self standards or guides [25]. Perhaps the advantages of the more egocentric period are not given up willingly. In order to feel competent and in control, it might be easier for children to guide someone else than to themselves abide by all the rules. Even if they do not follow all rules themselves, it may feel good to know someone who is worse. This might be supported by Brown's finding that people tend to overestimate their own positive characteristics and depreciate those of others [28]. Inventing a mischievous friend will provide children with a "favorable" comparison.

Self-Esteem Enhancement

The imaginary companions could work as ego support in different ways. In an increasingly complex world, where more and more responsibilities are placed on children, an imaginary world can constitute a transitional space where they can temporarily find protection from feelings of powerlessness. In this transitional space, the children experience a sense of control that is not possible in the external world.

Externalization of negative characteristics—A rather common function for these pretend playmates seemed to be as an object onto which bad traits could be projected. These make-believe friends were frequently ascribed characteristics

such as "stupid," "lazy," and "cowardly." Here is an example of how Rasmus described his Gremlin companions:

> Rasmus: They're so chicken, because . . .
> Interviewer: Don't they want to go with you to school?
> Rasmus: Yes, of course they want to come with me to school. But they're chicken.
> Interviewer: In what way are they cowardly?
> Rasmus: They're . . . they don't dare, well, they don't dare in that way. . . but they don't dare to take the first step.
> Interviewer: No.
> Rasmus: Or I don't know. They're always worrying about new things.

Rasmus had few friends and one could surmise that he was not always happy about going to school, but with very nervous imaginary companions perhaps his own insecurity could be deemed insignificant. Within the social cognitive tradition, this way of functioning has been described as externalizing [29, 30]. Sugarman and Jaffe maintained that through externalizing undesirable traits onto the make-believe friend, the self-representation can be "purged of imperfection" [14]. This in its turn facilitates an idealization of the self, which will become a part of the ego ideal. Nagera argued that externalization was a mechanism predominantly used by younger children [10]. However, among the participants in the present study, the sequences of externalizing took place at various periods in the children's lives. More recent research within social psychology [28, see also 31] has demonstrated that, for individuals with high self-esteem, self-worth is bolstered by perceiving the self to be better than others, a phenomenon termed "self-deceptive positivity" by Paulhus and Reid [32] . Likewise, negative attributes are believed to be more descriptive of others. Perhaps imaginary companions can "help" children form "unrealistically positive views" [26] of themselves, similar to the views self-confident adults possess. Brown also demonstrated that, in adults, this self-other bias promotes psychological well-being. Perhaps children can derive the same benefits from a self versus pretend playmate comparison [28].

Scapegoat—If the make-believe friends were not ascribed unfavorable characteristics, they could be used in a more concrete way as objects on which to blame mischief. In this interview collection, there were only a few tales about bow the children used their imaginary companions for this function. Harriet argued that she invented her companion in order to have someone on whom to blame mischief: "Well, I dropped a plate or something like that and then it was always my pretend playmate who had done what I had done. Once [giggle] I had a pair of scissors and then I cut my bangs, and then it was she who had done it, sort of."

In this material, however, this function was not very frequent, something that can be explained by age factors. Restoring self-worth by blaming someone

else is predominantly used by young children [10]. Harriet reported having her companion between four and six years of age. The ability to appraise both positive and negative self-features may be difficult for preschoolers, and imaginary companions could be handy for relinquishing responsibility for mischief in order to maintain a feeling of self-worth.

Protégé—For some children, the pretend playmate constituted someone in need of care and supervision. At least one part of the play involved acting as the caretaker of their companions. They were the mothers or fathers of their imaginary companions. This function is associated with the functions of externalization and moral guidance. However, having a protégé highlights the caretaking role. Aron acted as a math teacher for his imaginary companion: "Well, I could count a little, so I taught him that." Saga accounted for her relation with and the responsibility she felt for her pretend playmates:

> Saga: Pipip was childish. He was like a small child. I was older than him. I was a bit big compared with him, but the other was more like a mother. I felt childish together with her. But I could feel a bit big too, you know when I was taking care of those two.
> Interviewer: Yes, you could.
> Saga: Well, and even if it was, sort of, an adult hen that was an adult in "hen years," I still had to take care of it. It could not manage on its own. [. . .] If we moved, for example, then they wouldn't know themselves "now we have to move up to that trailer." We must help them.

According to Harter and Chou, this role of imaginary companions could constitute a mechanism for handling issues involving mastery and competence [6]. Having a friend who is less competent may allow children to feel more masterful in comparison. Having an imaginary companion as a protégé may also be a precursor of the above-mentioned propensity among normal healthy individuals to rate positive characteristics as more descriptive of themselves and to ascribe negative features to a larger extent to others [28, 31].

Harter and Chou asserted that girls predominantly develop this kind of a relation with their imaginary companions [6]. Among the interviewed children in the present study, 3 out of 10 boys and 6 out of 16 girls described their make-believe friends as someone who needed their caretaking.

Self-ideal—Not only did the imaginary companions assist the children in attaining certain self-ideal standards (as shown in the section about motivation and self-regulation), but they could also constitute ideals themselves through their action or appearances. Rasmus expressed his opinion that pretend playmates were much more fun to be with than were real ones: "They are almost perfect. They are absolutely perfect."

The propensity to idealize imaginary companions has been discussed by different scholars. Sugarman and Jaffe asserted that self-esteem might be regulated

through the fantasy of an *idealized object representation* [14]. The invention of idealized objects might also be an attempt to provide oneself with idealized caretaking, which in its turn can compensate for parental deficiencies. Kohut used the term *idealized self object* and argued that children have a need to idealize primarily parents, but also others [22]. The function of idealization is also related to identification. Children identify with their idealized objects. In the absence of real persons to idealize and identify with, perhaps imaginary companions might do.

According to Nagera, having imaginary companions as an ego ideal can be observed in children who for different reasons feel rejected. Having a companion who is clever or strong may alleviate the feelings of rejection [10]. The imaginary companion impersonates primitive ego ideals that may be beyond the child's reach. Seiffge-Krenke contended that the omnipotent feelings of the child are projected onto the imaginary companions in order to experience some longed-for competence [33]. Harter and Chou asserted that using the companion as an ideal is more common for boys [6]. However, among the children in this study, there was an equal occurrence of idealization among boys and girls.

Better than real friends—Several of the children described their play with imaginary companions as better and more fun than play with real friends. It seemed easier for these children to enjoy themselves if they were in full control of their games. Perhaps they were not yet ready to adapt to other children in order to play. Adrian said about the play that: "It looks different and it is a little cooler." Helga was of a similar opinion: She reported: "It's different 'cause for example when you talk to her then she understands. If I was sad then she always understood me. [. . .] Playmates could also do that, in a way, but maybe not really as well . . . and it felt good to talk to her. She understood everything."

The dream of being completely understood can be discerned in this statement. It can be reminiscent of early childhood when parents' caretaking was administered as soon as a need arose, for example a need for comfort or food. For the small toddler, this well-matched caretaking could be experienced as being completely understood and it might constitute a state longed for later in life [23].

Harriet accounted for why she preferred imaginary play: "With the pretend playmates you can do what you want to do, if you see what I mean. With playmates, you have to do something real. Then you can't go home to the castle and such things."

The play with imaginary companions can be an escape, away from the demands of social adjustment placed on children in middle childhood. The possibility to be in full control for a while, as Harriet was in her "castle," might be a necessary transitional occupation on her way to getting used to adapting to others. Children might also report that they prefer playing with their imaginary companions to playing with real friends because they cannot get along with actual playmates [3].

Attention seeking—For some children, the (invisible) pretend playmate actually worked as a means of getting attention. Helga had an imaginary companion in her stomach called "Kristina" and used her for more obvious attention-seeking reasons. She told her real friends about "Kristina" in order to get attention. Other children related that they received attention from parents or other playmates when they spoke their imaginary play out loud. They were not bothered by (or they enjoyed) having others in the vicinity while playing. They revealed their secret friend more or less voluntarily when others asked them about their play.

Coping with tragic or horrifying events—Sometimes the need to feel more competent was observed in situations where children felt powerless in the face of tragic or frightening events. The pretend playmates could alleviate these aspects in many different ways and make the children feel more in control. Amanda told me how she and "Bambina" went to war-torn parts of the world and cured the wounded:

> Amanda: Really, I want to become a doctor and then I fantasize that I'm already a doctor, so I go down there and then I help them there.
> Interviewer: Aha, and what does Bambina do, then? Does she help sort of.
> Amanda: Yes, she can help.
> Interviewer: Aha, so you imagine that you're a doctor. . . . Do you help the wounded?
> Amanda: Yes
> Interviewer: And what does Bambina do then?
> Amanda: She can, sort of, cure them. She has also special things with her.
> Interviewer: What kind of special things does she have?
> Amanda: She has such different medicines that, well, she maybe pours them onto the wounds and then suddenly everything's all right again.

Another child, Ania, had learned about her mother's miscarriage, "revived" this younger sibling and had her as an imaginary sister, perhaps as a way to deal with the loss.

Several scholars within the field of social cognition have discussed self-illusions in adults as a normal way of coping with the world. Langer showed how adults maintained an illusion of control in a series of experiments where factors ruled by chance were appraised as controllable [34]. A parallel can be discerned in children's imaginary companions, which might be a way of maintaining a feeling of being in control. The suffering of people in a war-torn country was probably experienced as more endurable when the child felt she had a magical power to cure them, at least in her fantasy.

Another example of this function was exemplified by Karin, who had invented a paracosm—a secret archipelago consisting of different islands of which Karin had made a drawing. It had been elaborated for many years. It was an idyllic world where nobody died and where no wars were fought. Karin's sister and grandmother lived on the main island called "Hummel." A nearby island was

called "Humse" and on this animals were kept. "Amse," a third island—half of which was covered with a flower meadow—provided the inhabitants with food. The inhabitants of the island could transport themselves to different places with a magical "transporter," which made the trips instant and without the need for any vehicle. The teacher accounted for Karin's problematic home-situation and her lack of psychological support. Perhaps her grandmother was very important for her. Karin created this paradisiacal island where no one could die and where there were no difficulties, such as transportation or food provision. This imaginary island was a way of establishing some control and stability for a child with a difficult life situation.

Nagera expressed the opinion that imaginary companions could prolong children's feelings of omnipotence and control [10]. He argued that children's process of accepting their limitations of control is slow, gradual, and difficult. In this achievement, fantasy play has a critical role.

Extended Personality

Extended gender roles—For some children with imaginary companions of the opposite sex, the companion appeared to give an opportunity to expand the children's gender roles. Rasmus had one male and one female pretend playmate, and he played predominantly with the girl. She was the most active of Rasmus' two make-believe friends. The boy was described as lazy and slept a lot.

The girl was also described as bigger than the boy and was in many ways idealized by Rasmus. Besides the perhaps non-stereotypical sex roles given to a girl, Rasmus also accounted for a favorite spare-time activity that he used to enjoy together with his pretend playmates, more often with the girl than the boy. This activity was cooking and baking cakes, which could be an example of how his imaginary companions assisted him in extending his personality potential beyond stereotypical male sex roles.

Imaginary companions with opposite characters—For participants who had two (or more) imaginary companions, the two figures often had opposite characteristics, something that might entail experimentation with possible selves. Oliver, for instance, had two male companions with different personalities. "Mårten" was tough and rebellious with tough-looking clothes and orange hair, and "Kurt" was old-fashioned and dressed in a suit and bow tie. "Mårten" preferred going on adventures, but "Kurt" preferred playing quiet games. Oliver told about his two friends, who were also different as regards being supportive or needing support:

> Interviewer: When one of you did not dare, who said "Come on now, let's do it?"
> Oliver: It was Mårten, then it was often him who, sort of, "Come on now, it's not dangerous."
> Interviewer: Could it happen that you felt: "No, I don't want to?"

> Oliver: Yes, but I was often talked into it. But when it came to Kurt who . . . I often had to push him "Okey, it's not dangerous, come on now, Kurt," sort of.

This is an example of how imaginary companions could function as transitional objects needed by children on their way to integrating their own positive and negative features into one person. Working out the opposite characters of their make-believe friends might aid children in their own gradual process of integrating positive and negative features into more lasting dispositional self-traits [25]. Having pretend playmates that are one-dimensional may also constitute transitional objects, offering momentary escape from the complex personalities of real persons—complexities that children in middle childhood have begun to observe [35].

Life Quality Enhancement

Imaginary companions are also a part of the imaginative world, which provides life with some extra color, Such imaginary companions and paracosms could be regarded as precursors of creative potential [1] or creative expression of children [3].

Paracosms—Sometimes the imaginary companions were associated with elaborated imaginary worlds. These worlds, termed paracosms by Silvey and Mackeith (as cited in [1]),were imaginary constructions of many different kinds. The play with imaginary companions and paracosms in particular was not necessarily connected to any specific need or deficiency. It can be seen as l'art pour l'art—or play for the sake of play. It might just be that imagination often serves as a way to enrich life, a function just as important as any other.

Silvey and Mackeith (as cited in [1]; see also Cohen & Mackeith [36]) distinguished five different kinds of paracosms among their 64 participants: a) toys, animals, and family; b) particular places and local communities; c) islands, countries, and people; d) systems, documents, and languages; and e) unstructured, shifting, and idyllic worlds.

Of the 13 instances of paracosms in the present study, only Elsa belonged to group (a); her paracosm constituted a jungle where Elsa ran with wild animals such as lions, moose, and birds. Three children's (Harriet, Hilda, Sibylla) paracosms were grouped in category (b), as the homes of their imaginary companions, and five (Adrian, Frida, Nadja, Otto, Karin) were grouped in category (c), comprising worlds with their own people, flora, fauna, and utilities used by the inhabitants of these worlds. Some children visited these paracosms with their pretend playmates. Frida had an intricately thought out world, that she described in the following way:

> Frida: You know it's sort of like that, that there are different, you know, there are a great many trees and grass and such things and then there's a gorgeous sea.
> Interviewer: Aha

Frida: And then there're different trees that are, you know, for example there's a tree which is, I don't know . . . like spaghetti, where you can sort of go and eat and so on.
Interviewer: Yes, a spaghetti tree.
Frida: Yes, and then there's, sort of, a small lake which is . . . It's filled with marmalade [giggle].
Interviewer: Yes, a lake that's made of marmalade.
Frida: And then, there's a tree that is ice cream.
Interviewer: Aha
Frida: And then if you go further away, then you meet some other, sort of, guys . . .
Interviewer: That also live in this world?
Frida: Yes.
Interviewer: What kind of guys are they?
Frida: You know, they have pretty big noses and they're very small. They have big feet and then they're very friendly. And after them comes one of those woods, sort of, and there they always say "No, don't go there because stones fall down there." And, that is, you know, jewels and all.

Adrian's companion arranged the trips to "the secret country," always with the departure-point in Adrian's room. With the imaginary panther's magic power, it took five seconds to go there. The landscape there did not look like our planet. Trees grew in the ocean. They jutted up above the surface. On the surface of the ocean there also grew flowers. In "the secret country" people lived on fruit and there were no apparatuses there.

In this sample, there were some difficulties in discerning the categories of Silvey and Mackeith. One example could belong to several. There were two possible instances of category (d). However, as these elaborated languages were parts of imaginary worlds, they could also be grouped under (c) (Hilda, Nadja). There were three representatives of (e) idyllic worlds (Frida, Nadja, Karin), which were also grouped under category (c) as these countries and islands were places where people could not die and where no wars were fought; they were also intricately elaborated worlds. As opposed to idyllic worlds, Ania played "the poor game" with her imaginary companion. It constituted a dark paracosm that took place in a rundown shed with rotten walls. In addition to the categories proposed by Silvey and Mackeith (as cited in [1]), I would like to suggest a sixth category, namely paracosms where the activity was most central: Action paracosms. Three of the participants described different kinds of worlds where certain activities were undertaken. Oliver had an imaginary haunted house, in which he sneaked around with his imaginary companions as a kind of adventure. Dora had four different people who were policemen and detectives who hunted thieves and solved mysteries. These constituted a world of their own, but the things around them were not particularly elaborated. The third example was Amanda who traveled to different countries and made poor countries look better and healed the wounded in war-torn places.

DISCUSSION

The present research demonstrated a multitude of forms that imaginary companions may have. Some findings were replications of earlier studies, such as the types and names of the companions [1, 2]. For example, there were individuals from "the little people," full-sized wild animals, and human children, The names were both creative inventions and ordinary human names. Several children had intricate stories of creation for their make-believe friends, but in contrast to earlier research no elaborate stories of disappearance were found [1, 2]. The other findings have not been elaborated in past research. The invention of a pretend playmate was predominantly influenced by actual friends, siblings, and television. The most frequent location of play was the children's own rooms. However, there were also children who brought their companions wherever they went, and others who had them in private cubbyholes. The imaginary quality of the companions did not seem to set any limits on the possible contents of play. The play with imaginary companions could both be enjoyed alone and with actual friends.

The Functions of Imaginary Companions

A multifaceted array of functions was discerned. Earlier the most cited presentations have had psychodynamic roots. In this presentation, self-theoretical and social cognitive frames have also been applied to the functions of pretend playmates. In a broader sense, the play with imaginary companions was found to be an identity forming activity, in which the make-believe friends worked as inner mentors. Five main functions were identified, each having several subfunctions:

Pretend playmates (i) compensated for a lack of comfort and company. Imaginary companions also performed (ii) the function of assisting *self-regulation* as well as being a *motivator* for both moral (in this sense working as a help-conscience) and mischievous deeds. The make-believe friends provided "cognitively oriented therapy" to overcome social problems, were "creativity consultants" and "school mentors," in addition to many other supportive functions. Imaginary companions also performed the function of (iii) *self-esteem enhancement.* Seven subcategories were identified: a) unfavorable features were externalized on make-believe friends. The companions were used as: b) scapegoats on whom one could pass on blame; c) as protégés to increase feelings of competence; and d) as self-ideals with whom to identify. Another strategy for maintaining self-respect was to: e) regard the pretend playmates as better than real friends in order to feel socially adequate. The make-believe friends were: f) means for seeking attention; and g) means for coping with tragic or horrifying events. Furthermore, the imaginary companions were utilized to (iv) try out alternative personality traits and non-stereotypical gender roles. Finally, make-believe friends and their worlds simply functioned (v) to enhance life quality, just as imaginary and pretend play often do. They were created for the fun of experiencing something that is not possible in reality. Pretend playmates were l'art pour l'art.

An interesting question is why imaginary companions have all these functions. Given that the children were the inventors and that they were fully conscious of the make-believe status of their companions, it is remarkable that the children still contended that their companions assumed active parts in all the observed supportive and coaching functions. They obviously experienced that it was easier to achieve different objectives with imaginary support.

One may speculate whether school-aged children are considered self-managing by parents and teachers, while the children themselves would prefer having less responsibility or at least receiving all the support and encouragement given to younger children. Might imaginary companions be an indication that too large a burden of responsibility is placed on too small children? Or should imaginary companions be considered normal transitional objects providing alleviation in a developmental shift? It was obvious that many children needed some extra help until they could grow into their oversized clothes. Having imaginary help was better than no help.

Illusion Forming

Many scholars have emphasized that healthy adults make use of self-deception—or adaptive illusions—in order to function well. Could there be a parallel between adults' positive illusions [26] and children's imaginary companions, which allow children to maintain their self-esteem? Taylor and Brown asserted that self-enhancing strategies promote psychological well-being [26]. Even though some research has demonstrated that children with imaginary companions are less socially well functioning [5, 6], and reported having more relational difficulties with friends [3], it can be fruitful to speculate on how these children would have coped had they not had these companions at all. Bouldin and Pratt found that preschoolers with imaginary companions had more anxiety compared with their peers [37]. However, the anxiety levels were within the normal range. Still, the anxiety might be regarded as so stressful as to warrant the formation of a pretend playmate to assist the children in the difficult situation. As I see it, by inventing make-believe friends children may cope with their anxiety in a constructive and possibly creative way, in a similar way as positive illusions work for adults.

One suggestion concerning how the illusion mechanisms operate is that individuals assimilate contradictory, negative, or ambiguous information into preexisting positive self-schemata. Another explanation is that positive illusions are maintained by social and cognitive filters that distort negative information. Having passed through these filters, negative information may be considered as not having general implications for the self [26]. For children with imaginary companions, this would mean, for example, that the signs of having difficulties with actual friends are distorted and that the positive experience of the imaginary friendship is exaggerated.

Some people may object to the idea of self-illusions among healthy individuals and ask where the impetus for growth and change is. If individuals believe they are perfect, why should they change? Taylor and Brown argued that positive illusions "may inspire people to make changes that might be avoided if the uphill battle ahead was fully appreciated" [26, p. 59]. They pointed out three kinds of adaptive illusions: unrealistic optimism about the future, an exaggerated sense of control, and excessive self-confidence. According to Paulhus and Reid, self-deception can be divided into two different tactics: enhancement of positive self-features and denial of negative traits [32]. Through factor analysis with several well-known self-assessment tests, the positively valenced characteristics (e.g., "I'm a saint") were related to adjustment, whereas the negative attributes (e.g., "I'm not a sinner") were related to impression management—that is individuals' wishes to appear better than they are [32]. If perceiving imaginary companions as instances of self-deception, children may use their imaginary companions for both adjustment and impression management purposes.

Imaginary companions might be predecessors of these adult forms of self-deception in many different ways. If children with social or regulatory difficulties had fully appreciated what the required adaptation in middle childhood entails, they might not have endured. That is, if they had not had the opportunity to go via imaginary companions (and experienced the illusion of being socially competent or capable of self-regulation) and gradually acquired a capacity for social adjustment and self-control, they might have given up. In Fraiberg's words:

> It can be demonstrated that the child's contact with the real world is strengthened by his periodic excursions into fantasy. It becomes easier to tolerate the frustrations of the real world and to accede to the demands of reality if one can restore himself at intervals in a world where the deepest wishes can achieve imaginary gratification [38, p. 23].

In opposition to these observations stands the clinical wisdom that people must appraise reality in an unbiased fashion in order to survive. How can these positive illusions be functional in the face of contradictory evidence from the environment?

In some respects, the social cognitive perspective on the area of self-illusions contradicts the psychodynamic view. Individuals do not become more and more realistic in their perceptions of reality—they do not completely leave the pleasure principle in favor of the reality principle, as Freud might have put it [39]—at least not healthy individuals, who rather continue to indulge in adaptive self-illusions throughout life [26]. Of course, individuals do need some realistic foundation so as not to be deemed psychotic. But if they are to remain happy, they should not be too objective when evaluating themselves. Depressed individuals were found to appraise their negative features and their lack of control in a less biased (more realistic) fashion than did healthy individuals. There are studies demonstrating that most non-depressed adults have less-than-realistic beliefs about their personal control over their environment [26, 31, 34]. My suggestion is that the imaginary

companion may be an early predecessor of this adaptive self-illusion and may assist children in establishing a sense of control. This idea, however, is contradicted by the fact that self-deceptive adults are described as well functioning, whereas children with imaginary companions have been found to have negative self-images, particularly regarding their mental well-being [3]. Does the imaginary companion illusion not work as well as the adult illusions described by Taylor and Brown [26]? Will the capacity for creating adaptive illusions develop further after the age of 10, such that individuals who had imaginary companions as children will have an advantage in terms of psychological health? Or are adult illusions and children's imaginary companion illusions unrelated? These questions need to be addressed through longitudinal research efforts.

Life Span Themes

Imaginary companions assist in several life span developmental tasks. According to many scholars, the *development of intimacy* and that of *autonomy* constitute two different life span themes [1, 23]. Stern maintained that, from the very beginning, infants construct schemas of themselves as differentiated from others [23]. Many different kinds of behavior can be described as signs of *autonomy;* for example, from averting the gaze or head at four months of age, to the more vociferous protests that adolescents often display in their struggle for independence. The development of *intimacy* can be discussed both in terms of attachment and in terms of sharing subjective states. Throughout life, people need attachment to other people. Imaginary companions can provide a means for dealing with both the development of *intimacy* and the development of *autonomy.* The ability to play alone for longer periods is gradually developed from the age of three [27]. This independence may evoke positive feelings for both children and parents. However, from time to time, parents may impose demands on their children to play alone when the children would rather have had company. In a similar way, school-aged children are expected to handle being home alone, which might sometimes be fun. However, sometimes the demands on children to be autonomous can be experienced as difficult to live up to. The invention of a make-believe friend may then alleviate children's longings to belong to someone or constitute a transitional phase during which children get used to being on their own, a capacity needed throughout life.

For other children, the imaginary companions functioned to make them more capable of interacting with other actual friends and of experiencing intimacy. The present results indicated that make-believe friends did not only have a social compensatory function, but also seemed to afford *social practice* within the child's own control. In her study, Gleason demonstrated that the social provisions of imaginary companions were similar to the provisions of real friends, indicating that relationships with real and pretend playmates may encompass a single cognitive schema [24]. She also argued that imaginary companions provide practice in negotiating and conceptualizing relationships. More over, the make-believe friends in this study were sometimes experienced to assist the children in overcoming social problems, for example prompting the children to make peace with their actual friends.

Many statements represented children's dreams of *being completely understood,* which is a third life span theme. This might be reminiscent of the proverbial stage of life [23]. Now and then, the parent may match the infant's wishes so well as regards provision of food when the infant is hungry and comfort when that is needed, that the child experiences being perfectly understood. The speaking child, on the other hand, often finds herself/himself in a situation where language is not enough to express certain thoughts and needs, and a sense of emptiness can be felt. This results in a dream of being understood more completely. Individuals of all ages can experience the discrepancy between what they want to express and what they can express given the general limitations of human language. In relation to imaginary companions, children can re-experience the feeling of total understanding.

Transitional objects are also a life span theme connected to imaginary companions. According to Winnicott [13], transitional objects are needed to create a place between inner and outer reality. The transitional world (and transitional objects) borrows material from both outer reality and the internal world. It exists in different forms throughout life. A transitional object in early childhood often exists in the real world in the form of a rag or a doll; however, to the child, it has a far greater internally constructed symbolic significance. In middle childhood, imaginary companions may constitute transitional objects. Finally, in adult life, cultural phenomena represent the transitional world. Even if imaginary companions often lack a reference in the real world, they are still more than mere inner experiences. The play often has external consequences, for example, the child experiences having "company" in real life occupations such as picnicking and playing cards. For some, the make-believe friend is involved in mutual play with actual playmates.

Developmental Differences

Imaginary companions exist in different age groups throughout childhood and adolescence and therefore one may suspect that make-believe friends perform different functions depending on the age of their inventors. In the present study, all interviewed children were 10 years old. Therefore no systematic comparison between different age groups could be made. However, some children accounted for their past experiences of pretend playmates and therefore the age differences reported in past research can be discussed in relation to the interview results.

Scapegoating is purported to be an immature kind of defense [10, 30], and, indeed, the interviewees reported few instances of this function. One of two examples of scapegoating reported in this study was reported to have taken place

between four and six. The other was a contemporary report, that is, took place at the age of 10.

Some scholars have maintained that older children predominantly keep their imaginary companions secret, whereas younger children are more likely to play overtly with theirs [1, 2, 10, 12]. Approximately half of the children in this study had told someone about their make-believe friends, but no differences were discernable between those who had companions earlier and those who had them presently. Nagera argued that the functions of older children's companions were more conflictual, which was supposed to be the reason for the secrecy. Among the children in the present study, these differences were not able to perceive. Many of the 10-year-olds did keep their companions secret, but they were there for giving love, affection, and comfort to the children, as well as serving as "someone" in whose company a sense of control (omnipotence) could be enjoyed.

In past research, shared pretend playmates have been described as indicative of younger children [1]. However, among the present participants, some shared their companions as late as at eight years of age.

To demonstrate possible age-specific functions, more work in different age groups or longitudinal studies are required. Moreover, the quality and possible change in the child-companion relationship could be followed in those cases where the companions remain with a child for several years. Another question is whether the characters of the make-believe friends develop over the years children have them.

Validity and Reliability

Methodological literature advocates different ways of securing the validity and reliability of interviews. One such method used in the present study was triangulation. In addition to myself, another researcher read through the transcribed interviews. The goal of this co-assessment was to obtain a second view on whether the result categories were representative of the interview material as a whole. The co-assessor also checked for interviewer mistakes such as leading questions. Another method used for scrutinizing the reliability of the interviews was to check for inconsistencies in the material. Did the children say things that could be confirmed in other ways? For example, there were some children who either had played with each other when they were younger or did so at the time of the interview. There was also a pair of twin boys who spoke about each other. All these stories matched, which was taken as an indication of reliability.

A third method involved taking different critical metaphors into consideration [40]. Alvesson advocated reflection concerning how *the context,* the *interviewee's interests,* and *language* influence the outcome.

(a) The context of the interview influences the answers. Preconceptions about the research project affect what the respondents present. The different, more or less powerful, roles of the interviewer and the interviewee influence the outcome.

Applicable to the context of the present study is the need to consider what happens when children speak to adults. They probably omit things that they discuss with their peers, and include other things that they would not tell a peer. When an adult addresses a child, the same selection probably occurs. However, it is difficult to foresee the impact of these factors. I tried not to play the role of a teacher, but as a woman in her thirties, I most likely resembled a teacher, though. Indeed, during the interviews, the children reported very little naughty behavior on their own part. The mischief reported was mostly harmless and some of the children actually emphasized that they did not make really wicked mischief, only the kind that does no harm. Few taboo topics were brought up, such as sexual themes or themes of companions as objects of romantic love. This selection might depend on the context, rather than on the fact that there were no such companions.

(b) The interviewee has an interest in presenting herself/himself in a positive way. The respondents might feel constrained by imperatives to express themselves in ways that maintain loyalty to their group. Included in this might also be a wish to appear competent or perhaps "not odd," something that Alvesson called "moral story-telling [40]. The above-mentioned selection of topics and children's efforts to present themselves as kind, normal asexual children might depend on moral story telling. They might have felt like representatives of their age group and therefore wanted to please the adult, teacher-like interviewer.

(c) Language can also include constraints for "truth-telling." Language is part of the construction of reality. Participants are believed to be easily influenced by verbal and non-verbal nuances expressed by the interviewer and can be led to answer in specific ways. There are probably such examples in the present study. However, in the interviews, evidence can also be found where this did *not* take place. Unfortunately for me as an interviewer, these proofs were sometimes found when I went beyond my role as an interviewer by asking leading questions. The kind of answer that I was looking for must have been evident even for 10-year-old children. But on many of these occasions, it was obvious that the children did not feel forced to answer in a specific way. If they did not relate to the situation I had imposed upon them, they just said "no, it was not like that." The information was omitted for the few occasions when the co-assessor felt that I had influenced the answers.

Considering the confounds that often interfere with what scholars intend to investigate, Alvesson proposed that the contribution a scientist can make is merely to provide "uncertain, but sometimes interesting clues for the understanding of social reality" [40, p. 23].

Conclusions

Through in-depth interviews, greatly varied descriptions of all kinds of imaginary companions were obtained. Furthermore, pretend playmates were found to be useful in many situations in assisting identity formation in middle childhood.

Earlier studies have pointed out several functions that imaginary companions may perform, such as providing comfort or company and moral guidance, as well as being "someone" to nurture, idealize, or project bad features onto [1, 2, 6, 10]. Apart from these functions, this study presented some functions not identified earlier, that is, children who had make-believe friends for attention seeking purposes and for personality expansion, such as gender role experimentation. Few scholars have attempted to systematize the functions. In this study, five main categories of functions have been proposed, including three categories of self-reinforcing roles: comfort or substitute for company, motivation and self-regulation, and self-enhancement. In addition to these, there was a fourth function, namely experimentation with personality expansion and a fifth, where the companions simply appeared to constitute life quality enhancement. The ability of imaginative children to invent an inner device that provides them with the psychological and emotional support that their outer environment has failed to adequately provide is a fascinating phenomenon. However, there are many questions within this field of research that still need to be investigated. In addition to the research tasks proposed earlier, the question of whether a child who has suffered psychological distress could, for one reason or another, gain psychological health through invention of an imaginary companion remains for future research to address.

Some ideas from modern self-theories and social cognition were applied to the different functions of imaginary companions in order to provide a complementary framework to the psychodynamic. For example, instead of discussing imaginary companions in relation to the id, the ego, and the superego, concepts as "a motivator for mischief," school and social motivation, and help-conscience was applied. As a complement to theories of defense mechanisms, adaptive illusions were discussed in relation to imaginary companions.

APPENDIX A
Interview Guide: 29 Structured Questions and Thereafter Open Questions

1. Do you have any brothers or sisters?
2. How old are they?
3. Do you know if they have had pretend playmates?
4. How long did you have it? Is it still around? How did it come to you? How did it disappear?
5. What do you call your pretend playmate?
6. Is the pretend friend A) a stuffed doll or a toy or B) invisible?

A. If the pretend friend is a stuffed doll or a toy:

7. How many?
8. What kind of a toy is it?

9. Sex?
10. Child or grown-up?
11. Does it have a voice of its own or is it your voice you hear?
12. How is this stuffed doll distinguished from other stuffed dolls?

B. If the pretend friend is invisible:

13. How many?
14. What does it look like?
15. Animal or human?
16. Size?
17. Child or grown-up?
18. Sex?
19. Does it always come in the same form?
20. Do you both see it and hear it as a voice or do you just hear it or just see it?

All pretend friends:

21. Does the pretend playmate have its own friends, relatives, and parents?
22. Do you play with other real playmates together with your make-believe friend?
23. How do you feel when you play with your pretend playmate (sad, bored, happy, etc?)
24. How do you think your pretend playmate feels (sad, bored, happy?)
25. Where are you when you play with your pretend playmate? Home? At school?
26. Do you play in a pretend world? Did you make up imaginary houses, woods, or anything that you played with?
27. Do you play with your real toys with the pretend playmate?
28. Have you told anybody about this pretend playmate?
29. What more do you know about your pretend playmate?

When does this pretend playmate show up? Anytime or when you want it to? Is it only when you are alone the pretend playmate appears?

What makes this pretend playmate different from other real friends?

What do you do when you meet? Adventure, mischief, talk, adult stuff?

When you talk, do you talk like adults or like children do? Do you share secrets?

Why do you think this pretend playmate came to you? Why is it good to have?

Roles: Who decides when you should meet? Who comforts whom? Who urges the other along? Who makes up the mischief, adventures? Who says so when the play has gone too far? Who says what is right to do? Have you ever learned anything from your pretend playmate or can she/he help you become a better person?

Were you told stories as a child? Many or few? Were there any about pretend playmates?

APPENDIX B
Descriptions of the Imaginary Companions

Name of participant	Name of i.c. (sex of i.c.)	Characteristics of i.c.	Contents of play	Influence	Play location	Secret i.c.
1. Rasmus	Salt (m) Pepper (f)	Gremlin, slept much Brisk, mischievous	All kinds, e.g. cooked, mischief	TV-show: Gremlins	School, own room, in attic	Yes
2. Oliver	Mårten (m) Kurt (m)	A cool guy, green hair Wimpish with a bow-tie	Adventure in imaginary ghost house, mischief, football	Sibling	Own room, garden	Yes
3. Ida	Knubbis (m) Josef (m)	A mouse walking on two legs from Mars Human boy	Played with actual mate related to squirrel Played with actual mate	Playmate	School, wardrobe and in classroom	No
4. Lisa	Michelle (f) Maria (f)	5 yrs, blond, blue eyes talked, adventure 6 yrs, brown hair and eyes	Played party games Talked and watched TV	Playmate	Home, at breaks, at school, cubbyhole, home	No
5. Helga	Josefin (f)	Inside the stomach, brown hair, blue eyes	Talked, H told actual mates about J in the stomach	TV, longed for sibling	Own room, outside home	No
6. Harald	— (m)	Like H	Played with toy cars	Sibling	Home inside	—
7. Frida	Nicki (f) Kicki (f)	Shy but also talkative, dark hair, had fiancé Her sister's i.c.	Talk, mischief, went to fantasy world with jam lakes and spaghetti trees	Sibling	At bedtime mostly, also other school, home	Yes
8. Jakob	Kurt (m)	Human boy	Picnicked, played with cars	Sibling	At bedtime, other	Yes
9. Harriet	Kristina (f)	Long red and white skirt	Cooked, mischief, played in imaginary castle		Bedtime, cubby-hole in own room	No
10. Saga	Kaka (f) Pipip (m) "Nobody" (f)	Hen clean, cartoonlike Chicken, mischievous, dirty, cartoonlike Twin sister	Talked, jumped in bed, drew pictures, mischief, played party games	Sibling	Anywhere	No
11. Herbert	— (m)	Boy, could fly	Mischief		Own room	Yes
12. Aron	Kalle (m)	Boy, blue clothes, kind	Mischief		Home, outside	No
13. Amanda	Bambina (f) Olof (m) Pelle (m)	Tiger, had relatives Tiger kitten Dalmation	Flew to different countries and healed		At bedtime mostly	Yes
14. Dora	Marko I (m) Tannya (f) Linky (f) Brunetten (f) Spinky (m)	D's hand, d's son D's hand, D's daughter Blond Brown hair Thin	Talked, D took care of i.c., D listened and watched what i.c. did		Bathroom, own room	Yes
15. Hilda	Madeleine (f)	Long hair, short skirt mischievous, kind	Played in i.c.'s house and garden, own language		Own room, home	Yes
16. Elsa	Matilda (f)	Rabbit, happy	Flew to jungle on carpet, ran with animals, mischief	Playmates at kindergarten	Own room, outside in cubbyhole	No
17. Miriam	Mimmi (f) Cissi (f)	Clever, kind, a tease Kind and mischievous	Talked, played with toys, cooked, cleaned the dishes	TV	Own room, in a cubbyhole	Yes
18. Magda	Delfi (m)	Dolphin				Yes
19. Adrian	Did not want to tell (m)	Panther	Performed tricks, played ball, went to "secret country" with talking teddies and butterflies		Home, school	Yes

APPENDIX B (Cont'd.)

Name of participant	Name of i.c. (sex of i.c.)	Characteristics of i.c.	Contents of play	Influence	Play location	Secret i.c.
20. Jesper	— (m)	Coon				Yes
21. Ania	"You" (f)	Looks like A	Played the "poor game," picked food in field and ate it	Miscarriage of her mum	Own room	No
22. Sibylla	Drakis (m)	Dragon, only those with brown eyes could see	Played "hide-and'seek," talked and played cards	Playmate	At school, home	No
23. Hans	Peter (m)	One of the little people i.c.'s brother is a bear			At school in bag	Yes
24. Otto	— (m)	Human	Talked, played in pretend country with coconut palms		Own room, home	No
25. Karin	— (f)	A secret archipelago Hummel, Amse, Humse	Played in the music room, transported between spots	Playmates		Yes
26. Nadja	Kristina (f) Malin (f)	Brown hair, cartwheeled	Went to "Fantasia" situated in the clouds, own language	Playmate		No

Note: I.c. means imaginary companion.

APPENDIX C
Functions of Imaginary Companions

Company/Comfort		Regulation/Motivation						Self-enhancement					Extended person		Life quality
		Influence of i.c. Good	Bad	Conscience/ s.o. to guide Conscience	Guide	Object of projection	Scape-goat	Protégé	Self-ideal	Better than real friends	Attention seeking	To cope with the tragic	Extended gender role	Opposite character	Paracosms
1. Rasmus	x	a	x	x	x	x	x	x	x	x		x	x	x	
2. Oliver	x	a	x		x	x		x	x	x				x	x
3. Ida	x	c							x	x	x		x		
4. Lisa	x	a							x					x	
5. Helga	x	c		x		x		x		x	x		x		
6. Harald	x														
7. Frida	x	b	x		x	x		x	x	x					x
8. Jakob	x	c			x	x					x				
9. Harriet	x	c	x		x		x			x					x
10. Saga	x	b	x	x	x	x		x			x			x	
11. Herbert	x								x						
12. Aron	x							x							
13. Amanda	x	b,d							x	x		x			x
14. Dora	x				x	x		x	x	x			x	x	x
15. Hilda	x														x

APPENDIX C *continues on next page*

APPENDIX C (Cont'd.)

Company/Comfort	Regulation/Motivation						Self-enhancement						Extended person Life quality		
	Influence of i.c.		Conscience/ s.o. to guide		Object of projection	Scape-goat	Protégé	Self-ideal	Better than real friends	Attention seeking	To cope with the tragic	Extended gender role	Opposite character	Paracosms
	Good	Bad	Conscience	Guide										
16. Elsa	x						x							x
17. Miriam	x	x		x				x						
18. Magda	x													
19. Adrian	x	b						x	x					x
20. Jesper	x													
21. Ania	x		x	x					x		x			x
22. Sibylla	x	b			x					x		x		x
23. Hans			x	x										
24. Otto														x
25. Karin	x										x			x
26. Nadja	x										x			x
Girls	16	8	5	3	7	1	5	6	7	4	4	4	3	10
Boys	8	4	2	2	4	3	1	3	3	1	1	1	2	3
Total	**24**	**12**	**7**	**5**	**11**	**8**	**2**	**8**	**10**	**5**	**5**	**5**	**5**	**13**

Note: "Good" and "bad" below "Influence" stand for positive and negative influence that the imaginary companions had. "Conscience" below the heading "Conscience/someone to guide" stands for help-conscience and "Guide" for someone to provide moral guidance.
a = Motivation to dare to do different tasks, b = Social coaching, c = Inspiration to become more imaginative, d = School mentor.

ACKNOWLEDGMENTS

The author would like to express her gratitude to the participating children, and to their teachers who have provided time to meet the children. She would also like to thank her mentor Ingegerd Carlsson for her support and Erica Fäldt for her help with the cojudgment. The staff of the Division of Developmental Psychology are also thanked for their general advice concerning the article.

REFERENCES

1. D. G. Singer and J. L. Singer, *The House of Make-Believe: Children's Play and the Developing Imagination*, Harvard University Press, Cambridge, Massachusetts, 1992.
2. M. Taylor, *Imaginary Companions and the Children Who Create Them*, Oxford University Press, New York, 1999.
3. E. V. Hoff, *Imaginary Companions, Creativity, and Self-Image in Middle Childhood*, manuscript submitted for publication, 2004.
4. I. Seiffge-Krenke, Imaginary Companions in Adolescence: Sign of a Deficient or Positive Development? *Journal of Adolescence, 20*, pp. 137-154, 1997.
5. P. Bouldin and C. Pratt, Characteristics of Preschool and School-Age Children with Imaginary Companions, *Journal of Genetic Psychology, 160*, pp. 397-410, 1999.
6. S. Harter and C. Chao, The Role of Competence in Children's Creation of Imaginary Friends, *Merrill-Palmer Quarterly, 38*, pp. 250-363, 1992.
7. M. Manosevitz, N. M. Prentice, and F. Wilson, Individual and Family Correlates of Imaginary Companions in Preschool Children, *Developmental Psychology, 8*, pp. 72-79, 1973.
8. L. Bender and B. F. Vogel, Imaginary Companions of Children, *American Journal of Orthopsychiatry, 11*, pp. 56-66, 1941.
9. M. Svendsen, Children's Imaginary Companions, *Archives of Neurology and Psychiatry, 32*, pp. 985-999, 1934.
10. H. Nagera, The Imaginary Companion. Its Significance for Ego Development and Conflict Solution, *Psychoanalytic Study of the Child, 24*, pp. 165-196, 1969.
11. W. A. Myers, Imaginary Companions in Childhood and Adult Creativity, *Psychoanalytic Quarterly, 48*, pp. 292-307, 1979.
12. L. B. Ames and J. Learned, Imaginary Companions and Related Phenomena, *The Journal of Genetic Psychology, 69*, pp. 147-167, 1946.
13. D. Winnicott, *Lek och verklighet* [Playing and reality], I. Löfgren (trans.), Natur och Kultur, Stockholm, 1995. (Original work published 1971.)
14. A. Sugarman and L. S. Jaffe, A Developmental Line of Transitional Phenomenon, in *The Facilitating Environment: Clinical Applications of Winnicott's Theory*, M. G. Fromm and B. L. Smith (eds.), I. U. P., Madison, pp. 81-129, 1989.
15. B. Carlsson, *Kvalitativa Forskningsmetoder för Medicin och Beteendevetenskap* [*Qualitative Research Methods for Medicine and Behavioral Sciences*], Almquist och Wiksell förlag, Stockholm, 1991.
16. S. Kvale, *Den kvalitativa forskningsintervjun* [*The qualitative research interview*], S. E. Torhell (trans.), Studentlitteratur, Lund, 1997.

17. J. Garbarino and F. M. Stott, *What Children Can Tell Us. Eliciting, Interpreting and Evaluating Critical Information from Children,* Jossey-Bass Publishers, San Francisco, 1992.
18. N. K. Denzin and Y. S. Lincoln, Introduction: Entering the Field of Qualitative Research, in *Handbook of Qualitative Research,* N. K. Denzin and Y. S. Lincoln (eds.), Sage, Thousand Oaks, California; pp. 1-17, 1994.
19. *En målares gåta: En berättelse som utgångspunkt för att studera kreativitet hos 10-åringar [A Painter's Mystery: A Story as a Starting-Point to Study Creativity in 10-Year-Old Children*], published manuscript in 2000.
20. A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques,* Sage, Newbury Park, California, 1990.
21. A. M. Huberman and M. B. Miles, Data Management and Analysis Methods, in *Handbook of Qualitative Research,* N. K. Denzin and Y. S. Lincoln (eds.), Sage, Thousand Oaks, California, pp. 428-444, 1994.
22. H. Kohut, *The Restoration of the Self,* I. U. P., New York, 1977.
23. D. Stern, *The Interpersonal World of the Infant,* Basic Books, New York, 1985.
24. T. R. Gleason, Social Provisions of Real and Imaginary Relationships in Early Childhood, *Developmental Psychology, 38,* pp. 979-992, 2002.
25. E. T. Higgins, Continuities and Discontinuities in Self-Regulatory and Self-Evaluative Processes: A Developmental Theory Relating Self and Affect, *Journal of Personality, 57,* pp. 407-445, 1989.
26. S. E. Taylor and J. D. Brown, Illusion and Well-Being: A Social Psychological Perspective on Mental Health, in *The Self in Social Psychology,* R. F. Baumester (ed.), Psychology Press, Philadelphia, Pennsylvania, pp. 43-67, 1999.
27. J. Piaget, *Play, Dream, and Imitations in Childhood,* C. Gattegno and F. M. Hodgson (trans.), Routledge & Kegan Paul Ltd., London, 1967. (Original work published 1951.)
28. J. D. Brown, Evaluating Self and Others: Self-Enhancement Biases in Social Judgements. *Social Cognition, 4,* pp. 353-376, 1986.
29. R. F. Baumeister, K. Dale, and K. L. Sommer, Freudian Defense Mechanisms and Empirical Findings in Modern Social Psychology: Reaction Formation, Projection, Displacement, Undoing, Isolation, Sublimation, and Denial, *Journal of Personality, 66,* pp. 1081-1124, 1998.
30. P. Cramer, Defense Mechanisms in Psychology Today: Further Processes for Adaptation, *American Psychologist, 55,* pp. 637-646, 2000,
31. A. G. Greenwald, The Totalitarian Ego: Fabrication and Revision of Personal History, *American Psychologist, 35,* pp. 603-618, 1980.
32. D. L. Paulhus and D. B. Reid, Enhancement and Denial in Social Desirable Responding. *Journal of Personality and Social Psychology, 60,* pp. 307-317, 1991.
33. I. Seiffge-Krenke, Ein Sehr Spezieller Freund: Der Imaginaere Gefaehrte [A Very Special Friend: The Imaginary Companion], *Praxis der Kinderpsychologie und Kinderpsychiatrie, 49,* pp. 689-702, 2000.
34. E. J. Langer, The Illusion of Control, *Journal of Personality and Social Psychology, 32,* pp. 311-328, 1975.
35. S. Harter, The Development of Self-Representations, in *Handbook of Child Psychology: Vol. 3. Social, Emotional and Personality Development* (5th Edition), W. Damon (series ed.) and N. Eisenberg (vol. ed.), Wiley, New York, pp. 553-617, 1998.

36. D. Cohen and S. A. Mackeith, *The Development of Imagination: The Private Worlds of Childhood,* Routledge, London, 1991.
37. P. Bouldin and C. Pratt, A Systematic Assessment of the Specific Fears, Anxiety Level, and Temperament of Children with Imaginary Companions, *Australian Journal of Psychology, 54*:2, pp. 79-85, 2002.
38. S. Fraiberg, *The Magic Years,* Scribner's, New York, 1959.
39. S. Freud, Formulations Regarding Two Principles of Mental Functioning, in *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 12), J. Strachey (ed. and trans.), Hogarth Press, London, pp. 213-226, 1986. (Original work published 1911.)
40. M. Alvesson, Beyond Neo-Positivist, Romantics and Localists: A Reflexive Approach to Interviews in Organization Research, *Academy of Management Review, 28,* pp. 13-33, 2003.

Direct reprint requests to:

Eva Hoff
Department of Psychology
Lund University
Box 213
SE-221 00 Lund, Sweden
e-mail: eva.hoff@psychology.lu.se

View publication stats

Dream Images: A Call to Mental Arms

Editors: *Jayne Gackenbach and Anees A. Sheikh*

IMAGERY AND HUMAN DEVELOPMENT SERIES *Anees A. Sheikh*, Series Editor

Renewed interest into serious dream investigation in recent years has supplied a number of conceptual and research applications into dream study. *Dream Images: A Call to Mental Arms* brings these works together, in one complete, comprehensive volume. Readers will appreciate that the psychology of dreaming is alive and thriving.

Table of Contents

Introduction *Jayne Gackenbach*

Part 1. Experimental Considerations

Sources and Consequences of Dreams: Experimental Manipulations and Their Effects *Joseph DeKoninck*

An Experimental Approach to the Anomalous Dream *Stanley Krippner*

Vestibular Involvement in the Neurocognition of Lucid Dreaming *Thomas J. Snyder and Jayne Gackenbach*

Linguistic and Literary Creativity in Dreams: A Psychoanalytic and an Experimental Approach *Frank Heynick*

Part 2. Clinical Considerations

Nightmares: Facts, Fictions and Future Directions *Kathryn Belicki and Marion A. Cuddy*

Nightmares and Childhood Dreams *Patricia Garfield*

The Meaning of Dreams: A Current Warning Theory *Robert C. Smith*

Dream Work with Dissociated Patients and a Self Curative Programming Function of Dreams *Roy D. Salley*

A Selective Mood Regulatory Function for Sleep *Milton Kramer*

The Psychobiology of Mental Illness: Changes in the Physiological and Psychological Aspects of Sleep *Milton Kramer*

Part 3. Interdisciplinary Considerations

Interdisciplinary Studies of Dreams: Finding a Common Ground *Don Kuiken*

Languages of Dreaming: Anthropological Approaches to the Study of Dreaming in Other Cultures *Waud H. Kracke*

Early Hindu Speculation about Dreams: Implications for Dream Yoga *George Gillespie*

Part 4. Considerations Beyond Dreaming

Reality Dreams and Their Effects on Spiritual Belief: A Revision of Animism Theory *Tore A. Nielson*

Lucid Dreaming as a Meditative State: Some Evidence from Long-Term Meditators in Relation to the Cognitive-Psychological Bases of Transpersonal Phenomena *Harry T. Hunt*

A Developmental Model of Consciousness in Sleep: From Sleep Consciousness to Pure Consciousness *Jayne Gackenbach*

Subject Index • Contributors

6" x 9", 320 pages, Cloth, ISBN: 0-89503-074-8, $51.95 + $5.50 p/h
Paper, ISBN: 0-89503-056-X, $39.00 + $5.50 p/h

Baywood Publishing Company, Inc.
26 Austin Avenue, P.O. Box 337, Amityville, NY 11701
call (631) 691-1270 • fax (631) 691-1770 • toll-free orderline (800) 638-7819

BAYWOOD JOURNALS–2005

Benefit from the experience of experts
Order through our web site: http://baywood.com

FREE SAMPLE ISSUE UPON REQUEST!

ABSTRACTS IN ANTHROPOLOGY

A CURRENT BIBLIOGRAPHY ON AFRICAN AFFAIRS

EMPIRICAL STUDIES OF THE ARTS

HALLYM INTERNATIONAL JOURNAL OF AGING

IMAGINATION, COGNITION AND PERSONALITY

THE INTERNATIONAL JOURNAL OF AGING & HUMAN DEVELOPMENT

INTERNATIONAL JOURNAL OF HEALTH SERVICES

INTERNATIONAL JOURNAL OF PSYCHIATRY IN MEDICINE

INTERNATIONAL QUARTERLY OF COMMUNITY HEALTH EDUCATION

INTERNATIONAL JOURNAL OF SELF HELP AND SELF CARE

JOURNAL OF APPLIED FIRE SCIENCE

JOURNAL OF COLLECTIVE NEGOTIATIONS IN THE PUBLIC SECTOR

JOURNAL OF COLLEGE STUDENT RETENTION

JOURNAL OF DRUG EDUCATION

JOURNAL OF EDUCATIONAL COMPUTING RESEARCH

JOURNAL OF EDUCATIONAL TECHNOLOGY SYSTEMS

JOURNAL OF ENVIRONMENTAL SYSTEMS

JOURNAL OF INDIVIDUAL EMPLOYMENT RIGHTS

JOURNAL OF RECREATIONAL MATHEMATICS

JOURNAL OF TECHNICAL WRITING & COMMUNICATION

NEW SOLUTIONS: A JOURNAL OF ENVIRONMENTAL AND OCCUPATIONAL HEALTH POLICY

NORTH AMERICAN ARCHAEOLOGIST

OMEGA–JOURNAL OF DEATH AND DYING

BAYWOOD PUBLISHING COMPANY, INC.
26 Austin Avenue, PO Box 337, Amityville, NY 11701
call (631) 691-1270 • fax (631) 691-1770 • toll-free orderline (800) 638-7819
e-mail baywood@baywood.com • web site http://baywood.com

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/356369293

Imaginary friend play in light of enactivism

Article *in* Theory & Psychology · November 2021

DOI: 10.1177/09593543211053815

CITATIONS
2

READS
47

1 author:

Zuzanna Rucinska
University of Antwerp

34 PUBLICATIONS **287** CITATIONS

SEE PROFILE

All content following this page was uploaded by Zuzanna Rucinska on 13 September 2023.

The user has requested enhancement of the downloaded file.

Imaginary Friend Play in light of Enactivism

Abstract

Review of psychological data of how children engage in imaginary friend play (IFP) shows that it involves a lot of explicit embodied action and interaction with the surrounding people and environments. However, IFP is still conceptualized as requiring the actor to mentally represent an absent entity in imagination, a capacity deemed necessary because the imaginary companion is absent or not real. Enactivism can propose to think of IFP in a fundamentally different way, as an explicitly embodied and performative act, where one does not need to mentally represent absent entities. Enactivism reconceptualises imagination involved in IFP as enacted and intersubjective, and proposes that play environments have present affordances for social and normative interactions that are re-enacted in IFP. The paper proposes a theoretical argument for an enactive approach to IFP as performed in the world, and not simply in the head, which captures the interactive nature of this play.

Keywords

Imaginary friend play, imaginary companions, enactivism, pretense, social affordances, absence, fictions

1 Introduction

The topic of imaginary friends or imaginary companions (ICs), and related to them alter-egos, ghosts, and hallucinations, has populated popular culture. When thinking about what it is to have an imaginary companion, many think of children who personify their dolls, or turn their stuffed animals into personified companions, such as an anthropomorphic tiger Hobbes (from *Calvin and Hobbes* comic strip, 1985-1995) who is seen by others than Calvin as merely a stuffed toy ("Calvin and Hobbes", 2021), or an anthropomorphic bear Winnie-the-Pooh, created from the imagination of Christopher Robin (both characters created by A. A. Milne in 1926). Some also think of having an illusory image not based on a toy or prop, such as a six-foot tall companion Harvey the rabbit from *Harvey* (1950) that only the main character can see, or an antihero man dressed in a large rabbit suit from *Donnie Darko* (2001), to mention but a few. Imaginary companions can take the shape of both anthropomorphized dolls or animals, as well as invisible companions (see Hoff, 2004, 2005).

To capture this phenomenon, theoretical accounts typically focus on the representational capacities of the individuals who engage in imaginary friend play (henceforth: IFP) (see Taylor, 2013). According to cognitivists, to engage in imaginary friend play is to represent an absent entity or relationship in imagination (Leslie, 1987; Carlson & White, 2013). In the case of personifying objects where a prop is present, it is the friendship, or other interpersonal relationship, that is said to take place in the imagination of the actor rather than in the external physical reality. In the case of the prop not being present, IFP is said to occur in the mind of the actor, in absence of relevant stimuli (Carlson & White, 2013). Without a prop, the imaginary friend is seen as a fictional entity not present in the immediate

environment, and so, in need of being represented in imagination. Imagination, in turn, is understood as a special mental state that involves cognitive architectures with mental representations that can act as "mental stand ins" for the absent entities (Kosslyn et al. 2006). From this account it follows that to understand how one represents absent entities in IFP, cognitive architectures of individual minds and workings of mental representational processes that provide mental stand ins of absent entities need to be studied.

IFP is one aspect of the rich repertoire of child's pretend play. Pretend play can involve treating a prop as if it is another (a banana is a pretend 'phone'), endowing it with qualities it does not possess (a dry towel is now 'wet'), or, instantiating an imaginary being (see Leslie, 1987). Proposals from enactivism (AUTHOR) and ecological psychology (Szokolszky, 2006, 2019) have provided explanations of pretend behaviors with objects, focusing on the role of active manipulation of props and their affordances in achieving pretend play. These theories have yet to account for imaginary play that does not involve props. Hence, in this paper the focus will be on children's play with invisible imaginary companions, as this form of play is said to be more challenging for the enactivist to explain (Foglia & Grush, 2011).

One way to explain how it is possible to play with an invisible imaginary companion is to say that one must invoke a mental image of an absent entity (Bouldin, 2006). Having an image of an absent entity should allow for the imaginary play with that entity to be done 'solely in one's head'. However, while such 'playing in the head' is possible, it does not reflect the available data on the the actual experiences of the pretenders who have ICs, or the interactive nature of their play practices (see, e.g., Singer & Signer, 1990; Taylor, 1999; Bouldin, 2006; Papastathopoulos & Kugiumutzakis, 2007; Gleason, 2013). It also does not settle the issue of whether the mental image is a mental representation (see section 6). Findings from these studies repot that children's IFP involves a lot of interaction with actual

people and available objects in the environment, even when the play does not involve the right kinds of props. The imaginary play makes heavy use of linguistic practices like verbalizations and negotiations, and lasts as long as the pretender is in control of the play. Therefore, children's IFP seems to involve children's embodiment and interaction in the physical and social environment much more than given credit for by the theoretical accounts of imaginary play, which focus mainly on individual cognitive capacities of the pretenders.

While cognitivists would not deny that past social interactions form an important context for the IFP, they still consider mentally representing absent entities to be the defining feature of this play *and* be a necessary part of its explanation. Cognitivists explain imaginary engagements though positing mental representations, because one deals with absent or fictional entities that must therefore be 'brought forth' in the mind. Mental representations are here understood as mental stand-ins that harbour propositional or imagistic contents. (Chemero, 2009) However, there an important difference between seeing an act as representational, and saying that it must involve mental representations to explain it. It is possible for an act to be a representational act by definition, without it invoking mental representations as explanatory tools that play a causal explanatory role in making the act possible (see AUTHOR). In this spirit, this paper will show that we can consider IFP as a form of enacted performance that represents (or presents) friendship routines in action, without assuming that the explanation of such play resides in the cognitive architecture of the pretender, in the mental representational processes providing mental 'stand ins'.

Enactivism focuses on explaining IFP by looking at socio-cultural norms of playing, and not individual mental representational processes. As the environment offers possibilities for normative actions like re-enacting friendship routines, one need not mentally represent an absent entity. And while IFP can be considered to be a representational phenomenon in the

performative sense – the actor represents in her performance what it means for her to have a 'friend' – it is something else than mental representations, understood as mental stand-ins that harbour contents, that cause this play. Enactivism argues that in our playful engagements such as in pretend and in imaginary play, we continue to act on material and social affordances of the environment (Gallagher, 2017; AUTHOR). It sees social and intersubjective aspects of pretend play as not only contextual, but as constitutive of the phenomena (Varga, 2011; AUTHOR). It sees social interactions as central to IFP, and further emphasizes the interactive nature of this form of play. It also reconceptualises imagination as a form of intersubjective and embodied *doing* (Medina, 2007; Hutto, 2015; Ilundáin-Agurruza, 2017), leaving room for non-representational explanation of acts that involve mental imagery (Gallagher & Rucińska, 2021).

Thus, the aim of this paper is to better understand the phenomenon of IFP by critically looking at theoretical positions that delineate this phenomenon and compare them against the psychological data. Cognitivist accounts of IFP maintain that IFP requires mental representations of absences to be entertained in the pretender's head, in spite of the rich literature from developmental and play psychology on the social and interactional nature of IFP. The paper does not aim to argue for the limitations of the cognitivist approach, but to put an 'enactivist reconceptualization' of IFP on the table, to be treated as a viable alternative conceptual model of IFP to the cognitivist model. This is motivated by the newly available understanding of imagination from the embodied and enactive cognitive science, and the better fit between enactivist theory and the social aspects of IFP.

This paper provides a first proposal of imaginary friend play (IFP) from the perspective of enactivism. Enactivism can reconceptualise IFP as a fundamentally social practice where the players re-enact friendship routines. It argues that the environments for play have present

affordances for social interactions, which the players explore in action. Following enactivist philosophy, this paper will argue that even interactions with absent imaginary companions are embodied and enacted in the world, not just in the head of the pretender.

The paper then shows what follows from the enactivist ideas on the fictional and 'absent' nature of the IC. By analyzing how it is possible to interact with ICs (entities that are considered by some definitions to be 'not real' or 'absent'), enactivism can further shed new light on how to understand the 'fictional' nature of ICs. The paper proposes that the fictional status of the imaginary friend is negotiated in play-acting, and therefore, 'brought forth' or presented in play. This position reverses the standard assumption in psychology that one must first know and 'quarantine' in the mind what is real and what is fictional (Leslie, 1987), proposing instead that what counts as fictional is pragmatically achieved in the worldly engagements.

This paper will unfold as follows. Section 2 describes what often happens before and during IFP, following some specific findings from developmental and social psychology that focus on actions of pretenders during IFP, and by providing examples of what IFP typically looks like in practice. Section 3 discusses the main tenets of enactivist philosophy and develops the proposal for the enactivist account of IFP. It focuses on the concept of social affordances (van Dijk & Rietveld, 2018) to show that contexts in which IFP occurs offer normative engagements like re-enacting friendship routines, even when there are no friends present. Section 4 supports the enactive take to IFP by introducing the view that imaginings taking place in IFP are intersubjective (Papastathopoulos & Kugiumutzakis, 2007) and enacted (Medina, 2007; Ilundáin-Agurruza, 2017; Gallagher & Rucińska, 2021), and by showing that imaginings can be communicated and shared in an intersubjective context (Stukenbrock, 2017). Section 5 follows up with an analysis of the fictional nature of ICs and a reconstruction

of the 'absence' criterion of IFP. It argues that IFP can be *immediately* afforded, does not deal with an *absence* of a friend to begin with, and the fictional friend exists *pragmatically* in the lives of the pretender and her family, thanks to the practices of negotiating the reality of the play. Section 6 concludes the paper with a discussion of the potential role of mental representations left to play in IFP. Section 7 follows-up by discussing limitations of the given analysis, and considers the benefits of the enactive take to IFP for psychological research.

2 What happens before and during IFP

From around the age of five and six (though earlier occurrences have also been reported)[1], many children, especially in the West, play with imaginary companions (Singer & Singer, 1990; Taylor, 1999; 2013). Children engage with toys, other objects or even empty spaces (such as unoccupied chairs) 'as if' they are other human or non-human beings. An imaginary companion or friend has been defined as "a very vivid imaginary character (person, animal) with which a child interacts during his/her play and daily activities" (Taylor, 1999, p. 28), or a character "that does not actually exist, although the child treats it as though it does" (Bouldin, 2006, p. 19). An imaginary companion can sometimes be entirely invisible, and sometimes take the form of a stuffed animal or doll (Taylor, 1999; Singer & Singer, 1990; see also Hoff, 2004, 2005). The imaginary companions can vary in the length of time, and they usually, but

[1] According to Hoff (2004), there is evidence that "the earliest imaginary companions appear at the age of two or three" (p. 151).

not always, are of a friendly character.[2] The companions are more than just comforting toys or beings; they often have distinct personalities.

IFP is an instance of pretend play (Leslie, 1987), and a common elaboration of early forms of pretend play. It has a developmental trajectory: first children pretend with external objects (e.g., one-to-two year olds engage in object substitution play, such as pretending that a banana is a phone), then children learn how to pretend without props, engaging in role-playing (e.g., three-to-four year olds start playing as if they are "bears"), and finally, imaginary friend play develops (Liao & Gendler, 2011). IFP is taken to be the most developed phase of pretending, occurring when children have learned to think symbolically and to decontextualize reality from fiction, in order to have pretend conversations with "absent entities" (Somers & Yawkey, 1984; see also Singer & Singer, 1990).[3] As the phenomenon is commonly thought of, this is when children pretend "alone with their thoughts".

IFP often involves props and toys. But even without explicit props (when the IC is invisible), non-mental entities such as the place, space, and people in the lives of the pretenders are be said to both determine and constrain the IFP phenomenon. Singer & Singer (1990) argue that aside the role of props, there are essential aspects of the *immediate environment* that facilitates imaginary friend play, such as "a key *person* in a child's life who inspires and sanctions play and accepts the child's inventions with respect and delight (...), a *place* for play, a "sacred space" (no matter how small), and *time,* open-ended and unstructured" (pp. 3-4).

---

[2] It must be acknowledged that not all ICs are friendly. Some ICs are antiheroes (as the Rabbit in *Donnie Darko*), others fill the role of scapegoats, protégé's, or idealized selves (see Hoff, 2004, p. 153), can behave like adults or commit transgressions without consequences. The play forms with these ICs do not reflect friendship routines (thanks to the anonymous reviewer for this clarification).

[3] Aside the development of intellectual cognition, imaginative play is also taken to be important for the emotional development (Taylor, 1999), and development of social cognition (Gleason, 2013).

Social factors also foster IFP. Looking at the development of the various imaginative practices, we see that they first and foremost occur in a social context. The imaginary friend character is often "borrowed" from what has been seen before on television or in movies (Taylor, 1999). Advanced social understanding also affects the imaginary relationships, which change with time and the environment. For example, as Gleason (2013) points out,

> almost all relationships with imaginary companions can be categorized in real relationship terms, such as friendships, parent–child relationships, or siblings (...). Social affordances of relationships, such as companionship, are regularly ascribed to imaginary companions [and] the dynamic nature of imaginary companions might mean that such an imaginary relationship would shift and change along with characteristics of the environment and/or development of the child (pp. 256-257).

The phenomenon of IFP is also highly culture-dependent. There are "substantial cultural differences in parental reaction to, and interpretation of, childhood fantasies that are likely to affect children's engagement in pretense, as well as their understanding of the distinction between fantasy and reality" (Taylor, 1999, p. 6.). For instance, Farver & Howes' (1993) cross-cultural study showed that the lower numbers of IFP of children in Mexico, relative to IFP of children in the USA, is likely due to different cultural acceptances of what is real and not real and what is culturally acceptable as play, not so much the imaginative skills of individuals. Their interpretation for finding cross-cultural differences was that American parents supported imaginary play of their children, justifying it as "healthy" (up until a certain age), while Mexican parents did not, considering it as "lying". Such difference in parental attitudes could be explained by the different cultural attitudes towards what counts as "real".

There are also explicit activities that children usually engage in during IFP. Some of those activities typical to IFP include explicit verbalization, negotiation of how the play unfolds, and keeping control in the play.[4]

Firstly, IFP is strongly linked to explicit verbalizations. Children with imaginary companions tend to "provide more detailed descriptions of their fantasies than did children without companions" (Bouldin, 2006, p. 25) and "tell more novel stories in responses to a set of verbal stimuli than children who are less interested in fantasy play" (Taylor, 1999, pp. 40-41). On average, girls were described by their parents as talking out loud to the make-believe friends at one time or another. Their ICs were more likely to be human (Singer & Singer, 1990; Hoff, 2005), and their IFP more cooperative and social (Papastathopoulos and Kugiumutzakis, 2007), than those of boys, keeping in with the generally greater verbal development of girls.[5]

Secondly, IFP involves negotiations. Papastathopoulos and Kugiumutzakis (2007) showed that children (mainly girls) with imaginary companions engaged twice as much in negotiations of pretense than other children. The negotiations aimed at the "organization and establishment of a joint and shared body of knowledge, meanings, rules and procedures for the collaborative maintenance and experience of an imaginary reality" (*ibid.,* p. 228). Moreover, imaginary companions (unlike, for instance, hallucinated imaginings) were not confused with reality; pretenders who had imaginary friends often made it very clear that they are aware of the pretense, and that the rule of this game is that the imaginary companion does not exist (Taylor, 1999, 2008). Children have been reported to help the interviewer by

[4] The following analysis is not complete, as there exist a myriad of types of IFP, including play that is not shared with others, or interactions with ICs that might be best described as occurring 'inside the minds of the pretender' (see Hoff, 2004). See section 7 for further discussion.

[5] This finding may be further connected to the practice of socializing, novel-reading and explicit verbal storytelling of school-age children, as the school environment fosters narrative practices as well as opportunities for forming new friendships.

saying, "It's just pretend, you know" or "She isn't real" (Taylor, 1999, p. 112).[6] It could be said that the child attempted to define and negotiate what happens in the IFP, including the reality of the imaginary companion, with the one who would listen.

Finally, what distinguishes IFP from having hallucinations or daydreams is the aspect of control of how the play unfolds. Children who engage in IFP are in total control of the IC, bossing them around, directing their activities or dictating how the others can communicate with their IC, resisting the adult's attempts to dictate the actions of imaginary companions. Taylor (1999) found that in IFP, “the child is the official commentator on the imaginary companion's activities and the translator of what it has to say” (p. 47) and that "if children's sense of control over the imaginary companion is diminished, the pretend friend sometimes disappears” (pp. 124-125). Children are thus in control of not only how the imaginary companion behaves, but also how the adult participates in the play. Similarly, Singer & Singer (1990) found that "when adults "butt in" uninvited, children not only lose a sense of privacy and possession but may also fear that adults are ridiculing them or have somehow stepped out of adult roles" (p. 110). Other, more recent work (Gleason, 2013) reports that control and authority over the play is a crucial aspect of the play itself, as "preschool age children rarely enjoy more knowledge and expertise than their parents" (p. 257). Keeping control of how the IFP unfolds and how others engage with the imaginary companion, including whether the person fostering pretend play accepts the game or not, is crucial part of the play itself.

[6] As Taylor reports, children "readily answer the question 'Do you have a *pretend* friend?' by describing imaginary companions that are substantiated by the reports of their parents. They know what we are talking about. In fact, we have the distinct impression that after children spend a period of time answering detailed questions about a pretend friend for a researcher who listens carefully and even takes notes, they begin to wonder if the adult might be confused" (1999, p. 112).

Consider the following illustration of IFP. According to Taylor (1999), some parents have reported not being able to sit down at a restaurant if the imaginary friend of their child did not have a seat. As Taylor reports: "One child who participated in my research expected the whole family to wait when they went out to a restaurant until a table large enough to accommodate her imaginary companions, as well as the rest of the family, was available" (*ibid.,* p. 87). Let's imagine that the child's name is Amy and that she saves a seat in a restaurant for her imaginary friend Max. What happens during such play? Following the findings mentioned above, it is likely that Amy controls the play by telling her parents "you can't sit here, Max is sitting here" or "this seat is reserved for Max". She is likely to speak out loud to Max, when facing the empty chair, indicating thereby that the chair is occupied by Max (explicit verbalization). Amy also does not allow the family to get a smaller table that does not have the extra chair for Max (negotiation of the play), objects when someone wants to sit on that specific empty chair and objects when others try to interact with Max directly (keeping control of the game). Even though there is no explicit prop that 'stands in' for Max, Amy utilizes her immediate environment (empty chair in the restaurant) and includes her family in the enactment of Max IFP.

The role of the immediate environment, linguistic verbalizations and social engagements in IFP reveal that this phenomenon not only develops in response to a social context, but importantly, could be seen as staying social. Not only is IFP vastly concerned with past social interactions, but it involves ongoing social interactions, and the imaginary companions can themselves be thought of as having a social existence as "they afford a malleable and dynamic forum for practicing, exploring, and managing relationship issues (Gleason, 2013, p. 252). These findings resonate with the enactivist proposal that cognition is

social and interactive. The next section will elaborate on the relevant tenets of this theoretical approach.

3 Enactivist account of pretend and imaginary play

What is *enactivism?* Enactivism is a philosophical framework that stresses the constitutive role of the environment and interactions of the agent with the environment in shaping cognition. While there are different variants of enactivism in the literature,[7] they share the commitment to understanding cognition through dynamical and reciprocal interactions between the agent and its physical, social and cultural environment. This holds for pretense engagements as well. For example, the enactive account of pretend play (AUTHOR) proposes that engagement in pretense is highly expressive: rather than being 'inside one's head', it mainly involves manipulations of objects in the world and exploring new possibilities for action in engagements with others. The ability to see playful affordances *in* objects, combined with embodied, sensorimotor actions, is sufficient for some types of pretend play like object-substitution play. Enactivist account of pretend play negates the distinction between offline and online imaginative capacities, cutting across this dichotomy, and provides support for perceptual-imaginative pretense capacities (AUTHOR). The outcome of this approach is a *performative* notion of pretense, where pretend scenarios are *presented*, performed, or enacted in the world, not represented in detachment from the world.

[7] For example, enactivism varies between autopoietic, (Varela, Thompson & Rosch, 1991), autonomous (Di Paolo et al., 2010), sensorimotor (O'Regan and Noë, 2001) and radical (Hutto & Myin 2013, 2017).

Enactivism as a philosophical framework reframes cognition, including imagination, into action and interaction (see also section 4). Its main unit of analysis is the dynamic agent-environment coupling, not the sub-personal cognitive architectures of individual agents (Gallagher, 2017). This is relevant to how to think of IFP as well. Enactivists will conceive of IFP as not happening in the head first, only to be later projected onto behavior, but as happening in actual playing in the world with the others in it. Hence, the enactivist take to IFP, in contrast to the cognitivist account, will argue for the direct role of the actions and interactions in this play.

Following enactivism, we can propose a new definition of Imaginary Friend Play: IFP is a performative act of representing friendship scenarios in action, where the representation is achieved through action (verbalization, negotiation, establishing control dynamics).[8] The representation of the imaginary entity in IFP is better conceptualized as an act of *presenting* a friend*, or* making the friend *present in action* and *in experience*, where in the presenting, the child creates and negotiates the reality of the imaginary companion, in the world, and with others.

One might contest this proposal, and argue that there is no immediate environment that affords friendship routines present in IFP, such as in the context of Amy's play. A restaurant with a chair might be seen as not sufficiently inviting specific re-enactments of friendship routines, which is why a mental representation of an imagined friend must still be brought forth. However, enactivism can make use of resources from ecological psychology and pragmatism to argue that human environments within which we play directly afford

[8] Not all imaginary play re-enacts friendship scenarios, hence, this new definition is limited to imaginary *friend* play specifically. However, an enactive account of imaginary play could build on this definition and propose a more general definition that would include imaginary play with non-friendly companions as well.

various socio-normative engagements. As a *philosophy of nature* (Gallagher, 2017), enactivism brings together different theoretical frameworks and can introduce its own explanatory concepts.[9] To make sense of the idea that IFP is situated in and structured by the immediate environment, enactivism can make use of the concepts of *embeddedness* or *participatory sense-making*.[10] In this paper, I will, however, focus on the ecological concept of *social affordances*.

Social affordances are affordances structured by sociocultural environments (Rietveld & Kiverstein, 2014). They match a conception of affordances understood as relations between the ecological niches and the whole animal species (which, for humans, extends to cultural environments), not just between individual animals and their environments. The rich sociomaterial notion of affordances proposes that affordances are possibilities available in a form of life, tied to practices and patterns of behavior, rather than to individual abilities (van Dijk & Rietveld, 2018).[11] This notion of affordances can be further grounded in Vygostkian understanding of cognitive development as situated in socio-cultural environment. According to Valenti & Gold (1991, p. 93),

[9] As explained by Godfrey-Smith (2001), philosophy of nature "can use its own categories and concepts, concepts developed for the task of describing the world as accurately as possible when a range of scientific descriptions are to be taken into account, and when a philosophical concern with the underlying structure of theories is appropriate" (p. 284).

[10] Embeddedness refers to the idea that cognitive acts, including IFP, are constrained by the existing environmental materials as well as social interactions and situations. Sense-making refers to the organism's or agent's active creation and appreciation (or *enaction*) of meaning when engaging with the world, which makes the world inherently significant and normative to that agent (Di Paolo, Rohde, & De Jaegher, 2011).

[11] Van Dijk and Rietveld (2018) present a framework in which situations and their affordances "unfold" and "can be thought of as continuing a history of practices into a current situational activity". They call this a "process-based account of affordances", which makes affordances an *open-ended* concept, as affordances become open to include new possibilities for actions as practices change.

> In contemporary Vygotskian approaches to understanding cognitive development (...),"many objects and settings in the human environment are outright products of culture, constructed by human beings for some meaningful purpose"(...). The child acts within these socioculturally structured environments, transforming these contexts and in doing so, continually creating new opportunities for action (i.e., new affordances) (...). Cultural practices, therefore, hide some affordances, reveal others, and allow the child to participate in the creation of still others, providing a range of directions for the development of behavioral competence.

The cultural practices of friendship scenarios will reveal what affordances are available in play. Some of those affordances can be pertaining to friendship practices, such as 'saving a seat for a friend when you're in a restaurant'. The concept of social affordances supports a position that friendship pretend play can be made immediately available even in unobvious places like a restaurant, because of the restaurant's social affordances to re-enact the seat-saving routine.

It might be thought that the environment's materiality constraints, rather than opens, the possibilities for action. For example, a cup of coffee has an affordance of 'pickupability' due to the cup's material properties like its shape and size, relative to a human hand, and coffee has an affordance of 'drinkability' due to its liquidity relative to the human digestive system. However, a cup of coffee that belongs to another person does not have either of these affordances. As social norms dictate that we do not drink from other people's cups, another person's cup of coffee will not invite picking up or drinking from. Van Dijk and Rietveld (2018) point to the socio-cultural context in which our practices have developed, which

determine how we interact with the material world, and propose that even material constrains do not pre-exist our practices.[12]

Hence, social affordances, or affordances of the environment that pertain to social conducts, are also at play during IFP. When playing the game of Max who "sits" on the empty chair in a restaurant, Amy re-enacts the social practice of saving a seat for a friend. She acts on one of the social affordances of chairs: objects that can be saved for friends. The empty chair affords a "saving" routine to Amy.[13] Amy experiences that particular chair in the restaurant as "not-sit-able" for others, just as we experience other people's cups of coffee as "not-drinkable-from" in everyday contexts. To an observer of Amy's play, the empty chair in a restaurant might still afford sitting-on, but to those who participate in Amy's play, the chair will not afford sitting-on by others as well. The interaction between Amy and her family, where the chair's "sitability" is negotiated, constructs a situation where the imagining of Max can be shared (see section 4).

To summarize, the enactivist view of IFP sees this form of pretense as a *presentation*: the child, in her behaviour, *presents* to oneself and the others around her what it is to have a friend. Instead of mentally representing the absent friend, she is presenting one of the normative possibilities for friendly action as experienced in a given situation. The acting out of the "saving a seat" norm does not necessitate forming a mental representation of Max on

[12] Van Dijk and Rietveld (ft. 7) give an example of the materiality of the table. Knocking on the tabletop makes it seem that the hardness of the material is an undeniable physical reality of the tabletop, solely determined by its materiality and not our practices. However, they point that this way of determining its materiality - knocking on it - is precisely a practice, which influences how we think of tables. In different practices (ones of carpenters or scientists engaging in a chemical analysis), the same table can be used differently (e.g., sawing into two, burning), and thereby thought of differently (as relatively soft materials or specific constellation of isotopes), as different aspects of the table's materiality come forth.

[13] Why this particular affordance, and not another one, is solicited and acted on by one child and not another further stems from emotional and social needs of the child in play. That is a matter of individual differences between children. To read more about how affordances can solicit pretend play, see (Author).

the chair, or a counterfactual thought with content "if Max were real, he could be really sitting on that chair, and the chair would be occupied."[14] These social norms are themselves constitutive of the socio-material environment,[15] and afforded in that environment for play purposes. Acting on one social affordance of the chair instead of another (its "saveability" and not "sitability") further reinforces the commitment of the child and her co-players to the imaginary game.

Therefore, the enactivist take to IFP impacts how we should think of imaginary companions: from mental representations of absent entities, to actively created 'fictions' with which we can jointly interact. The following section 4 furthers the enactivist take to ICs by clarifying how enactivists think of imagination. Then, section 5 explains how it is that ICs are jointly created fictions that do not require mentally representing absent entities.

4 Enactive and intersubjective imagination

This section aims to show that an account of *intersubjective* and *enactive imagination* can capture the social nature of IFP discussed above. I will first refer to a classic, cognitivist account of imagination and its role in IFP, and contrasts it with intersubjective imagination

[14] While it is possible that enacting the friend routine is also compatible with mentally representing an absent friend, radical enactivism argues that positing explanatory notions as mental representations is no longer necessary once we can explain them away; they are also costly theoretical posits, due to 'hard problem of content' (see Hutto & Myin, 2013).

[15] Factors play a constitutive role when they are present in the same time frame as the phenomenon and form the process that produce the phenomenon (De Jaegher et al., 2010). While not many would disagree that to be able to play with an imaginary friend this way, one has to have experienced friendships and seat-saving practices in the past, enactivism can defend the idea that social factors involved in IFP play not only a historical or contextual role, but a stronger, *constitutive* role (Author).

(Papastathopoulos & Kugiumutzakis, 2007; Stuckenbrock, 2017), and enactive imagination (Medina, 2013; Ilundáin-Agurruza, 2017), which will be explored in more detail.

To recap, typical way to account for imaginary play is to study the player's individual capacity for representational thought and study the cognitive mechanisms underlying imaginative acts. Cognitivist theories propose that mental representations are needed to make sense of playing with imaginary entities. They take for granted that we need to create mental representations in imagination to stand in for those absent entities, as mental representations are stand-ins with contents that specify what the world is like in its absence. Carlson & White (2013) refer to imagination as "cognitive representations of objects, feelings, or events that are not perceptually available in the here and now. Components of imagination include symbolic thought, pretend play, and imaginary companions" (p. 162).

Imagination is also associated with the ability to produce, maintain and manipulate mental images, and according to a cognitivist analysis of visual mental imagery, it occurs "when a representation of the type created during the initial phase of perception is present, but the stimulus is not actually being perceived" (Kosslyn et al., 2006, p. 4). On this view, to engage in IFP, such as playing that Max is sitting on a chair, or to forbid others from sitting on an empty chair, Amy presumably must entertain a mental representation with a content "imaginary friend" or "Max", or invoke a mental image of "Max sitting on the chair" first.

However, according to Papastathopoulos and Kugiumutzakis (2007), this classical notion of imagination takes it to be an individual, inner and private faculty, locked in the head, which "imprisons [it] in the subjectivity of the individual, rendering it fundamentally asocial and non-intersubjective" (p. 219).[16] Instead, they propose an account of *intersubjective*

[16] As they elaborate, "In the history of psychology, imagination has been considered mainly as a private psychological state, a compensatory way of fulfilling individual wishes, an escape from reality, 'a thing in the

imagination to make sense of IFP. Intersubjective imagination is the view that imagination is formed through intersubjective and dialogical experiences, and has a fundamentally dialogical and intersubjective *nature*:

> [I]magination is not a secluded "castle" in the head, that function apart from reality and especially intersubjective reality (...). Imagination is dialogically and intersubjectively constituted and functioning. [It] springs from the same soil as with engagements with the real world: that is intersubjectivity (*ibid.*, pp. 229-230).

Papastathopoulos and Kugiumutzakis continue to conceive of imagination as mental imagery, or an ability to produce, maintain and manipulate mental images, but they link this ability to perception, understood in Gibsonian terms as "active picking up of meaning afforded in the environment [through] one's corporeal action in the world" (p. 221). They argue that bodily exploratory activity contributes to formation of imaginings, which have image-schematic form: we "move" and process meanings in imagination, much in the same way we move and process meanings in the real world. These authors also recognize that intersubjectivity plays a constitutive and not just a contextual role for pretense. They propose that imagination has a dialogical structure ("Its creation is guided by the same innate motives that guide and support the communication and interaction between persons" (p. 225)) and that dialogicity might be even be inherent in the mind, as shown by the developmental

mind' (...) For Piaget (1951) imagination is equated with the subjective assimilation of reality to the ego, is purely egocentric and serves only the satisfaction of the individual's ego. In recent developmental theories, imagination is depicted as a form of mental representations that 'do not have as a purpose to represent reality accurately' (Woolley, 1995)" (p. 219).

findings of infants having expectations about how an interacting partner should respond (Reddy, 1991).

This influences how to think of ICs. As Papastathopoulos and Kugiumutzakis (2007) write,

> We hypothesize that the creation of an imaginary companion is an expected and direct outcome of the intersubjective nature of development and the dialogical structure of imagination. (...) The imaginary companion is a person and has the functional equivalence in imagination of a real person. It has psychological and physical attributes, and is a being with which the child can form various interpersonal relations (p. 225).

This position resonates with Gleason's (2013) findings that children form enduring bonds with their imaginary companions, who can be categorized in real relationship terms.

Imaginary friends can also be shared. Papastathopoulos and Kugiumutzakis acknowledge that an implication of their position is that imagination can be communicated and shared in an intersubjective context. [17] The idea that imaginings can be shared has also been developed by Stukenbrock (2017), who introduced the concept of *intercoporeal imagining* in self-defense training. In this context, the trainer and the trainees were observed to engage in jointly imagining bodies (the attackers) and aligning their own bodies with the imaginings to practice self defense routines. According to Stukenbrock,

[17] The limits of Papastathopoulos and Kugiumutzakis' proposal is that they understand imagination happening in detachment from exploration of the environment: "The subject produces certain predictions and expectations about how something might be, while refraining from searching the “real world” to perceive" (2007, p. 221). This does not reflect the findings that when children engage in IFP, they respond to and explore their environments, both in active perception of the immediate environment and by moving around in it.

> the participants can be observed to verbally and bodily interact with their constructs as if they were there, coordinating to and with them (...) and thereby constituting an intercorporeal relationship with those phantom phenomena. This can be done because we have an inherent, sedimented, embodied 'feel for' those entities from innumerable interactions with them in our lives (2017, p. 241).

The concept of intersubjective imagination meshes with enactivist take to imagination. The concept *enactive imagination* (Medina, 2013) and *corporeal imaginings* (Ilundáin-Agurruza, 2017) can further capture the active, embedded and embodied aspect of IFP.

Enactive imagination is the view that imagining is a kind of a *doing* that involves acting in the world, not a process that happens in the brain. "[I]magining is not something that happens to us, it is something we do, something that requires active participation, a form of action and interaction" (Medina, 2007, p. 319). Imagining is integrated with perception and action in an ongoing dynamical pattern; it is a construction defined by the movements of the body that is informed by perception (Gallagher, 2017), and is thereby not reducible to simply having an image in one's head. Medina conceives of imagination as a mastery of a technique of enactment and re-enactment of experience, which remains to be an embodied and interactive exploration of the world. He gives an example of imagining a room in visual imagination, which he characterizes as "enacting or re-enacting seeing that room, moving our eyes around it and checking things out, and not simply seeing picture-like mental image of the room with out mind's eye, or having a mental model of it" (2007, p. 319). Just as visual imagination involves re-enactment of seeing, "other forms of imagining are likely to involve other forms of enactments: for example, enacting touching or being touched, enacting speaking or listening, or enacting specific movements such as grasping, holding or releasing

through the activation of the sensorimotor patterns involved in the preparation of those actions" (*ibid.,* p. 320). Following Medina's proposal, we can conceive of IFP not just as involving reenactment of seeing (see the enactive approach to motor imagery as proposed by Gallagher & Rucińska, 2021), but also of speaking (talking to the imaginary companion) and moving, in a way that reflects social practices of friendship, such as saving a seat. This active aspect of imagining is further explicated by Ilundáin-Agurruza's (2017) notion of muscular or *corporeal* imaginings understood as "fully embodied, animate processes" (p. 97) that are constrained by the environment-embodiment relation. They are re-enactments of perceptual activities based on both past and present material and bodily engagements.

Moreover, imaginings understood as perceptual and embodied re-enactments "are wholly interactive and non-representational in character" (Hutto & Myin, 2017, p. xxi). They are non-representational in the sense that "there are no conditions of truth, functionality or veridicality that need obtain" (Ilundain-Araguzza, 2017, pp. 93-94). Mental imagery has been recently shown to be understandable as a motoric and bodily process, rather than a representational state, in line with neuroscientific descriptions of imagery (Gallagher & Rucińska, 2021). These studies show that studying cognitive architectures that harbour mental representations is not the only way to approach imagination. Hence, enactivists need not deny that IFP involves imagination, but they can replace the cognitivist notion of imagination with intersubjective and enactive one, which foregrounds the social nature of imaginary play.

5 Re-conceptualizing fiction and reformulating absence

The critics of enactivism say that non-representational positions like enactivism are said to have trouble accounting for imaginative phenomena (Spaulding, 2010; Foglia & Grush, 2011), or forms of cognition that deal with less immediate, wholly absent or non-existent items (Clowes & Mendoça 2016).[18] Enactivism is said to be unable to deal with phenomena that do not allow direct interaction (the *non-immediacy condition*), are not there (the *absence condition*) or are fictional to begin with (the *non-existence condition*).[19] IFP seems to apply to all of these conditions. The friendship routine seems to not be immediately given in experience in a restaurant setting with an empty chair; Max the imaginary friend seems to be absent relative to the chair's presence; and Max's non-existence is part of the definition of IFP that take imaginary companions to be fictional beings (Taylor, 1999; Bouldin, 2006). This suggests that the enactivist approach to IFP is a non-starter: enactivism, a theory whose main quality is accounting for mentality via direct engagement with the environment, is unlikely to account for a phenomenon where the interactions are taking place "in the actor's head".

By now, however, I have hoped to show that these worries are ill-conceived. The previous sections have argued that it is incorrect to assume that interactions with the imaginary friends are taking place "inside the actor's head", and showed that the enactive account of IFP is possible, when the focus is on the *doing* involved in IFP, and when imagination is seen as achieved in action and intersubjective engagements. This responds to

[18] As Hutto & Myin (2017) notice, "many still incline to the view that [Radical Enactivism] has inherent limitations and that some version of [Cognitivism] is needed to understand cases of cognition in which what is thought about is not present—as in paradigmatic instances of dreaming, imagining, planning, and deliberating. Making precisely this assessment, Clowes and Mendonça (2016) (...) hold that [Radical Enactive] approaches are well suited for understanding those forms of cognition in which what is thought about is immediately accessible, whereas [Cognitivism] is needed for understanding forms of cognition that require dealing with items of thought that are less immediate, wholly absent, or nonexistent" (pp. 177-178).

[19] Being "less immediate", "wholly absent" and "nonexistent" can be seen as referring to different degrees of absence, or *non-present-ness*, of an item, rather than its different metaphysical statuses. Thanks to Luke Roelofs for this insight.

the worries that enactivism is a non-starter. What follows is that the enactivist take to IFP promotes a new way to think of the fictional nature of the IC, as it proposes a different approach to seeing them as *absent*.

Firstly, the concept of social affordances showed that IFP can be *immediately* afforded. IFP was presented as a form of engagement with the immediate environment, in which the player is re-enacting friendships routines and acts in accordance with shared norms of friendship, all part of the socio-material environment in which the playing takes place. In agreement with ecological psychology, enactivists can agree that the environment is rich with relevant possibilities for action, including possibilities for friendship re-enactments. This means that the possibility for re-enacting friendship routines is part of the immediate environment's affordances. What makes the child act on that affordance and not another requires further analysis (such as consideration of the emotional needs of the child), but there is no need for additional structures like mental representations to 'fill in a gap' in the environment.

Secondly, IFP is not a situation where there is an *absence* of a friend to be dealt with in the first place. The notion of absence seems to refer here to absence of relevant stimuli, like a real friend. But it is a misconception to think that when playing with an imaginary friend, one is dealing with an absent *real* friend. There was no real friend to begin with. To take an example from object-substitution pretense, in playing with a toy block as if it is a car, there is no absence of a real car either. It is in the manipulation of the object at hand that new possibilities for action with the object are explored and acted out, thereby making it "car play". It is not a mental representation with a content "car", but the entire activity of the actor with the toy block, that presents a "car". Similarly, in IFP, there is presenting of the imaginary friend in action and in speech, to the others as much as to oneself. IFP takes place in the

talking out loud to the imaginary friend, in the pointing at the space the imaginary friend occupies, in correcting the others not to sit on the chair where the imaginary friend is sitting, etc. It is this overt action that makes the imaginary friend *present* in experience, without dealing with absence of real friends.

Finally, the imaginary friend *exists* in a pragmatic sense of the term, in the lives of the pretender and those who participate in the play, with whom the reality of the imaginary friend is negotiated. For example, when Amy says the seat is taken because Max is sitting there, she is negotiating this new reality with others. Will her family follow her proposal or resist? Following it gives Max a pragmatic role in their lives: the family has to look for a table with an extra chair for Max. We can compare the reality of the imaginary companion with the reality of money. Some fictions we do not negotiate, like treating some kinds of paper as money, even though "money" could also be seen as a fiction. The difference between "money" and "Max" is that the former is part of an institutionalized practice, while the latter is negotiated on the spot. The child is an active participant to social reality, where she learns that money isn't just paper but something with 'real' value. Hence, the child may be negotiating that if money is real, then her imaginary friend Max is also (temporarily and contextually) real.[20]

Pretend play is not the only context in which we negotiate what is pretend and what is real. There are many entities people of various cultures believe to be real that others treat as fictions. There is no simple-minded distinction between pretend and non-pretend; there are plenty of fictions in life, such as money, nations, institutions, stories, avatars, or deities, whose metaphysics leave a lot to desire. It's often what is agreed upon that determines what

[20] Wanting to exert control, the child engages in a negotiation along the lines of: if I follow you in your reality, will you follow me in mine? Thanks to Thomas Fondelli (private correspondence) for this insight.

we call real and fictional. This follows the enactivist and pragmatist thought that we enact our realities. Just like there are political structures that bring forth existence of money into our lives, or religious structures that bring forth deities and sacral objects into our lives (churches, holy books), there are non-institutionalized practices like pretend playing that bring forth imaginary friends into our lives. As Geurts (2017) proposes, "fiction is a social practice, a form of interaction between speakers and hearers, or writers and their readers" (p. 56), and fictional entities, including ICs are occupants of that social space.

Doesn't the fact that children are in control of the play suggest that they *know* what is real and not real? Even though pretenders can show that they *know* that their imaginary companions are not real by making appropriate assertions (see section 2), this speaks more to their capacity to act in accordance with social expectations of what we count as real, rather than to their private beliefs or experiences about the reality of the companion. By making assertions that the imaginary friend is "not real", children show that they are enculturated into the society. They act in accordance with social expectations, which is that imaginary friends are fictional. This knowledge is shaped by the ongoing interactions. For example, Singer & Singer (1990) report that children's sense of reality of the imaginary companion and their awareness that it is "just play" is often shaped by the responsiveness of the adult: "Children probably hover between a sense of the reality of these playmates and an awareness that they are part of their pretend play" (p. 110). Even though imaginary friends can feel phenomenally present to children, and even be experienced as *autonomous* (Taylor et al. 2007; Gleason 2013), or experienced in a phenomenally immediate way (just as fiction authors have conversations with phenomenally present characters of their stories), children are still able to understand that these friends are to be taken as fictional.

To summarize, enactivism reconceives the idea that "absence of a friend" is involved in IFP. It shows that in IFP we make, bring about, or enact the imaginary friend in virtue of acting as if it existed. The imaginary friend is enacted into a pragmatic reality and occupies a real space in the interaction. It may be even proposed that the imaginary friend is never necessarily a private entity, as others can interact with it too, if only its creator lets them. Therefore, the challenge of non-immediacy, absence or non-existence of the imaginary friend is moot: the friendship routines can be immediately afforded, and once enacted, the fictional entities are "real fictions" in a pragmatic sense of the term: they make a difference in the life of the pretender and her family. As there are no "absent friends" to begin with, we need not mentally represent them "as present" in our minds.

6 Conclusion: Towards IFP without mental representations

To summarize, enactivism is well suited to account for the fact that children's IFP involves interactive behaviors, like talking to the imaginary companions aloud, giving detailed descriptions of their fantasies, controlling how others participate in IFP, or resisting the attempts of parents to dictate the actions of their imaginary friends. This paper has shown that in IFP, the child typically responds to immediate environmental factors, such as presence of another person (parent), place where the activity takes place (restaurant), time of the activity (dinnertime) and props to be found (chair). Enactivism acknowledges that actual people, objects or places play a crucial role in triggering, structuring and shaping IFP, and proposes a reconceptualization of IFP that does not continue to seeing it as "done in the head". Enactivism can propose to think of IFP as a performative act of presenting friendship

scenarios in action (verbalization, negotiation, establishing control dynamics). In IFP, actors re-enact friendship routines, guided by their enactive and intersubjective imaginings (understood as forms of *doing,* which can also be shared), and constrained by social affordances of the environment that are themselves normative in a given culture. The paper has thereby reversed the standard cognitivist position, by proposing that we need not conceive of the imaginary friend as an absent entity in need of mentally representing in one's mind's eye.

This opens the door for enactivists to further the account of IFP, possibly without positing mental representations *altogether*.[21] Without the need to mentally represent absent entities, what role are mental representations left to play in accounting for IFP? While mental representations might be seen to be needed for various processes connected to IFP,[22] I will focus on the assumption that mental representations are needed for storage, abstraction and retrieval of friendship experiences, as this assumption was brought up by Papastathopoulos and Kugiumutzakis (2007) themselves. According to these authors, the intersubjective experiences and communication achievements formed together by the pretenders "form and organize generalized representations or internal working models that contain expectancies about the way that another person is expected to act" (p. 228). They follow Stern's (1985) proposal that cognitive structures such as RIGs (*Representations of Interactions that have been Generalized*) are needed in IFP. RIGs are mental representations that combine slightly

[21] The motivation for doing without mental representations altogether follows from positions that question the roles and nature of mental representations in cognitive science, such as Ramsey's (2007) "job-description challenge" or Hutto & Myin's (2013) "hard problem of content" challenge.

[22] Those processes can include processes of quarantining reality from fiction, or specifying how a play can be with *this imaginary friend* (Max) and not *another one* (Jill). How the enactivist can respond to these objections is beyond the scope of this paper, but to be explored in future research that studies the background knowledge and linguistic referential practices in IFP, respectively.

different series of events and transform them into “event-ides”, which function as prototypes of concepts, and lays the ground for later conceptual categorization of events (e.g., a specific instance of a peek-a-boo game can be later recognized as the game “peek-a-boo”). Hence, even if one agrees with enactivists that IFP involves re-enactments of friendships, it seems like the child still needs to represent the concept "*friendship*" somehow, and RIGs are one way of making sense of how the individual child's friendship experiences are "stored" and made into concepts, to be brought back in play.

This is an interesting challenge for enactivists that, in my view, requires careful analysis of what it means to "internalize" experiences, or "store them" as concepts. One alternative to invoking processes like RIGs is to better understand Vygotsky's (1978) proposal of how knowledge is internalized. It is likely that Vygotskian internalization is not a process by which public representations are transferred in the head, but it is a way of enculturation.[23] Leontiev (Vygotsky's student) proposed that “the process of internalization is not the transferal of an external activity to a preexisting, internal 'plane of consciousness': it is the process in which the internal plane is formed (Leontiev, 1981, p. 163; see also Podolskiy, 2012). Understood as a process of an aquisition of a capacity or technique, what follows is that there is no *thing* in the head that gets internalized, like the concept of RIG suggests. Similarly, Noë's (2015) analysis of concepts as skills of access proposes a notion of concept that does not "store" representational contents. Also, the enactivist account of remembering seen as the sensorimotor activation of familiar patterns does not require manipulation of informational content (Peeters & Segundo-Ortin, 2019), just as the enactivist account of mental imagery does not require finding a 'picture' in one's consciousness (Gallagher &

[23] Thanks to Karim Zahidi for the references to Vygotsky's and Leontiev's work on internalization.

Rucińska, 2021). How such enactivist and sensorimotor accounts of concept formation or remembering can be applied to IFP is a topic of future research, needed for better understanding of the full scope of the enactive account of IFP.

However, whether we need mental representations to entertain concepts like "friendship", or remember friendship routines for their re-enactment, is a different question than whether we need mental representations of *absent entities* to account for IFP. Questioning this latter assumption was one of the targets of this paper, and if the debate shifts to other functions that mental representations may play in IFP, then this paper has already reached its goal. Importantly for the enactivist position, even complex forms of imaginary activities like IFP need not invoke mental representations of absent entities. These activities could remain to be characterized as *representational* phenomena, but their representational status, according to enactivism, is not to be derived from mental representations. It is something else than mental representations that makes IFP an act of presenting the experiences in action.

7 Follow-up: Limitations and Benefits

Finally, it is useful to mention the potential limitations as well as added values of the enactive account of IFP to psychological literature.

Firstly, this paper acknowledges that ICs provide all manner of relationships, not just friendships; thus, the scope of the analysis of this paper was limited. The paper presented a single illustration of a form of IFP and did not engage in the discussion of other instances of imaginary play, such as play with non-friendly ICs, or other forms of imaginary engagements

that are perhaps not playful at all. This was done for the sole purpose of presenting the novel theoretical claims in an extensive and clear way. However, future research should expand on different forms of engagements with ICs and contrast their enactivist and cognitivist explanations.

There are also instances of IFP in which the IC is merely described to a family member. In these instances, no enactment is taking place beyond the child relaying the IC's behavior descriptively. Also, IFP can be hidden from a family member. Hoff (2004, 2005) documents numerous ICs of ten-year-olds that seem to exist primarily within their creators minds and are not shared with others.[24] This is another limitation of the presented view, which is that it has focused mainly on the psychological data that showed explicit engagements of pretenders who play with ICs, and did not discuss instances of children's IFP that are unobservable to the researchers or parents. These are interesting situations that require further analysis. Enactivist perspective can accept that it is possible that one can play out and imaginary engagement solely 'in the head' of the pretender, without explicit action. However, the enactivist perspective will contest that such instances must require re-instantiating mental representations. As explained in section 6, we need to better understand how play is habitualized and internalized. Also, we need to understand better how imagining works. Some forms of imagining need not require explicit bodily action, but are embodied nonetheless, as they activate neural substrates of movement (Gallagher & Rucińska, 2021). In short, explaining unobservable IFP does not automatically call for invoking mental representations. This is to be discussed in future research.

The potential benefits of the enactivist conceptualization of IFP to psychological theory is that aside its contribution to clarifying the concepts for the psychological literature

[24] Thanks to the anonymous reviewer for highlighting this point.

on imagination and children's play, it allows for the broadening of the scope of the empirical research on IFP. Enactivism changes the explanandum from what happens in the cognitive architecture of individual pretenders, to what happens in the social situations in which the pretenders enact their ICs. Enactivism could be seen as a research heuristic, or, as Gallagher (2017) writes, "even if enactivism were to be considered a philosophy of nature, it wouldn't be right to conclude that it cannot offer concrete hypotheses or raise novel scientific questions" (p. 24). It could suggest that the future psychological studies on IFP move away from studying IFP in artificial settings, from analyzing IFP through experiments that assume fictional or representational status of the IC, or that focus mostly on providing children and their parents with questionnaires about their individual imaginary experiences. Instead, it could suggest ecologically valid observational and longitudinal studies (*ala* Sawyer's (1997) analysis of pretend play of children) that acknowledge not only children's individual differences, but changes in social practices and cultural differences. The unit of analysis should be the interactional space in which the IFP occurs, not the individual mind of the pretender, and some potential methods of studying IFP should include embodied pretend reenactments (Gallagher & Lindgren, 2015), joint IC play (Hoff, 2004) or a qualitative analysis of a pretend episode (Szokolszky, 2006).

References

Bouldin, P. (2006). An investigation of the fantasy predisposition and fantasy style of children with imaginary companions. *The Journal of Genetic Psychology: Research and Theory on Human Development, 167*(1), 17–29. https://doi.org/10.3200/GNTP.167.1.17-29

Carlson, S. M., & White, R. E. (2013). Executive Function, Pretend Play, and Imagination. In M. Taylor (Ed.), *The Oxford Handbook of the Development of Imagination* (pp. 161-174). Oxford University Press.

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge, MA: MIT Press.

Clowes, R. W., & Mendonça, D. (2016). Representation Redux: Is there still a useful role for representation to play in the context of embodied, dynamicist and situated theories of mind? *New Ideas in Psychology*, 40, 26-47. https://doi.org/10.1016/j.newideapsych.2015.03.002

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in cognitive sciences*, *14*(10), 441-447. https://doi.org/10.1016/j.tics.2010.06.009

Di Paolo, E., Rohde, M., & De Jaegher, H. (2010). Horizons for the enactive mind: Values, social interaction, and play. In H. De Jaegher, H., & M. Rohde (Eds.), *Enaction: Towards a new paradigm for cognitive science* (pp. 33-88). Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9780262014601.003.0003

Farver, J. M., & Howes, C. (1993). Cultural differences in American and Mexican mother-child pretend play. *Merrill-Palmer Quarterly, 39 (3),* 344-358.

Foglia, L., & Grush, R. (2011). The limitations of a purely enactive (Non-representational) account of imagery. *Journal of Consciousness Studies*, *18*(5-6), 35- 43.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford: Oxford University Press.

Gallagher, S. & Lindgren, R. (2015). Enactive Metaphors: Learning Through Full-Body Engagement. *Educational Psychology Review*, *27,* 391– 404. https://doi.org/10.1007/s10648-015-9327-1

Gallagher, S., & Rucińska, Z. (2021). Prospecting performance: Rehearsal and the nature of imagination. *Synthese*. https://doi.org/10.1007/s11229-020-02989-2.

Geurts, B. (2017). Fictional commitments. *Theoretical linguistics*, 43(1-2), 53-60. https://doi.org/10.1515/tl-2017-0003

Gleason, T.R. (2013). Imaginary relationships. In M. Taylor (Ed.), *The Oxford Handbook of the Development of Imagination* (pp. 251-271). Oxford: Oxford University Press.

Godfrey-Smith, P. (2001). On the status and explanatory structure of developmental systems theory. In P. E. Griffiths and R. D. Gray (eds.), *Cycles of Contingency: Developmental Systems and Evolution* (pp. 283–298). Cambridge, MA: MIT Press.

Hoff, E. (2004). A friend living inside me — The forms and functions of imaginary companions. *Imagination, Cognition and Personality*, *24*(2), 151-189. https://doi.org/10.2190/4M9J-76M2-4Q4Q-8KYT

Hoff, E. (2005). Imaginary companions, creativity, and self-Image in middle childhood. *Creativity Research Journal*, *17*(2-3), 167-180. https://doi.org/10.1080/10400419.2005.9651477

Hutto, D. D. (2015). Overly enactive imagination? Radically re-imagining imagining. *The Southern Journal of Philosophy*, *53*, 68-89. https://doi.org/10.1111/sjp.12122

Hutto, D. D. & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content.* Cambridge, MA: MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

Ilundáin-Agurruza, J. (2017). Muscular Imaginings—A Phenomenological and Enactive Model for Imagination. *Sport, Ethics and Philosophy*, *11* (1), 92-108. https://doi.org/10.1080/17511321.2017.1294197.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The Case for Mental Imagery*. Oxford: Oxford University Press.

Leontiev, A. N. (1981). *Problems of the Development of Mind*. Moscow: Progress.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Liao, S. Y., & Szabó Gendler, T. (2011). Pretense and imagination. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 79-94. https://doi.org/10.1002/wcs.91

Medina, J. (2013). An enactivist approach to the imagination: Embodied enactments and "fictional emotions". *American Philosophical Quarterly*, *50*(3), 317-335.

Noë, A. (2014). Concept pluralism, direct perception, and the fragility of presence. In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 27. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570597

Papastathopoulos, S., & Kugiumutzakis, G. (2007). *The intersubjectivity of imagination: The special case of imaginary companions.* In S. Bråten (Ed.), *Advances in consciousness research: Vol. 68. On being moved: From mirror neurons to empathy* (pp. 219–233). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.68.17pap

Peeters, A., & Segundo-Ortin, M. (2019). Misplacing memories? An enactive approach to the virtual memory palace. *Consciousness and Cognition*, *76*, 102834. https://doi.org/10.1016/j.concog.2019.102834.

Piaget, J. (1951). *Play, Dreams and Imitation in Childhood*. London: Routledge & Kegan Paul Ltd.

Podolskiy A. (2012) Internalization. In: Seel N.M. (Eds.), *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4419-1428-6_313

Ramsey W. M. (2007). *Representation Reconsidered*. New York: Cambridge University Press.

Reddy, V. (1991). Playing with other's expectations: Teasing and mucking about in the first year. In A. Whiten (Ed.), *Natural Theories of Mind* (pp. 143–158). Oxford: Blackwell.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology, 26*(4), 325–352. https://doi.org/10.1080/10407413.2014.958035

Sawyer, R.K. (1997). *Pretend Play As Improvisation: Conversation in the Preschool Classroom* (1st ed.). New York: Psychology Press. https://doi.org/10.4324/9780203774359

Singer, D. G., & Singer, J. L. (1990). *The house of make-believe: Children's play and the developing imagination.* Harvard: Harvard University Press.

Somers, J. U., & Yawkey, T. D. (1984). Imaginary play companions: Contributions of creative and intellectual abilities of young children. *The Journal of Creative Behavior*, 18(2), 77-89. https://doi.org/10.1002/j.2162-6057.1984.tb00370.x

Spaulding, *S*. (2010). Embodied cognition and mindreading*. Mind & Language 25,* 119–140. https://doi.org/10.1111/j.1468-0017.2009.01383.x

Stern, D. (1985). *The Interpersonal World of the Infant.* New York: Basic Books.

Stukenbrock, A. (2017). "Intercorporeal Phantasms: Kinesthetic Alignment with Imagined Bodies in Self-Defense Training." *Intercorporeality. Emerging Socialities in Interaction*, 237-263. https://doi.org/10.1093/acprof:oso/9780190210465.001.0001

Szokolszky, A. (2006). Object use in pretend play: Symbolic or functional? In A. Costall & O. Dreier (Eds.), Doing things with things: The design and use of everyday objects (pp. 67–86). London, UK: Ashgate Publishing

Szokolszky, A. (2019). Perceiving metaphors: An approach from developmental ecological psychology. *Metaphor and Symbol*, *34*(1), 17-32. https://doi.org/10.1080/10926488.2019.1591724

Taylor, M. (1999). *Imaginary companions and the children who create them*. New York: Oxford University Press.

Taylor, M. (Ed.). (2013). *The Oxford Handbook of the Development of Imagination*. New York: Oxford University Press.

Taylor, M., Carlson, S., & Shawber, A. (2007). Autonomy and control in children's interactions with imaginary companions. *Proceedings of the British Academy*, 147, 81–100. https://doi.org/10.5871/bacad/9780197264195.003.0004

Valenti, S. S., & Gold, J. M. (1991). Social affordances and interaction I: Introduction. *Ecological Psychology*, *3*(2), 77-98. https://doi.org/10.1207/s15326969eco0302_2

van Dijk, L., & Rietveld, E. (2018). Situated anticipation. *Synthese*, 1-23. https://doi.org/10.1007/s11229-018-02013-8

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge, MA: MIT Press.

Varga, S. (2011). Pretence, social cognition and self-knowledge in autism. *Psychopathology, 44*(1), 46–52. https://doi.org/ 10.1159/000317777

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: MIT Press.

Woolley, J. D. (1995). Young children's understanding of fictional versus epistemic mental representation: Imagination and belief. *Child Development, 66,* 1011–1021. https://doi.org/10.2307/1131795

View publication stats

The International Journal of Indian Psychology
ISSN 2348-5396 (e) | ISSN: 2349-3429 (p)
Volume 4, Issue 1, No. 77, DIP: 18.01.073/20160401
ISBN: 978-1-365-51571-2
http://www.ijip.in | October-December, 2016

Friendship and Psychological Well-Being

Shaheen Falki[1]*, Fauzia Khatoon[2]

ABSTRACT

"Some people go to priests. Others to poetry.I to my friends". This quotation by Virginia Woolfe would be the perfect way to start this study. "A faithful friend is the medicine of life." Again a quotation but this time by Apocrypha is apt to define the purpose of this study. Everyone has friends – sometimes only one, while at other times, many. But what is the importance of friendship or friends to a person? How does having a friend help us in our time of need, problems and sorrow? Is it possible that friendship relaxes us or reduces our tensions? Does friendship have effect on our well-being and especially psychological well-being? The aim of the current study was to find out what role does Friendship play in Psychological Well-Being. The study was carried on the sample of 200 individuals all aged from 17 to 35. Dimensions of Friendship Scale (SunandaChandna and N.K. Chadha, 1986) and Psychological Well-Being Scale (Bhogle and Jai Prakesh, 1995) were used for the purpose of measurement. Results indicated that there is a significant positive correlation (r = .723) at 0.01 level of significance between friendship and psychological well-being.

Keywords: *Friendship, Well-Being, Psychological Well-Being*

It has been rightly said that a friend is someone who can see the truth and pain in you even when you are fooling everyone else. Sometimes it is a friend's shoulder that we need the most during the low phase of our life. At times like these, even our spouse, children or other family members fail to do what a friend does. A friend is the only person who sees us in and out and still loves us the way we are, never trying to change us. And the best part is, unlike our blood relations, it is a mutual choice that makes a friendship work. It has been rightly said by Walter Winchell that, "A real friend is one who walks in when the rest of the world walks out". May be that's why if we have a friend by our side, life doesn't feel lonely anymore. A friend not only shares our problems and gives us the solutions, but also makes us laugh at our problems until we realize that no problem is big enough to make us sad. That's what friends are and that's what friends do. Everyone have friends at all or some stage of our lives. Friendship tends to change over a period

[1] Research Scholar, Dept. of Psychology, AMU, Aligarh, Uttar Pradesh, India, India
[2] Research Scholar, Dept. of Psychology, AMU, Aligarh, Uttar Pradesh, India, India
*Responding Author

© 2016 S Falki, F Khatoon; licensee IJIP. This is an Open Access Research distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any Medium, provided the original work is properly cited.

of time as we age. Sometimes friendship gets stronger with the passage of time while at other times, friendship tends to fade away gradually over time.

There is no particular definition of friendship. It is a relationship with very vague boundaries. It can have a different meaning or emotional value to different people, and also different meaning or emotional value to the same person at different times. We do not have a specific type of friendship. We have true friends, BFF (Best Friends Forever), good friends, casual friends, work friends, etc. and no matter what name we assign to them, they are our friends. Several characteristics are present in all types of friendships. These include affection, sympathy, empathy, honesty, mutual understanding, compassion, enjoyment of each other's company, trust, and the ability to be oneself, express one's feelings, and make mistakes without fear of judgment from the friend. Friendships doesn't have any code of conducts or legal boundaries or any obligations, it's free from all these ties. Maybe that's why we feel free to be ourselves when we are with friends. It have been perfectly quoted by William Shakespeare that, 'A friend is one that knows you as you are, understands where you have been, accepts what you have become, and still, gently allows you to grow.'

Since 1960's friendship became one of the most talked about and written about topic among relationship scholars. Friendship is very essential element which affects our psychological well-being.

In layman language, well-being is a state of being comfortable, healthy and happy with ourselves and our life as whole. But what actually is well-being? Wellbeing, welfare or wellness is a general term for the condition of an individual or group for example their social, economic, psychological, spiritual or medical conditions; While Friedli (2009) defines well-being as feeling good and functioning well. This includes having a fair share of material resources, influence and control, a sense of meaning, belonging and connection with people and place and the capability to manage problems and change.

Psychological well-being refers to how people evaluate or perceive their lives. It is more than simple joy or happiness and it is also more than absence of depression and anxiety. In short, it is the feeling that we are living our life in enjoyable and rewarding way. According to Diener (1997), these evaluations may be in the form of cognitions or in the form of affect. The cognitive aspect includes all information based assessment of one's life which means when a person gives conscious evaluative conclusions about one's satisfaction with life as a whole. The affective aspects deals with a hedonic evaluation guided by emotions and feelings such as frequency with which people experience pleasant/unpleasant moods in relation to their lives. According to Shek (1992), Psychological Well-Being is a state of mentally healthy person who possess a number of positive mental health qualities such as active adjustment to the environment, and unity of

personality." Carol Ryff had broken down psychological wellbeing into 6 different components like:

- **Self-acceptance** –It is how we accept ourselves as we are and how happy and satisfied we are with ourselves.
- **Personal growth** – It is a feeling of continued development and potential and being open to new experiences, in life.
- **Purpose in life –**It is having aim in life or having a thing to look forward to everyday.
- **Environmental mastery –** It is feeling competent and able to manage a complex environment; choosing or creating personally suitable contexts.
- **Autonomy –** It means evaluating ourselves by personal standards.
- **Positive relations with others –** It is having warm, satisfying & trusting relationships.

A World Happiness Database (2010) study found that people with close friendships are happier. It has also been documented that friendship is the most important relationship in an emotional life of an individual between the end of early childhood and onset of full adulthood, and it has also been recorded that friendship is more intense than relationships later in life (Sparks & Glenn, 2007). It has also been seen that the absence of friends can be emotionally damaging (Williams & Alex, 2012). As an individual develop emotionally, it has a sequence in which it is noted that friendship comes after parental bonding and before pair bonding.

Psychological Well-Being in Relation to Friendship

There are numerous studies which prove that friendship has a positive effect on psychological well-being of a person. Friends are important source of positive socialization across the life span (Hartup & Stevens, 1997) and effects of social relationships on well-being have often been conceptualized in terms of social support that reduces the psychological impact of exposure to stress across the life span (Cohen & Wills, 1985). As adolescents make the transition into young adulthood, friends are especially important sources of such support (Carbery & Buhrmester, 1998; Kon & Losenkov, 1978). It has also been observed that support from friends may buffer the effects of stress experienced by college-aged women perhaps because friends are more available than family members for this aged group, both geographically and psychologically (Martin & Burks, 1985). People appear to develop fewer cross-gender friendships throughout the life span than same-gender friendships, but a substantial number of both children and adults report having such friends (Booth & Hess, 1974; Kovacs, Parker & Hoffman, 1996; Wright, 1989) and it appears that these friendships are related to psychological well-being (Hartup& Stevens, 1997)

Objective

- To find out correlation between Friendship and Psychological Well-Being.

Hypothesis

- There will be positive effect of Friendship on Psychological Well-being, i.e. those who score more on Dimensions of Friendship Scale will show high level of Psychological Well-being while those who score low on Dimension of Friendship Scale will show low level of Psychological Well-being.

METHOD

Participants

The samples of this study were citizens of Aligarh. The sample included 200 subjects. Their ages ranged from 17 to 35 years. The data was collected by the means of questionnaire method for Dimensions of Friendship and Psychological Well-Being.

Measures

1. ***Dimension of Friendship Scale*** developed by Sunanda Chandna & N.K. Chadha (1986) was used to measure the status of friendship in individuals. According to this scale friendship is divided into 8 dimensions such as enjoyment, acceptance, trust, respect, mutual assistance, confiding, understanding and spontaneity. This scale consists of 64 items out of which 37 were positively scored and 27 were negatively scored. The test-retest reliability of the scale was found to be .78, while the split-half reliability found out in case of even-odd items was .72 while for first-second half was .76. The validity of the scale was found to be .78.
2. ***Psychological Well-Being Questionnaire*** developed by Bhogly & Jai Prakesh (1995) was used in this study to measure Psychological Well-Being of students. This scale consists of 28 items with alternative response categories – Yes and No, designed to measure 13 dimensions of Psychological Well-Being including meaninglessness, somatic symptoms, selflessness, positive effects, daily activity, life satisfaction, suicidal ideas, personal control, social support, tension, wellness, general efficiency and satisfaction. The reliability was found to be .717. in terms of validity it correlates well, both with Subjective Well-Being Scale (r = .622) and with General Well-Being Scale (r = .484).

Procedure

The present study is based on citizens of Aligarh, Uttar Pradesh. The subjects in this study were between the ages ranging from 17 to 35. All the subjects were given both Scales, i.e. Dimension of Friendship Scale and Psychological Well-Being Scale to complete. The set of instructions were read out to the subjects for both scales and they were asked to inform if they couldn't understand any statement in any of the scale. In both the scales there was 2 response categories 'Yes' and 'No'. Subjects were also asked to fill the scale truthfully and to the best of their knowledge and finally they were reminded that there are no right and wrong answers. After completing the scale all the subjects were thanked and the completed scales were collected from them. Then the scoring of the scales was started.

DATA ANALYSIS

The mean score for Psychological Well-Being was calculated by adding all the scores obtained by individuals divided by the number of individuals (**N=200**). Mean scores of Psychological Well-Being was **17.92** while SD was **4.782**. The mean score of Friendship was **37.32** while SD was **11.350**. Bivariate Pearson Correlation was used for the present study to know whether there is a relationship between Friendship and Psychological Well-Being and if there is, than what is the direction of that relationship – whether it is positive relation, i.e. if there is high scores on Friendship than there is high score on Psychological Well-Being (direct relationship), or whether there is negative relation, i.e. if there is high score on Friendship than there is low score on Psychological Well-Being (inverse relationship). The value obtained by Bivariate Pearson Correlation is .723 which is significant at 0.01 level.

Table 1. Indicates Mean & SD of Friendship and Psychological Well-Being

Descriptive Statistics

Variables	Mean	Std. Deviation	N
Psychological Well-Being	17.92	4.782	200
Friendship	37.32	11.350	200

Table 2. Indicates correlation between Friendship and Psychological Well-Being

Correlations

Variables		PWB	Friendship
Psychological Well-Being	**Pearson Correlation**	1	.723**
	Sig. (2-tailed)		.000
	N	200	200
Friendship	**Pearson Correlation**	.723**	1
	Sig. (2-tailed)	.000	
	N	200	200

*****. Correlation is significant at the 0.01 level (2-tailed).***

RESULT AND DISCUSSION

The result revealed by Table 2 shows that there is a significant correlation between Friendship and Psychological Well-Being. The value of correlation is .723 which is significant at 0.01 level of significance. It also shows the direction of relationship, which is positive meaning that when friendship increases than Psychological Well-Being also increases. This means our null hypothesis is rejected, i.e. there is significant correlation between the scores of Friendship and Psychological Well-Being.

On the basis of correlation, it appeared that Friendship has an on overall positive effect on Psychological Well-Being. The purpose of this paper is to contribute to a theoretical understanding of role of friendship in promoting psychological well-being. The findings of the present study clearly indicate that friendship plays an important role in psychological well-being of an individual. Though the sample of the present study was quite small, previous researches in this area supports the result so this can be generalized.

It is seen that friendship is a very important part of an individual's life. Whether its childhood friendship, teenagers' friendship or a friendship in adulthood and old age. Irwin Sarason rightly quoted that – 'Good friends are good for your health.' Many a times it is only friendship that gives us strength to walk in darkness. 'I would rather walk with a friend in the dark, than alone in the light.' A quote by Helen Keller explains this feeling very well. And it's not only in sadness that a person needs friends but also in their happiness, friends don't lose their importance. We all want to share our happiest moments with our friends – whether it's our results, our first job, our love confessions, or our retirement news.

Acknowledgments
The author appreciates all those who participated in the study and helped to facilitate the research process.

Conflict of Interests
The author declared no conflict of interests.

REFERENCE

Booth, A., & Hess, E. (1974). Cross-sex friendship. *Journal of Marriage and the Family, 36,* 38- 47.

Carbery, J., &Buhrmester, D. (1998). Friendship and need fulfillment during three phases of young adulthood. *Journal of Social and Personal Relationships, 15,* 393-409.

Cohen, S., & Wills, T. A. (1985). Stress, social support, and buffering hypothesis. *Psychological Bulletin, 98,* 310-357.

Diener, E., & Fujita, F. (1997). Social comparisions and subjective well-being. In B. Buunk& R. Gibbons (Eds.), *Health, coping, and social comparison.* (pp. 329-357). Mahwah, NJ: Eribaum.

Friedli, L. (2009) 'Mental health, resilience and inequalities', *World Health Organization*.

Hartup, W. W., & Stevens, N. (1997). Friendships and adaptations in the life course. *Psychological Bulletin, 121,* 355-370.

Kon, I. S., &Losenkov, V. A. (1978). Friendship in adolescence: Values and Behavior. *Journal of Marriage and Family, 40,* 143-155.

Kovacs, D. M., Parker, J. G., & Hoffman, L.W. (1996). Behavioral, affective and social correlates of involvement in cross-sex friendship in elementary school. *Child Development, 67,* 2269-2286.

Martin, B., & Burks, N. (1985). Family and nonfamily components of social support as buffers of stress for college women. *Journal of Applied Social Psychology, 15*, 448-465.

Revisited from Boardofwisdom.Comhttp://boardofwisdom.com/togo/?viewid=1005&listname=Friendship&start=51#.Vd8ILvNViko

Revisited from En.Wikipedia.Org https://en.wikipedia.org/wiki/Well-being

Revisited from Thoughtcatalog.Com http://thoughtcatalog.com/mandy-zucker/2013/10/50-quotes-about-friendship/

Shek, D. (1992). Meaning in life and psychological well-being: an empirical study using the Chinese version of the purpose in life questionnaire. *Journal of Genetic Psychology, 153*, 185-190.

Sparks, Glenn. (2007). *Study shows what makes college buddies lifelong friends.* Purdue University.

Williams, Alex. (2012). "*Friends of a Certain Age: Why Is It Hard To Make Friends Over 30?"*. The New York Times. Retrieved October 25, 2012.

World Happiness Database. Cited in Brace, N. & Byford, J. (Ed.) (2010) Discovering psychology: *What is friendship*. The Open university. ISBN 1-84873-466-2.

Wright, P. H. (1989). Gender differences in adults same and cross-gender friendships. In R.G. Adams & R. Blieszner (Eds.), *Older Adults Friendship: Structure & Process* (pp. 197-221). Newbury Park, CA: Sage.

How to cite this article: S Falki, F Khatoon (2016), Friendship and Psychological Well-Being, International Journal of Indian Psychology, Volume 3, Issue 4, No. 77, ISSN:2348-5396 (e), ISSN:2349-3429 (p), DIP:18.01.073/20160304, ISBN: 978-1-365-51571-2