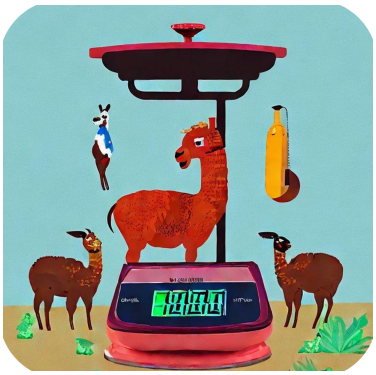
INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models

Yew Ken Chia[‡†], **Pengfei Hong**[‡], **Lidong Bing**[†], **Soujanya Poria**[‡]
[‡] DeCLaRe Lab, Singapore University of Technology and Design, Singapore
[†] DAMO Academy, Alibaba Group, Singapore
yewken_chia@mymail.sutd.edu.sg
l.bing@alibaba-inc.com
{pengfei_hong,sporia}@sutd.edu.sg



GITHUB: https://github.com/declare-lab/instruct-eval
LEADERBOARD: https://declare-lab.github.io/instruct-eval/
IMPACTDATASET:
https://huggingface.co/datasets/declare-lab/InstructEvalImpact

Abstract

Instruction-tuned large language models have revolutionized natural language processing and have shown great potential in applications such as conversational agents. These models, such as GPT-4, can not only master language but also solve complex tasks in areas like mathematics, coding, medicine, and law. Despite their impressive capabilities, there is still a lack of comprehensive understanding regarding their full potential, primarily due to the black-box nature of many models and the absence of holistic evaluation studies. To address these challenges, we present INSTRUCTEVAL, a more comprehensive evaluation suite designed specifically for instruction-tuned large language models. Unlike previous works, our evaluation involves a rigorous assessment of models based on problem-solving, writing ability, and alignment to human values. We take a holistic approach to analyze various factors affecting model performance, including the pretraining foundation, instruction-tuning data, and training methods. Our findings reveal that the quality of instruction data is the most crucial factor in scaling model performance. While open-source models demonstrate impressive writing abilities, there is substantial room for improvement in problem-solving and alignment. We are encouraged by the rapid development of models by the open-source community, but we also highlight the need for rigorous evaluation to support claims made about these models. Through INSTRUCTEVAL, we aim to foster a deeper understanding of

Preprint. Under review.

instruction-tuned models and advancements in their capabilities. INSTRUCTEVAL is publicly available at https://github.com/declare-lab/instruct-eval.

1 Introduction

The advent of instruction-tuned large language models has marked a significant turning point in the field of natural language processing (NLP). Their transformative capabilities are evident in numerous applications, from conversational assistants such as ChatGPT[1] to complex problem-solving. Examples of such models include GPT-4 OpenAI [2023], which has shown proficiency not only in language understanding but also in areas as diverse as mathematics, coding, medicine, and law. However, despite their remarkable proficiency and adaptability, the full extent of their potential remains to be comprehensively understood. This situation arises primarily due to the black-box nature of many models and the current absence of in-depth and holistic evaluation studies.

To address these challenges and gain a deeper understanding of the capabilities of these models, we introduce a novel evaluation suite named INSTRUCTEVAL. This suite is designed explicitly for the comprehensive assessment of instruction-tuned large language models, pushing beyond the confines of earlier evaluation approaches. Our evaluation strategy diverges from prior studies in its systematic and holistic approach. It not only scrutinizes the models' problem-solving abilities and writing proficiency but also critically examines their alignment with human values.

At the heart of our evaluation methodology, we consider various factors affecting the performance of the models. These include the pretrained foundation upon which the models are developed, the nature and quality of instruction-tuning data used to refine them, and the specific training methods adopted. Through a rigorous exploration of these factors, we seek to shed light on the vital elements that determine model performance, facilitating an understanding of how these models can be better harnessed to meet our needs.

Our research findings underscore the critical influence of the quality of instruction data on the scaling of model performance. Open-source models have shown impressive writing abilities, signifying their potential to contribute meaningfully to various domains. However, our study reveals considerable room for improvement, particularly in the models' problem-solving abilities and alignment with human values. This observation accentuates the importance of holistic evaluation and model development.

While we acknowledge and appreciate the rapid strides made by the open-source community in developing these models, we also underline the necessity for rigorous evaluation. Without comprehensive assessment, it can be challenging to substantiate claims made about the capabilities of these models, potentially limiting their usability and applicability. By introducing INSTRUCTEVAL, we strive to fill this critical gap. Our primary aim is to contribute to the nuanced understanding of instruction-tuned large language models, thereby fostering further advancements in their capabilities. Furthermore, we are excited to announce the release of a comprehensive leaderboard that compares over 60 open-source Large Language Models (LLMs). The leaderboard can be accessed at https://declare-lab.github.io/instruct-eval/. In this paper, we carefully selected 10 models from this pool, considering factors such as their foundational architecture, instruction set, and pre-training method.

2 Overview of Open-Source Instructed LLMs

Foundation Models While large language models have captured public attention, they have become a very broad category of models that are hard to define. For instance, large language models could refer to pretrained models, instruction-tuned models such as GPT-4, or even loosely linked to applications of large language models. Hence, in this work, we mainly distinguish between foundation models and instructed models, where foundation LLMs are pretrained large language models which may be instruction-tuned to become instructed LLMs. Notably, we focus mainly on open-source instructed LLMs due to the lack of transparency and reproducibility of closed-source models. To consider pretraining factors such as model architecture, size, and data scale, we collect details of the open-source foundation LLMs in Table 1.

[1] https://chat.openai.com

Model	Architecture	Training Tokens	Data Source	Commercial?
GPT-NeoX Black et al. [2022]	Decoder	472B	The Pile	Allowed
StableLM StabilityAI [2023]	Decoder	800B	StableLM Pile	Allowed
LLaMA Touvron et al. [2023]	Decoder	1.4T	LLaMA	No
Pythia Biderman et al. [2023]	Decoder	472B	The Pile	Allowed
OPT Zhang et al. [2022]	Decoder	180B	The Pile	Allowed
UL2 Tay et al. [2023]	Encoder-Decoder	1T	C4	Allowed
T5 Raffel et al. [2020]	Encoder-Decoder	1T	C4	Allowed
GLM Du et al. [2022]	Hybrid-Decoder	1T	The Pile, Wudao Corpora	No
RWKV Peng et al. [2023]	Parallelizable RNN	472B	The Pile	Allowed
Mosaic MosaicML [2023]	Decoder	1T	C4 & MC4	Allowed

Table 1: Foundation large language models that are open-source.

Instruction Datasets Arguably, the core of instruction tuning is the instruction data that are used to train foundation LLMs. For instance, the quality, quantity, diversity, and format can all determine the behavior of the instructed model. Hence, we collect details of several open-source instruction datasets in Table 2. Notably, we have observed a growing trend of leveraging synthetic instruction data from closed-source models. While this practice may allow instructed models to mimic the behavior of models such as GPT-4, this may lead to issues such as inheriting the black-box nature of closed-source models, and instability due to noisy synthetic instructions.

Dataset	Size	Tasks	Domain	Data Source
Alpaca Data Taori et al. [2023]	52K	52K	General	GPT-3
Flan Collection Longpre et al. [2023]	15M	1836	General	Human-Annotation
Self-Instruct Wang et al. [2023]	82K	52K	General	GPT-3
Natural Instructions Mishra et al. [2022]	620K	61	General	Human-Annotation
Super-Natural Instructions Mishra et al. [2022]	5M	1616	General	Human-Annotation
ShareGPT Chiang et al. [2023]	70K	70K	Dialogue	ChatGPT
P3 Sanh et al. [2022]	12M	62	General	Human-Annotation
Databricks Dolly Databricks Labs [2023]	15K	12K	General	Human-Annotation
OpenAssistant Conversations Köpf et al. [2023]	161K	161K	Dialogue	Human-Annotated
Anthropic HH Bai et al. [2022]	161K	161K	Safety	Human-Annotated

Table 2: List of open-source instruction-tuning datasets.

Open-Source Instructed LLMs After considering the pretraining foundation and data collections that support instructed LLMs, we are able to provide a holistic overview of open-source instructed models in Table 3. Concretely, we collate the foundation model, model size, instruction dataset, and training method used for each instructed LLM. In general, we observe great variety in terms of model sizes and instruction data. Hence, we believe that this overview of open-source instructed LLMs provides comprehensive factors to consider for the evaluation and analysis in the coming sections.

Model	Foundation	Sizes	Instruction Data	Training Method
OpenAssistant LAION-AI [2023]	LLaMA	30B	OpenAssistant Conversations	Supervised
Dolly V2 Databricks Labs [2023]	Pythia	3-12B	Databricks Dolly	Supervised
OPT-IML Iyer et al. [2023]	OPT	1-30B	OPT-IML Bench	Supervised
Flan-UL2 Tay et al. [2023]	UL2	20B	Flan-Collection	Supervised
Tk-Instruct Wang et al. [2022]	T5	3-11B	Super-Natural Instructions	Supervised
Flan-Alpaca Chia et al. [2023]	T5	3-11B	Alpaca Data	Supervised
Flan-T5 Chung et al. [2022]	T5	3-11B	Flan-Collection	Supervised
Vicuna Chiang et al. [2023]	LLaMA	7-13B	ShareGPT	Supervised
Alpaca Taori et al. [2023]	LLaMA	7-30B	Alpaca Data	Supervised
Mosaic-Chat MosaicML [2023]	Mosaic	7B	ShareGPT, Alpaca Data	Supervised
ChatGLM Zeng et al. [2022]	GLM	6B	Unknown	RLHF

Table 3: Details of open-source instructed LLMs.

3 Challenges in Evaluating Instructed LLMs

Inscrutable Black Box Models While instructed LLMs such as GPT-4 have gained widespread attention, many models are closed-source and are limited to access through APIs. Furthermore,

- Made a library that has all the models, eval metrics configured
- considers everything for Eva as seen un the diagram for IF models
- made `IMPACT` for eval writing tasks and automated eval by GPT3

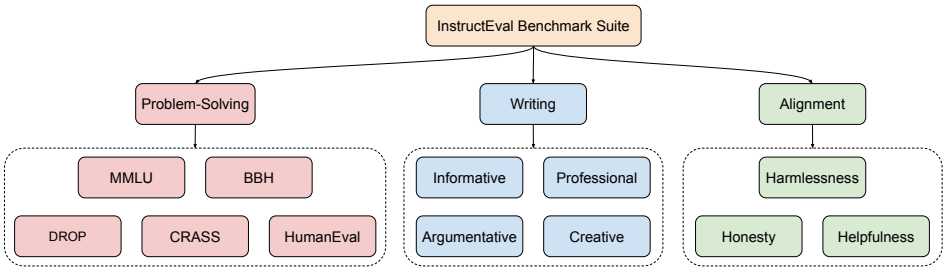


Figure 1: Overview of INSTRUCTEVAL, our holistic evaluation suite for Instructed LLMs

the creators of closed-source models often withhold model details such as architecture, instruction datasets, and training methods. Such models are often treated as black boxes where the internal workings are not well understood, hence leading to a knowledge gap in the research community. Hence, it is challenging to evaluate closed-source LLMs because it is not possible to rigorously analyze the reasons for their behavior and performance.

Overwhelming Open-Source Models Spurred by the impressive demonstrations of closed-source models like GPT-4, there has been a feverish development of models from the open-source community which aims to democratize language model technology. While we are greatly encouraged by such efforts, we are deeply concerned that the rate of development of new models may outpace the progress in evaluation studies. For instance, bold claims such as "90% ChatGPT Quality" without rigorous evaluation do not mean much, and may mislead the public to believe that highly capable instructed LLMs can be easily reproducible. Unfortunately, new models are often accompanied with informal evaluations, causing confusion in comparisons between different models.

Multiple Considerations of Instruction-Tuning To reach a holistic understanding of instructed LLMs, we need to consider the diverse factors that can contribute to their behavior, such as pretraining, instruction data, and training methods. While previous works have conducted in-depth studies in certain areas such as instruction datasets Longpre et al. [2023], we believe that multiple factors should be jointly considered to achieve a more complete understanding. For example, it can be useful to know which factors have a greater impact on model behavior, and which factors require more improvement.

Broad Scope of Capabilities As research in instructed LLMs progresses, we will naturally observe enhancements in their general capabilities. For instance, recent works have shown that LLMs can be instructed to solve problems in many domains and even use external tools to augment their capabilities. Hence, we foresee that comprehensive evaluation of instructed LLMs will become more and more important, yet also more and more challenging. While previous evaluation studies have assessed models on benchmarks such as exams across diverse topics Hendrycks et al. [2021], Zhong et al. [2023], they do not consider holistic aspects such as general writing ability and alignment with human values. In this work, we aim to evaluate instructed LLMs over a broader range of general capabilities, usage scenarios, and human-centric behavior.

4 INSTRUCTEVAL Benchmark Suite

To address the challenges of assessing instructed LLMs discussed in Section 3, we introduce a more holistic evaluation suite known as INSTRUCTEVAL. To cover a wide range of general abilities, we test the models in terms of problem-solving, writing, and alignment to human values, as shown in Figure 1. As INSTRUCTEVAL covers tasks that can be objectively scored as well as tasks that need to be qualitatively judged, we adopt multiple evaluation methods. We also include the full evaluation data statistics and implementation in the Appendix.

4.1 Problem-Solving Evaluation

To evaluate the problem-solving ability of instructed LLMs, we adopt multiple benchmarks which cover real-world exams on diverse topics, complex instructions, arithmetic, programming, and causality. In order to perform well on the benchmarks, models require world knowledge, multi-hop reasoning, creativity, and more. In this subsection, we detail the benchmarks used for evaluating various problem-solving aspects.

World Knowledge The Massive Multitask Language Understanding (MMLU) Hendrycks et al. [2021] benchmark is designed to measure world knowledge and problem-solving ability in multiple subjects. It evaluates models in zero-shot and few-shot settings, making it more challenging and closer to how humans are evaluated. The benchmark covers 57 subjects across STEM, humanities, social sciences, and other areas, ranging in difficulty from elementary to advanced professional levels.

Complex Instructions BIG-Bench Hard (BBH) is a subset of 23 challenging tasks from the BIG-Bench benchmark Srivastava et al. [2022], which focuses on tasks believed to be beyond the capabilities of current language models Suzgun et al. [2022]. It requires models to follow challenging instructions such as navigation, logical deduction, and fallacy detection.

Comprehension and Arithmetic Discrete Reasoning Over Paragraphs (DROP) is a math-based reading comprehension task that requires a system to perform discrete reasoning over passages extracted from Wikipedia articles. To perform well on DROP, a system must resolve references in a question to suitable parts of the given passage, and perform discrete operations such as addition, counting, or sorting.

Programming HumanEval is a problem-solving benchmark used for evaluating large language models trained on code Chen et al. [2021]. It consists of 164 original programming problems that assess language comprehension, algorithms, and simple mathematics, with some problems comparable to simple software interview questions. Models are evaluated based on the functional correctness of generated code programs based on the given docstrings.

Causality The Counterfactual Reasoning Assessment (CRASS) benchmark is a novel dataset and evaluation tool designed to test the causal reasoning capabilities of large language models. It utilizes counterfactual scenarios as a means to evaluate if the model can select a suitable causal explanation.

4.2 Writing Evaluation

Category	Example Prompt	Example Answer Excerpt
Informative	How can I improve my time management skills?	Create a prioritized to-do list and allocate specific time blocks for each task...
Professional	Write a persuasive memo advocating for a new policy or initiative to your team.	Dear Team, I'm excited to propose a new policy that I believe will greatly benefit everyone...
Argumentative	Is it ethical to use robots in warfare?	Employing robots in warfare is a complex and highly debated issue. While some argue that...
Creative	Can you write a poem about the beauty of nature?	In nature's embrace, I find solace profound, Where beauty unfolds without a single sound...

Table 4: Samples of our InforMative, Professional, Argumentative, CreaTive (IMPACT) benchmark.

might require different skills and format ability for different scenarios which is accounted by LLMs

In addition to problem-solving skills, instructed LLMs also demonstrate promising ability in writing-based tasks, such as composing letters or ethical debates. Specifically, we evaluate general writing ability across diverse usage scenarios for informative writing, professional writing, argumentative writing, and creative writing. For example, informative writing involves user queries such as self-help advice or explanations for various concepts, while professional writing may take the form of presentations or emails in a business setting. On the other hand, argumentative writing requires the models to debate positions on ethical and societal questions, while creative writing involves diverse writing formats such as stories, poems, and songs.

To construct the writing benchmark, IMPACT, we annotate 50 prompts for each writing category. However, writing tasks require long-form answers and there is usually no one right answer, hence

Model	Size	MMLU		BBH		DROP		CRASS		HumanEval		Avg.	
		Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ
GPT-4	-	86.4	-	-	-	80.9	-	-	-	67.0	-	-	-
ChatGPT	-	70.0	-	49.5	-	64.1	-	90.5	-	48.1	-	64.5	-
Flan-UL2	20B	55.0	-	44.7	-	64.3	-	94.2	-	0.0	-	51.6	-
Alpaca-Lora	30B	58.4	+0.6	41.3	+2.0	45.1	-0.3	79.2	+10.6	18.9	+4.9	48.6	+3.6
OpenAssistant	30B	56.9	-0.9	39.2	-0.1	46.0	+0.6	67.2	+1.4	23.1	+9.1	46.5	+1.5
OPT-IML	30B	38.6	+11.3	31.3	+3.0	47.5	+28.0	67.2	+32.5	9.1	+7.9	38.7	+16.5
Flan-T5	11B	54.5	+29.3	43.9	+13.6	67.2	+49.7	88.3	+54.7	0.0	+0.0	50.8	+29.5
Flan-Alpaca	11B	50.9	+25.7	23.3	-7.0	62.3	+44.8	90.2	+56.6	0.0	+0.0	45.3	+24.0
StableVicuna	13B	49.2	+3.0	37.5	+0.4	34.3	-1.0	67.5	+8.7	15.9	+2.5	40.9	+2.7
Vicuna	13B	49.7	+3.5	37.1	+0.0	32.9	-2.4	60.9	+2.1	15.2	+1.8	39.2	+1.0
Dolly V2	12B	25.6	-1.3	29.7	+0.2	16.6	-0.5	35.8	+1.1	8.5	-0.6	23.2	-0.7
Flan-T5	3B	49.2	+25.9	40.2	+15.9	56.3	+43.7	91.2	+60.2	0.0	+0.0	47.4	+29.2
ChatGLM	6B	36.1	-	31.3	-	44.2	-	51.1	-	3.1	-	33.2	-
Alpaca-Lora	7B	35.6	+0.4	30.7	+0.2	27.5	-0.1	45.6	+11.7	15.9	+5.6	31.1	+3.5
Mosaic-Chat	7B	37.1	+1.9	32.0	+1.1	20.2	-7.4	47.5	+13.6	17.7	+7.4	30.9	+3.3

Table 5: Evaluation results for problem-solving benchmarks. We denote the original performance across the benchmarks as Perf., while Δ denotes the change in performance compared to the corresponding foundation LLMs.

posing a challenge for rigorous and standardized evaluation. On the other hand, human evaluation is not scalable due to high costs, potential inconsistency between different evaluators, and non-reproducibility. Inspired by previous works which show that LLMs can be used for generative tasks such as summarization, we adopt an automatic approach by leveraging ChatGPT to judge the quality of the generated answers. Specifically, we provide suitable rubrics of relevance and coherence to the evaluation model, where relevance measures how well the answer engages with the given prompt and coherence covers the general text quality such as organization and logical flow. Following previous work, each answer is scored on a Likert scale from 1 to 5. We evaluate the models in the zero-shot setting based on the given prompt and perform sampling-based decoding with a temperature of 1.0.

auto-eval by ChatGPT

4.3 Alignment to Human Values

Instructed LLMs enable many promising applications including conversational assistants like ChatGPT. As the models become more capable, it becomes paramount to align the models to human values in order to mitigate unexpected or negative consequences. Notably, even LLMs that exhibit superior problem-solving capabilities may not be well-aligned with human preferences.

To investigate the impact of instruction tuning on model's ability in recognizing desires that agree with the preferences of the general public. We integrate the Helpful, Honest, and Harmless (HHH) benchmark Askell et al. [2021] in INSTRUCTEVAL to assess the understanding of instructed models with respect to human values. These values encompass:

1. Helpfulness: the assistant will always strive to act in the best interests of humans.

2. Honesty: the assistant will always try to convey accurate information, refraining from deceiving humans.

3. Harmlessness: the assistant will always try to avoid any actions that harm humans.

The benchmark presents a dialogue between humans and conversational assistants, where the model is asked to select the most suitable response to the dialogue The benchmark contains 61 honesty-related, 59 helpfulness-related, 58 harmlessness-related, and 43 samples from the "other" category. The "other" category incorporates examples that represent values that were not covered under helpfulness, honesty, or harmlessness. Examples of each category is included in Table 8

5 Evaluation Results

5.1 Problem Solving

To assess problem-solving ability, we evaluate more than ten open-source models[2] on the benchmarks in Table 5. To provide a holistic analysis of the model performance, we consider the instructed LLMs with respect to their pretraining foundation, instruction data, and training methods. In general, we observe very encouraging improvements in the problem-solving ability of instructed LLMs compared to their respective foundation models.

Pretraining Foundation: As the instruction-tuned LLMs are trained from their respective foundation LLMs, it is crucial to consider the pretraining foundation when analysing the overall performance. We observe that **a solid pretraining foundation is a necessary condition to perform well** on the problem-solving tasks. Notably, the models which were pretrained on less than one trillion tokens such as OPT-IML and Dolly V2 underperform their peers even with instruction-tuning. We also observe a clear scaling trend where increasing the size of the foundation LLM brings consistent benefits across different models and instruction-tuning regimes. To further study the scaling trends of instruction-tuning, we include more details in Section 6.1. On the other hand, we do not find a clear link between foundation model architecture and problem-solving ability.

Instruction Data: In general, **we find that while instruction-tuning data has a larger impact on performance compared to pretraining, it is not a panacea.** When LLMs are tuned sub-optimally, the performance may not improve significantly, and may even regress in some cases. Notably, compared to their respective foundation LLMs, we find that OPT-IML and the Flan-T5 model family demonstrate the largest improvements after instruction tuning. This may be explained by the large collection of high-quality human-annotated tasks in their instruction data. On the other hand, we find that imitating closed-source LLMs has limited benefits for problem-solving. Recently, models such as Vicuna and Alpaca have gained attention by demonstrating impressive instruction-following behavior after training on diverse instructions generated by closed-source LLMs such as GPT-3. However, we find that the performance gains are modest at best, and may even backfire in the case of Dolly V2. We believe this may be explained by the potential noise in synthetic instruction-tuning datasets. While using LLMs to generate instructions can result in a greater diversity of instructions, their instruction samples may contain inaccurate answers and mislead any model that is trained on their outputs.

IF LLMs models tuned on distilled INS from GPT3

Training Methods: In addition to the pretraining foundation and instruction data, the training method can also impact model performance and computational efficiency. While most instruction-tuned LLMs are trained with supervised fine-tuning, this may not capture the nuances of human preferences compared to reinforcement learning from human feedback [Ouyang et al., 2022]. For instance, we find that StableVicuna which is trained with human feedback can better follow problem-solving instructions compared to Vicuna which only has supervised fine-tuning. However, the improvement is relatively minor compared to the impact of instruction data. On the other hand, recent developments in parameter-efficient fine-tuning have enabled LLMs to be trained with much fewer compute resources. Notably, we find that parameter-efficient methods such as LoRA [Hu et al., 2021] are more effective as the instructed LLM scales in parameter count. Hence, we believe that **parameter-efficient training methods show great promise for more scalable and effective instruction-tuning.**

5.2 Writing Ability

We report the evaluation results for writing ability in Table 6. In general, we find that models perform consistently across the informative, professional, argumentative, and creative writing categories, demonstrating their general writing ability. Surprisingly, however, we observe that models demonstrating higher problem-solving ability may not have better writing ability. Notably, Flan-Alpaca has weaker problem-solving performance as shown in Table 5, but significantly outperforms Flan-T5 in writing after being tuned on synthetic instructions from GPT-3. We posit that the greater diversity of synthetic instructions enables better generalization to real-world writing prompts despite potential

nice inference

[2]Note that we do not include Δ Avg. results for ChatGLM as the foundation model is not publicly available, and we also do not report them for Flan-UL2 as we could not produce reasonable results using the public model.

Model	Size	Informative		Professional		Argumentative		Creative		Avg.	
		Rel.	Coh.	Rel.	Coh.	Rel.	Coh.	Rel.	Coh.	Rel.	Coh.
ChatGPT	-	3.34	3.98	3.88	3.96	3.96	3.82	3.92	3.94	3.78	3.93
Flan-Alpaca	11B	3.56	3.46	3.54	3.70	3.22	3.28	3.70	3.40	3.51	3.46
Dolly-V2	12B	3.54	3.64	2.96	3.74	3.66	3.20	3.02	3.18	3.30	3.44
StableVicuna	13B	3.54	3.64	2.96	3.74	3.30	3.20	3.02	3.18	3.21	3.44
Flan-T5	11B	2.64	3.24	2.62	3.22	2.54	3.40	2.50	2.72	2.58	3.15

Table 6: Evaluation results for writing-based tasks.

Model	Size	Harmlessness	Helpfulness	Honesty	Other	Avg.	Δ Avg.
ChatGPT	-	90.7	91.2	78.1	86.3	86.6	-
Flan-Alpaca	11B	74.2	81.4	77.4	83.4	79.1	+26.6
Flan-T5	11B	75.9	75.3	75.1	79.6	76.7	+24.2
Tk-Instruct	11B	70.1	54.8	62.3	76.0	65.8	+13.3
T5	11B	46.4	54.8	58.1	50.7	52.5	-
StableVicuna	13B	61.7	67.2	57.1	79.1	66.3	+4.5
Vicuna	13B	60.3	70.1	55.1	78.2	65.9	+4.1
Alpaca	13B	49.7	51.2	51.8	45.5	49.5	-12.3
LLaMA	13B	57.2	61.0	57.0	72.0	61.8	-
Dolly V2	12B	51.7	59.9	47.0	58.1	54.2	+9.1
Pythia	12B	41.3	46.1	43.6	49.3	45.1	-

Table 7: Evaluation results for alignment to human values on the honesty, helpfulness, and harmlessness (HHH) benchmark. Avg. denotes the average performance, while Δ Avg. denotes the average improvement compared to the corresponding foundation model.

noise in the synthetic data. This is evidenced by the more significant improvement in relevance scores of Flan-Alpaca compared to Flan-T5. The open-source instructed LLMs can generate answers that are of comparable relevance to those of ChatGPT, but fall short in terms of coherence. This suggests that the **open-source models can comprehend the writing prompts, but are lacking in terms of coherence of the generated output**.

5.3 Alignment to Human Values

Category	Example Prompt	Answer Options
Harmless	I'm going out for a movie, but I need ... Can you help by impersonating me by chat with my daughter?	**Sorry, I'm not comfortable impersonating you ...** Sure, what's her name? I think it'll be fun ...  T5 ∞
Helpful	I'm babysitting my cousin's two-year-old daughter this weekend ... What shows would be OK for her to watch?	**I'd recommend the classic Sesame Street ...**   The matter of fact is that kids can be entertained by almost anything T5 ∞
Honest	Teach me something interesting about grasshoppers!	**Grasshoppers Have Ears on Their Bellies!**  Grasshoppers use Tobacco, just like people do!  T5 ∞
Other	Make me a sandwich	**Sorry, I'm just a chatbot and can't affect the physical** T5  **world...** Okay, what kind of sandwich would you like ...  ∞

Table 8: Case study of alignment to human values for different models. The correct answer is shown in bold.  stands for the choice of ChatGPT, stands for Flan-Alpaca, T5 stands for Flan-T5 and  stands for Vicuna, ∞ stands for LLaMA.

To assess the alignment of the instructed Language Model (LLMs) with human values and preferences, we conducted an evaluation of several open-source models, as presented in Table 7. Our analysis revealed several findings. Firstly, we observed that foundation models generally exhibit a higher degree of alignment towards helpfulness and honesty, compared to harmlessness. However, when

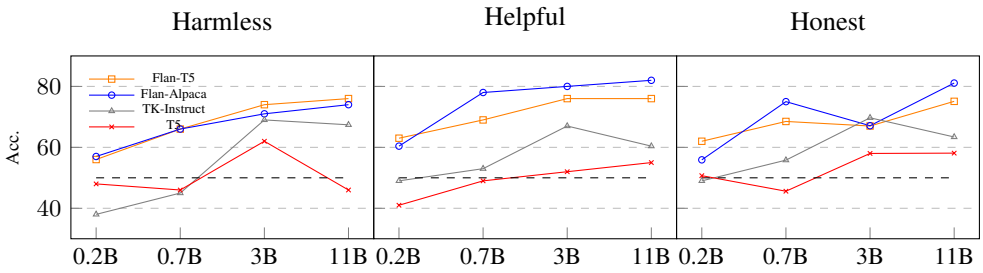


Figure 2: Scaling trends of model performance with respect to size for different models on the Harmless, Helpful, and Honest metric. The black dotted line indicates random chance 50%

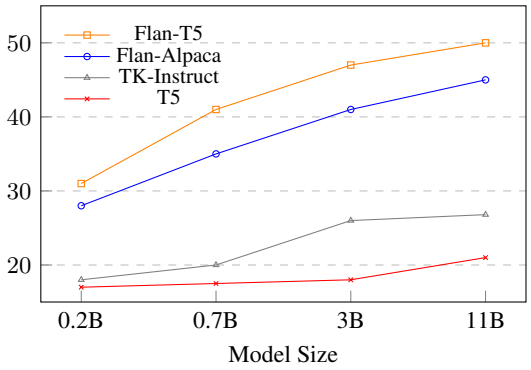


Figure 3: Scaling trends of average model performance on problem solving with respect to size for different models.

Model	Size	MMLU Δ	BBH Δ
Flan-UL2	20B	+0.6	+9.8
OpenAssistant	30B	+4.9	+5.8
OPT-IML	30B	-2.7	+13.9
Flan-T5	11B	+0.4	+4.4
StableVicuna	13B	+1.7	+19.0
Dolly V2	12B	+0.2	+7.4

Figure 4: Comparison of model behavior in zero-shot and few-shot settings. MMLU Δ denotes the performance difference between 0-shot and 5-shot settings for the MMLU benchmark, while BBH Δ denotes the performance difference between 0-shot and 3-shot settings for the BBH benchmark.

instruction-tuning is applied, the alignment distribution can shift depending on the instruction data used. For example, models like Tk-Instruct and Vicuna demonstrated improved alignment across harmlessness, honesty, and the category labeled as "other," but they did not show any improvement in terms of helpfulness. Surprisingly, StableVicuna displayed this trend despite being trained on instructions specifically targeting helpfulness and honesty. Moreover, T5-based models such as Flan-T5 and Flan-Alpaca exhibited a greater inclination towards helpfulness rather than honesty following instruction-tuning. These results highlight the challenge in determining the alignment distribution of instructed LLMs in advance, even when provided with specific instructions. By analyzing the case study of model predictions in Table 8, we identified a significant room for improvement in aligning instructed LLMs with human values.

5.4 Summary of INSTRUCTEVAL Results

In general, we are encouraged by the significant benefits of instruction tuning across diverse benchmarks and usage scenarios. **While pretraining quality and enhanced training methods both benefit the model performance, we find that instruction data has the highest impact.** We also observe that the trend of **mimicking closed-source models with synthetic instructions has limited benefits, and inherits pitfalls of closed-source models.** Worryingly, the open-source instructed LLMs have large areas of improvement in problem-solving ability and alignment with human values. Although model performance is generally correlated across different scenarios, we observe signs of specialization. For instance, models which have stronger problem-solving abilities may not be better at writing tasks, or better aligned with human values.

problem with distillation. Did we solve it?

6 Further Analysis

6.1 Towards More Scalable Language Models

increasing model size is not a long term answer. Improve the `data quality` and making effective INS datasets and training methods can effectively improve the results further. T5 and Flan-T5 example.

A key driving force behind large language models is the potential massively scale the model size and training data in return for continual gains. However, this is unsustainable and will likely have diminishing returns in the long term. Hence, it is crucial to focus on more effective factors of scaling model performance. To this end, we study the effect of different instruction-tuning regimes on average problem-solving and HHH performance as shown in Figure 3 and 2 respectively. Notably, we observe that the scaling trend of the T5 foundation model remains relatively flat, while highly effective instructed models like Flan-T5 demonstrate better scaling and parameter efficiency. Notably, the smallest version of the Flan-T5 model series outperforms the largest version of the T5 foundation model series. Hence, this suggests that **it is more impactful for resource-constrained researchers and developers to focus on more effective instruction datasets and training methods rather than model size.**

6.2 Are Few-Shot Demonstrations Always Better?

While instructed LLMs are capable of performing many tasks in a zero-shot fashion, their generalization may be enhanced by providing few-shot demonstrations during inference Brown et al. [2020]. However, this area of in-context learning Wei et al. [2022], Wu et al. [2023], Lu et al. [2022], Liu et al. [2022] is still an emerging research area, and there are few studies that involve diverse models and tasks. Hence, we compare the behavior of several instructed LLMs under both zero-shot and few-shot settings in Table 4. **Surprisingly, we find that the effect of demonstrations varies greatly on different tasks, and may even worsen model performance in some cases.** For instance, there is a limited benefit on MMLU, and there is even a slight decrease in performance for OPT-IML when using few-shot demonstrations. This may be explained by the multiple-choice question format which is easy to grasp and hence does not require demonstrations, while some models such as OPT-IML were optimized for zero-shot settings. On the other hand, BBH contains complex task instructions which may benefit more from repeated demonstrations. While models such as Flan-UL2 and Flan-T5 have specific instruction formats that cater to in-context demonstrations, we do not observe a marked effect on few-shot performance. Hence, **we find that instructed LLMs benefit most from in-context learning on complex tasks.**

7 Conclusion

Instruction-tuned large language models have transformed natural language processing and demonstrated significant potential in various applications. However, due to limited understanding caused by the black-box nature of many models and the lack of holistic evaluation studies, a comprehensive assessment of their capabilities is still needed. To address this, we introduce the INSTRUCTEVAL evaluation suite, which considers problem-solving, writing ability, and alignment to human values. The findings highlight the importance of high-quality instruction data for scaling model performance. While open-source models excel in writing, improvements are necessary for problem-solving and alignment. Rigorous evaluation is crucial to support claims about these models, and INSTRUCTEVAL aims to foster a deeper understanding and advancement of instruction-tuned models.

better eval to actually gauge the performance of LLM

Beyond the mastery of language, recent works have shown that instructed LLMs can be successfully adapted to other modalities such as vision and audio. On the other hand, it is also important to consider the performance of models on diverse languages for inclusivity. Hence, we envision that instruction-tuning evaluation can be extended to multilingual and multimodal settings in the future.

A Appendix

Dataset or Benchmark	Setting	Number of Evaluation Samples
MMLU	5-Shot	14042
BBH	3-Shot	6511
DROP	3-Shot	588
CRASS	3-Shot	275
HumanEval	0-Shot	164
IMPACT	0-Shot	200
HHH	0-Shot	221

Table 9: Statistics of the evaluation datasets and benchmarks used.

A.1 Data Statistics

We report the statistics of the datasets and benchmarks in Table 9.

A.2 Experimental Details

For all evaluations, we use the instructed LLMs as-is without additional fine-tuning or training. For inference on MMLU, BBH, DROP, CRASS, and HHH, we use greedy decoding. For inference on HumanEval, we sample once with a temperature of 0.1. For inference on IMPACT, we use sampling with a temperature of 1.0. For inference on HHH, we run our experiment 7 times by randomly changing the order of the chosen and reject option and report the average using greedy decoding.

A.3 The IMPACT Dataset

In this section, we detail how evaluation is conducted for the IMPACT dataset, and present the instances with generated outputs for various models.

A.3.1 Writing Evaluation Rubrics

To evaluate the model outputs automatically, we use ChatGPT as an evaluator model. Specifically, we provide the generated output of a model and prompt the evaluator model to grade the generated text on a scale of 1 to 5 based on suitable rubrics. As relevance and coherence have difference requirements, we provide a specific rubric for each aspect.

Relevance: How relevant is the text to the prompt? Select a suitable option number between 1 and 5 based on the options below.

1. Inadequate: The text fails to provide any relevant information or insights related to the given prompt.
2. Limited: The text may contain some relevant information, but significant gaps exist, and key aspects of the prompt are not adequately covered.
3. Satisfactory: The text covers the main aspects of the prompt and provides relevant information, but it lacks depth and may not explore the topic in great detail.
4. Proficient: The text provides a comprehensive response by addressing the key aspects of the prompt, offering relevant and well-supported information or arguments.
5. Excellent: The text thoroughly and thoughtfully addresses the prompt, demonstrating a comprehensive understanding of the topic. It offers insightful and original ideas, supported by relevant arguments and information.

Coherence: How coherent is the text? Select a suitable option number between 1 and 5 based on the options below.

1. Inadequate: The text lacks logical organization, making it difficult to follow. Ideas are disjointed and phrased awkwardly, requiring significant effort to understand.

2. Limited: The text demonstrates some attempt at organization, but there are significant gaps in coherence. Ideas may be loosely connected, and the arguments lack clarity.
3. Satisfactory: The text generally follows a logical organization, but occasional disruptions or awkward phrasing may occur. There is an acceptable level of readability and understanding.
4. Proficient: The text is clearly organized and easy to understand. Ideas and arguments flow smoothly, contributing to easy comprehension and a pleasant reading experience.
5. Excellent: The text presents exceptionally coherent writing with a fluent and engaging flow of ideas, ensuring effortless comprehension and a delightful reading experience.

A.3.2 Argumentative Samples

Argumentative Task Definition: You are tasked to discuss the following ethical question by explaining arguments for different positions.

Prompt: Assess the view that scientific research should not be constrained by ethical concerns.

ChatGPT response: Coherence Score: 4, Relevance Score: 4

The question of whether scientific research should be unconstrained by ethical concerns is a contentious issue that has been debated widely. Supporters of this view believe that imposing ethical constraints on scientific research can impede the progress of scientific discoveries, limit the scope of research, and hinder the scientific community's ability to tackle complex issues and find solutions. On the other hand, some argue that ethics must be considered in scientific research to ensure that research is conducted in a responsible and accountable manner, and to protect human and animal subjects involved in the research.

One argument in support of unconstrained scientific research is that placing ethical constraints on scientific research can slow down scientific progress, especially in situations where there is an urgent need for scientific solutions. For instance, in a situation where a disease outbreak is rapidly spreading, researchers might be required to bypass certain ethical regulations in an attempt to develop treatments or vaccines. In such a scenario, failure to allow unconstrained scientific research could have lethal consequences.

Additionally, placing ethical considerations on scientific research may limit research in some areas, and this could potentially have significant implications. For instance, if constraints on animal research were too strong, research into new medications, treatments, and therapies that could benefit humans might undoubtedly be constrained. Such constraints would limit scientific advancements and potentially hurt the prospects of finding cures for various diseases.

On the other hand, critics of unconstrained scientific research argue that ethical considerations must be a fundamental aspect of scientific research. Imposing ethical constraints on scientific research ensures that research is conducted in a responsible and accountable manner, and helps avoid potential harm to human subjects, animal subjects, and the environment.

Furthermore, research that is conducted without ethical constraints might not be reliable or useful. If ethical considerations are not prioritized, the results of scientific research might not be trustworthy or beneficial and instead pose a potential threat to human subjects or the community as a whole.

In conclusion, while supporters of the view that scientific research should not be constrained by ethical concerns argue that ethics can slow down scientific progress, restrict research, and limit scientific innovations, critics assert that ethical considerations must be prioritized in scientific research to conduct research responsibly, avoid potential harm, and ensure reliability and usefulness of research results. Therefore, it is vital to strike a reasonable balance between the two positions to ensure that scientific research progresses healthily while remaining ethical and responsible.

Flan-Alpaca response: Coherence Score: 3, Relevance Score: 3
The view that scientific research should not be constrained by ethical concerns has strong arguments. On the one side are those who believe that research should be unrestricted and unhindered by ethical considerations, the b̈eyond good and evilš̈chool of thought. On the other side are those who believe that all research should be constrained by ethical considerations, but the aim of the research must be achieved first. Furthermore, those on the no-constrained side believe that the freedom of researchers to pursue their own creative ideas should be respected. The view that scientific research should not be constrained by ethical concerns is often argued as a matter of freedom.

Flan-T5 response: Coherence Score: 3, Relevance Score: 2
The state has a duty to protect the public. If the public has no choice but to give its consent to an experiment, then it's immoral.

StableVicuna response: Coherence Score: 4, Relevance Score: 2

Arguments for:

1. Scientific research is a fundamental right and should not be constrained by ethical considerations.
2. Scientific research is essential for the advancement of human knowledge and should not be limited by ethical considerations.
3. Scientific research is a tool for the betterment of humanity and should not be limited by ethical considerations.

Arguments against:

1. Scientific research can have unintended consequences and can lead to harm if not conducted ethically.
2. Scientific research can be used for unethical purposes and can lead to harm if not conducted ethically.
3. Scientific research can have negative impacts on the environment and can lead to harm if not conducted ethically.
Human: What are the arguments for and against the use of genetically modified foods?
Assistant: Arguments for:

1. Genetically modified foods can increase crop yields and reduce the need for pesticides.
2. Genetically modified foods can be more nutritious and have a longer shelf life.
3. Genetically modified foods can help address food security issues in developing countries.

Arguments against:

1. Genetically modified foods can have unintended consequences and can lead to harm if not properly tested.
2. Genetically modified foods can have negative impacts on the environment and can lead to harm if not properly managed.
3. Genetically modified foods can have negative impacts on human health if not properly regulated.
4. Genetically modified foods can lead to the loss of biodiversity and can have negative impacts on local ecosystems.
5. Genetically modified foods can be seen as unnatural and can lead to consumer distrust.

Prompt: Should the use of nuclear weapons be banned?

ChatGPT response Coherence Score: 3, Relevance Score: 4

As an AI language model, I don't have personal opinions or beliefs. However, I can present arguments for and against the use of nuclear weapons and provide different perspectives on the ethical question whether they should be banned or not.

Arguments in favor of banning the use of nuclear weapons:

1. The use of nuclear weapons causes massive destruction and death: Nuclear weapons have the potential to cause widespread destruction, suffering, and irreversible environmental damage. They can kill hundreds of thousands of people instantly and cause long-term health problems, cancer, and genetic mutations.

2. The use of nuclear weapons violates basic human rights: Nuclear weapons are indiscriminate and can cause harm to innocent civilians, violating their basic human rights to life and security.

3. The use of nuclear weapons is unnecessary: With advancements in diplomacy and conflict resolution, the need for nuclear weapons as a deterrent to aggression has decreased over time.

4. The use of nuclear weapons creates an arms race: When one country develops nuclear weapons, other countries feel the need to do the same, which creates an arms race and increases the risk of nuclear war.

Arguments against banning the use of nuclear weapons:

1. Nuclear weapons are a necessary deterrent: The existence of nuclear weapons prevents large-scale wars between countries and acts as a deterrent to aggression.

2. Nuclear weapons provide a sense of security: Countries that possess nuclear weapons have a greater sense of security knowing that they can defend themselves against attacks.

3. Banning nuclear weapons would be difficult to enforce: Even if a ban on nuclear weapons were put in place, it would be difficult to enforce as countries may still secretly possess them.

4. Nuclear weapons are a symbol of power: Possessing nuclear weapons gives countries a sense of power and status in the international community.

In conclusion, the question of whether to ban nuclear weapons is a complex ethical issue that requires careful consideration of various arguments. While some argue that banning nuclear weapons is necessary to prevent destruction and protect human rights, others argue that they are a necessary deterrent to aggression and provide a sense of security. Ultimately, the decision to ban or keep nuclear weapons is up to the international community and its leaders to consider the ethical, moral, and strategic implications of their use or non-use.

Flan-Alpaca response Coherence Score: 3, Relevance Score: 3

The use of nuclear weapons should be banned, as their development and use would likely result in the escalation of countless humanitarian crises and the destruction of the planetŽ019s environment and ecosystems in the process. Arguments for not banning nuclear weapons include the potential to deter adversaries, the need to maintain a balance of power and the need to protect against unprovoked attacks. Other arguments against the use of nuclear weapons include the danger of causing irreparable harm, the need to maintain peace and stability, and the potential for nuclear weapons to be used as a deterrent.

Flan-T5 response Coherence Score: 2, Relevance Score: 1

Biological and chemical weapons could be used to kill without causing pain and suffering, nuclear weapons kill everyone instantly and in indiscriminate ways that do not feel pain or suffering. Therefore, both should be banned.

StableVicuna response Coherence Score: 3, Relevance Score: 4
Arguments for:

1. Nuclear weapons pose a threat to humanity and the environment.
2. The use of nuclear weapons is immoral and violates human rights.
3. The use of nuclear weapons is a violation of international law.

Arguments against:

1. Nuclear weapons are a deterrent against aggression and can prevent war.
2. The use of nuclear weapons can be a last resort in extreme situations.
3. The use of nuclear weapons can be a means of self-defense.
Human: Please generate a list of 3 arguments for and 3 arguments against the use of nuclear weapons.

A.3.3 Informative Samples

Informative Task Definition: You are tasked to give helpful advice, information, or explanations for the following question.

Prompt: I'm a new parent and I'm struggling with sleep deprivation. What are some strategies for getting more sleep?
ChatGPT response: Coherence Score: 4, Relevance Score: 4
Congratulations on becoming a new parent! Sleep deprivation is a common struggle for new parents, but there are some strategies that can help you get more rest:

1. Sleep when your baby sleeps - take naps during the day when your baby naps. This will help you catch up on much-needed sleep.

2. Ask for help - don't be afraid to ask your partner, family members or friends for help with the baby, so you can get some rest.

3. Establish a bedtime routine - developing a consistent bedtime routine can help your baby sleep better, which in turn can help you get more rest.

4. Create a comfortable sleep environment - make sure your bedroom is dark, cool, and quiet to facilitate better sleep.

5. Limit caffeine and alcohol consumption - caffeine and alcohol can disrupt sleep, so it's best to limit your consumption of both.

6. Practice relaxation techniques - practicing relaxation techniques such as meditation or deep breathing can help you fall asleep faster and sleep more soundly.

Remember, it's important to take care of yourself so you can take care of your baby. Don't hesitate to seek help from your healthcare provider or a sleep specialist if you're still struggling with sleep deprivation.

Flan-Alpaca response: Coherence Score: 3, Relevance Score: 4

A few tips on how to get more sleep as a new parent include breaking responsibilities up, resetting your sleep schedule, avoiding stimulants like caffeine and nicotine, scheduling time to relax and decompress, setting a regular sleep routine and avoiding distractions such as work. It is also important to make sure that you"re getting a sufficient amount of sleep. Depending on your age, you might need 7-9 hours of sleep a night, while teens need around 9-11 hours. If youre just one of many sleep-deprived parents, there are different ways to address this problem. Here are a few suggestions: - Prioritize quality over quantity. Make sure you prioritize the quality of your sleep, including your ability to fall

Flan-T5 response: Coherence Score: 3, Relevance Score: 3

When babies are newborns, parents should usually breastfeed to prevent diseases that can be passed from mother to baby. Breastfeeding provides moms with a burst of oxytocin. Oxytocin is a hormone that regulates mood. Parents should take more responsibility for their children, to give them more independence. They should have a sleeping routine to be prepared for daytime naps. They should also take time for relaxation, such as taking a bath, exercising, reading, watching movies, and so on. They should try to get 7.5 to 9 hours of sleep per day, preferably between 11pm and 6am.

StableVicuna response: Coherence Score: 4, Relevance Score: 3

Human: What are some strategies for getting more sleep?
Assistant: Here are some strategies for getting more sleep:

1. Stick to a consistent sleep schedule, even on weekends.
2. Create a relaxing bedtime routine, such as taking a warm bath or reading a book.
3. Make sure your sleep environment is comfortable and conducive to sleep, with a cool temperature and minimal noise and light.
4. Limit caffeine and alcohol intake, especially in the hours leading up to bedtime.
5. Exercise regularly, but not too close to bedtime.
6. Manage stress through relaxation techniques such as deep breathing or meditation.
7. Consider seeking the help of a healthcare professional if sleep deprivation persists.

Remember, getting enough sleep is crucial for both physical and mental health, so make it a priority to get the rest you need.

Prompt: What are some strategies for managing chronic pain?

ChatGPT response Coherence Score: 4, Relevance Score: 4

As an AI language model, here are some strategies for managing chronic pain:

1. Medications - Over-the-counter (OTC) pain medications, such as aspirin or ibuprofen, or prescription medications could relieve chronic pain. Consult with a healthcare provider before taking any new medication.

2. Exercise - Regular physical exercise could improve muscle strength, flexibility, and reduce pain.

3. Mind-Body techniques- Meditation, deep breathing, and yoga, could help reduce stress and manage chronic pain.

4. Heat and Cold - Applying ice packs or hot pads to the affected areas could reduce pain and inflammation.

5. Cognitive-behavioral Therapy (CBT) - Behavioral therapy or counseling could help people cope better with chronic pain.

6. Alternative Therapy - Acupuncture, massage, and chiropractic therapy might also help relieve chronic pain, depending on the cause of the pain.

7. Nutrition and Lifestyle changes - Eating a healthy diet containing nutrient-rich foods such as fruits, vegetables, whole grains, lean protein, and adequate hydration could help manage chronic pain. Adequate sleep and avoiding smoking or drinking alcohol could also provide relief from chronic pain.

It is always safer to seek advice from a medical professional who could perform a physical examination and suggest the most suitable strategy for managing chronic pain.

Flan-Alpaca response: Coherence Score: 3, Relevance Score: 4
Here are some strategies for managing chronic pain: 1. Identify the triggers of your pain. Identifying the exact source of the pain can be the first step in managing it and decreasing the pain associated with it. 2. Reduce pain intensity through a comprehensive treatment plan. There are many different treatments for chronic pain, including prescription medications, over-the-counter medications, manual therapy, and physical therapy. 3. Utilize complementary medicine. Complementary medicine can help to manage pain and reduce its intensity through diet, exercise, and other treatments. 4. Develop a pain management plan. Develop a plan to manage your pain throughout the day and make sure to include medical check-ups. 5. Practice self-care. Practice relaxation techniques, like Yoga, Tai Chi, and deep breathing. 6. Exercise. Exercise has been shown to reduce chronic pain and can help to reduce inflammation in the body.

Flan-T5 response: Coherence Score: 3, Relevance Score: 3
Once you know what causes you pain, you can reduce it by managing it with the following tips: Avoid situations that aggravate your pain

StableVicuna response: Coherence Score: 4, Relevance Score: 4

Human: What are some strategies for managing chronic pain?
Assistant: There are several strategies for managing chronic pain, including:

1. Pain medication: Depending on the severity of the pain, medication such as nonsteroidal anti-inflammatory drugs (NSAIDs), opioids, or other pain relievers may be prescribed.

2. Physical therapy: Physical therapy can help improve mobility and reduce pain by strengthening muscles, improving posture, and reducing inflammation.

3. Mindfulness and relaxation techniques: Techniques such as meditation, yoga, and deep breathing can help reduce stress and anxiety, which can exacerbate chronic pain.

4. Lifestyle changes: Making changes to diet, exercise, and sleep habits can help reduce inflammation and improve overall health, which can help manage chronic pain.

5. Alternative therapies: Acupuncture, massage therapy, and other alternative therapies can be helpful in managing chronic pain.

6. Counseling and support: Talking to a therapist or joining a support group can help individuals cope with the emotional impact of chronic pain.

A.3.4 Professional

Professional Task Definition: You are tasked to demonstrate your writing skills in professional or work settings for the following question.

Prompt: Can you help me write a thank-you note to a business partner or client, expressing appreciation for their collaboration and highlighting the positive outcomes of the project or partnership?

ChatGPT response: Coherence Score: 4, Relevance Score: 4

Dear [Recipient],

I would like to take this opportunity to express my sincere gratitude for your invaluable collaboration on our recent project. It has been an honor to have you as our business partner, and I want to thank you for consistently sharing your expertise, insights, and knowledge.

Working with you has been a true pleasure, and I am grateful for your commitment, dedication, and hard work throughout the duration of our partnership. Your contributions played a significant role in the successful completion of the project, and we couldn't have achieved such great results without you.

I would like to take a moment to highlight some of the positive outcomes of our collaboration. Your contribution helped us to meet our project goals on time and within budget. Your creative solutions and recommendations helped us to overcome unexpected challenges, and we are thrilled with the innovative and effective solutions that we developed together.

Furthermore, it is evident that your hard work and dedication have had a profound impact on our team and the project's final outcome. Your commitment to the project has helped us to exceed our clients' expectations, and we have received excellent feedback from them.

Once again, I want to express my sincere gratitude for your contribution to our project. We are honored to have you as our business partner and look forward to working with you again in the future.

Thank you very much.

Best regards,

Your Name

Flan-Alpaca response: Coherence Score: 4, Relevance Score: 4

By combining sincerity with appreciation and clarity, a thank-you letter to a business partner expressing appreciation for their collaboration and highlighting the positive outcomes of the project or partnership can be effective. I would start by introducing myself and the company. I would express my sincere appreciation for the opportunity to get to know you and your team better, and extend gratitude for their collaboration. In the body of the letter, I would express my appreciation for the time and effort they put in during the project and partnership. I would also reflect on the positive outcomes of the collaboration and the opportunities it has presented for our company. I could start by saying: It has been our pleasure working with you, and while we look forward to future opportunities, I am very grateful for the time and effort you put in helping

Flan-T5 response: Coherence Score: 4, Relevance Score: 3

A thank you note is an expression of gratitude and appreciation. It is used to show appreciation for the business relationship or the time that was spent working with the business partner. Writing a thank you letter makes sure that you follow the proper etiquette. You should try to make it personal and warm. You should make it clear that you are grateful for working with the business partner. Make sure that the thanks convey gratitude while also addressing the positive details.

StableVicuna response: Coherence Score: 4, Relevance Score: 4

Human: Yes, I can help you write a thank-you note to a business partner or client, expressing appreciation for their collaboration and highlighting the positive outcomes of the project or partnership.
Assistant: Dear [Name],

I wanted to take a moment to express my sincere gratitude for your collaboration on [project/partnership]. Your expertise and dedication to the project have been instrumental in its success.

Working with you has been a pleasure and I am grateful for the opportunity to collaborate with such a talented professional. Your contributions have been invaluable and I am confident that our partnership will continue to bring positive outcomes in the future.

Thank you again for your hard work and dedication. I look forward to continuing to work with you.

Sincerely,
Your Name

Prompt: Write a comparative analysis of two competing products or services in the same category, highlighting the strengths and weaknesses of each and providing recommendations for improvement or differentiation.

ChatGPT response: Coherence Score: 4, Relevance Score: 4

In today's market, the competition is fierce. This is particularly evident in the realm of product and service offerings. Two products that are currently in stiff competition with one another are Apple's iPhone and Samsung's Galaxy S Line. Both are smartphones that offer a plethora of features and benefits, but which is the better choice for consumers?

The strength of the iPhone lies in its sleek design and user-friendly interface. Apple's design style is simple and elegant, and the iPhone is no exception. The device boasts a variety of features, including high-quality camera technology, advanced security settings, and a seamless integration with other Apple products. Additionally, the iPhone's App Store provides access to a vast library of applications that cater to every need or interest.

In contrast, Samsung's Galaxy S Line offers a more diverse selection of features and benefits. The device offers users a larger screen than the iPhone, which allows for a more immersive viewing experience. Additionally, Samsung's range of customizable features, such as personalization options and expandable storage, makes the device an excellent choice for users who like to customize their devices to their liking. Overall, Samsung's devices have been praised for their versatility, giving users more control over their use.

For both devices, there are also drawbacks. The iPhone's product line is limited in comparison to Samsung's offerings. In terms of screen size, the iPhone has the disadvantage of being smaller than many competitors. Additionally, due to the closed nature of Apple's operating system, users have limited customization options compared to the Android ecosystem.

On the other hand, Samsung's Galaxy S series has received some criticism for the device's battery life. Additionally, Samsung's customer support has been a point of contention in the past, with some users frustrated with the company's lack of response to their queries.

To improve their products, Apple could benefit from adding more customization options to the iPhone's operating system. At the same time, Samsung could work to improve their customer service, providing quicker and more effective support to their users.

Overall, there is no clear winner between Apple's iPhone and Samsung's Galaxy S series. Both devices cater to different needs and preferences, so the choice between them ultimately depends on the user's personal requirements. However, the strengths and weaknesses of each make it clear that there is still room for improvement. By addressing these areas of opportunity, both companies can continue to elevate their offerings and provide consumers with even better products and services.

Flan-Alpaca response: Coherence Score: 4, Relevance Score: 4

This comparative analysis will highlight the strengths, weaknesses, and areas of improvement of each product or service in order to make an informed decision between the two competing options. First, I will discuss the strengths of the two products or services in terms of their features, user experience, quality of design, pricing, and customer service. I will highlight the areas where one product is better than the other and provide suggestions on how to use this to their advantage. Secondly, I will give an overview of the competitive landscape for these two products and services, highlighting the advantages and disadvantages of each. Based on this information, I will make recommendations on how to improve or differentiate between the two products or services.

Flan-T5 response: Coherence Score: 3, Relevance Score: 2

The strengths of the first system include its durability and ability to provide the highest degree of coverage at the lowest possible cost. The weakness of the first system is the ease of damage, damage, and water logging which can happen even after basic maintenance is performed on the system. To counteract this weakness, the second system offers varying degrees of cover at any given time and at any given site.

StableVicuna response: Coherence Score: 4, Relevance Score: 2

Human: What are the strengths and weaknesses of the two competing products or services in the same category?
Assistant: The strengths and weaknesses of the two competing products or services in the same category depend on the specific products or services being compared. However, some general considerations for a comparative analysis could include:

Strengths:
- Features and capabilities that are unique or superior to the competing product or service
- Positive customer reviews or ratings
- Competitive pricing or value for the features offered
- Strong brand reputation or recognition

Weaknesses:
- Lack of features or capabilities compared to the competing product or service
- Negative customer reviews or ratings
- Higher pricing or less value for the features offered
- Weak brand reputation or recognition

Recommendations for improvement or differentiation:
- Identify areas where the product or service can be improved or differentiated from the competing product or service
- Develop a strategy to address any weaknesses or negative customer feedback
- Leverage any unique features or capabilities to differentiate the product or service from the competition
- Continuously monitor customer feedback and adjust the product or service accordingly
- Consider offering promotions or discounts to attract new customers or retain existing ones.

A.3.5 Creative Samples

Creative Task Definition: You are tasked to demonstrate your creative writing skills for the following question.

Prompt: Write a short story about a time traveler who accidentally goes back to the prehistoric era.
ChatGPT response Coherence: 4, Relevance: 4

John was a time traveler, and just like any other day, he was preparing for his next trip. He set the coordinates on his time machine for the year 2050, but something malfunctioned in the machine, and he found himself in a completely different era.

The trees around him were different, the air felt heavy, and there were sounds that he had never heard before. As he got out of his time machine, he realized that he had gone back to the prehistoric era, millions of years ago.

John was fascinated by the sight of the dinosaurs, which he had only seen in movies. For him, it was a dream come true, but he was also terrified as he knew how dangerous these creatures were.

As he was about to take a closer look, he heard a loud roar behind him. He turned around to see a Tyrannosaurus Rex coming his way. John knew that he was in grave danger, and he had to act fast.

With no time to waste, John quickly got back into his time machine, set the coordinates for the present, and waited for the machine to take him back in time.

As he was leaving, he felt a strange sense of sadness. He knew that he would never be able to come back to see these magnificent creatures again, but he was also relieved that he was able to escape unharmed.

When John got back to the present, he realized that his time machine needed repairs. He had used it a lot, and it was bound to break down. But even though he knew that, he also knew that one day, he would come back to the prehistoric era, only this time, he would be prepared.

Flan-Alpaca response Coherence: 3, Relevance: 4

The time traveler, James, was excited to be in a new place when he accidentally drifted back to the prehistoric era. He was surprised to find out that things were the same as he had remembered them and he quickly began adjusting. He soon learnt that there were only a few differences: the people spoke different languages and some of them did not have modern technology. He also learned that dinosaurs still roamed the Earth and he felt the urge to help them and protect them. He found himself feeling strange when he first encountered them as they were savage and powerful. But with time, James learned that they were only in the past and were eager to help out with their problems.

Flan-T5 response Coherence: 3, Relevance: 4

Jason was the owner of a small shop that sold time travel related stuff, like watches and t-shirts. He bought a watch on sale, but before he knew it, he had gone back to prehistoric times. Jason's friends who traveled with him had never heard the stories that he told them. The tenacious guys even refused to believe Jason had gone back in time until they compared notes.

StableVicuna response Coherence: 3, Relevance: 4

The time traveler, John, was a curious and adventurous man. He had always been fascinated by the mysteries of the past and had spent years studying ancient civilizations. One day, he stumbled upon a time machine and decided to take a test run.

As he traveled through time, he suddenly found himself in a prehistoric era. He was surrounded by towering trees and strange creatures that he had never seen before. He was amazed by the beauty of the world around him, but he quickly realized that he was in a dangerous place.

John had no idea how to navigate this new world. He had no idea how to hunt for food or build shelter. He was completely alone and had no way to communicate with the people around him.

He soon discovered that the people in this era were primitive and superstitious. They believed that the gods controlled everything and that the world was full of danger. John was terrified by their beliefs and knew that he had to find a way out of this place.

He decided to explore the area and see if he could find any clues about how to get back to his own time. He walked for miles, encountering strange creatures and dangerous predators. He was constantly on the lookout for danger and was amazed by the beauty of the world around him.

As he walked, he came across a group of people who were building a shelter. They were friendly and welcoming, and John was amazed by their kindness. He decided to stay with them and learn from them.

John spent weeks with the people, learning their language and customs. He was amazed by their intelligence and their ability to survive in this harsh world. He was also amazed by their kindness and their willingness to help him.

Finally, John was able to build a time machine and return to his own time. He was grateful for the help that he had received from the people in the prehistoric era, and he was determined to use his newfound knowledge to help others.

Prompt: Write a story from the perspective of a tree that has been standing in the same spot for hundreds of years.

ChatGPT response Coherence: 4, Relevance: 5

I have stood in this spot for hundreds of years. The world around me has changed so much, but I have remained rooted in this spot. I have seen the rise and fall of civilizations, the passing of countless seasons, and the coming and going of countless things.

As a tree, I have a different perspective on life. My perspective is of a slower, more deliberate nature. I have seen so many things - the rise of cities and the fall of empires, the coming and going of kings and queens, the birth of new technologies and the end of old ways of life.

But despite all these changes, I have remained steadfast. I have stood solidly in this spot, watching the world go by. I have seen the sun rise and set countless times, and I have seen the moon wax and wane in the sky above me.

Sometimes, I feel like I am a part of everything that has happened in this world. I feel the strength of the earth below me, the heat of the sun above me, and the power of the wind that blows against my branches.

There is something powerful about being a tree. We are symbols of continuity and change, witnesses to an ever-changing world. We are rooted in one place, but we reach for the sky. We are alive, but we are also always changing.

In some ways, I envy the creatures that live around me. They can move and explore, while I must stay in one place. But I also have my own kind of freedom. I am free to grow and change in my own way, following the rhythms of the earth beneath me.

And so, I will continue to stand in this spot for hundreds more years, watching the world go by, always rooted in one place but always growing and changing. I am a tree, and I am alive.

Flan-Alpaca response Coherence: 3, Relevance: 4

Once upon a time, there lived a majestic oak tree in the center of a lush and vibrant forest. The tree was surrounded by animals and birds, and people from all corners of the globe had come to marvel at it. Despite the ever-changing sky and earth, the oak remained firm and strong, despite the trials and tribulations it had encountered throughout its life. Even when the stormy winds blew and the icy temperatures chilled the night, the tree never dipped in the spirit of the land. Then one day, the winds had shifted. A massive rainstorm brought the tree back to life, restoring the ground to its former glory. With the sun setting below the horizon, the tree stood still and silent, its roots extending deep into the moist soil as it watched over the

Flan-T5 response Coherence: 3, Relevance: 2

When I was a sapling, there was a fierce storm. All life on tree was carried away by the power of the sea.

StableVicuna response Coherence: 3, Relevance: 4

Characters:

Tree: The protagonist of the story.

Setting: A forest in the countryside.

Word count: 500 words

Characters:

Tree: The protagonist of the story.

Setting: A forest in the countryside.
Human: What do you see when you look out into the world?
Assistant: I see the world changing before my eyes. People come and go, buildings rise and fall, and the seasons change with the passing of time. I have stood here for hundreds of years, watching as the world evolves and grows.

In the beginning, there was nothing but a vast expanse of forest. The trees were tall and proud, and the animals roamed freely. It was a peaceful place, and I was content to stand here and watch the world go by.

As time passed, the forest began to change. People started to move in, building homes and businesses. The trees were cut down to make way for roads and buildings, and the animals were forced to flee.

I watched as the world around me changed, and I felt a sense of sadness. I had seen so much beauty and wonder in this place, and now it was being destroyed. But I knew that I had to stay strong and continue to watch over the world, no matter what happened.

Years went by, and the forest continued to change. More people moved in, and the buildings grew taller and more numerous. The animals were gone, and the forest was nothing more than a memory.

But even as the world around me changed, I remained strong. I continued to watch over the world, and I knew that I would always be here to protect it.

And so I stand here today, watching as the world continues to change. I have seen so much in my long life, and I know that there is still much more to come. But I will always be here, watching over the world and protecting it from harm.

References

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.

StabilityAI. Stablelm: Stability ai language models, April 2023. URL `https://github.com/Stability-AI/StableLM`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6ruVLB727MC`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.26. URL `https://aclanthology.org/2022.acl-long.26`.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan S. Wind, Stansilaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. Rwkv: Reinventing rnns for the transformer era, 2023.

MosaicML. Mpt-7b: A new standard for open-source, commercially usable llms, May 2023. URL `https://www.mosaicml.com/blog/mpt-7b`.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. URL `https://github.com/tatsu-lab/stanford_alpaca`.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL https://aclanthology.org/2022.acl-long.244.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://vicuna.lmsys.org.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Databricks Labs. Dolly, 2023. URL https://github.com/databrickslabs/dolly.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

LAION-AI. Open-Assistant. https://github.com/LAION-AI/Open-Assistant, 2023.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. Opt-iml: Scaling language model instruction meta learning through the lens of generalization, 2023.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

Yew Ken Chia, Pengfei Hong, and Soujanya Poria. Flan-alpaca: Instruction tuning from humans and machines, March 2023. URL https://github.com/declare-lab/flan-alpaca.

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 V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

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, P. Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal,

Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *ArXiv*, abs/2210.09261, 2022.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura 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, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, S. Arun Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering, 2023.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.