Blending Is All You Need: Cheaper, Better Alternative to Trillion-Parameters LLM

Xiaoding Lu[§] **Adian Liusie**[§] **Vyas Raina**[§] **Yuwen Zhang**[¶] **William Beauchamp**[§]

[§]University of Cambridge [¶]University College London

Chai Research

Abstract

In conversational AI research, there's a noticeable trend towards developing models with a larger number of parameters, exemplified by models like ChatGPT. While these expansive models tend to generate increasingly better chat responses, they demand significant computational resources and memory. This study explores a pertinent question: Can a combination of smaller models collaboratively achieve comparable or enhanced performance relative to a singular large model? We introduce an approach termed *Blending*, a straightforward yet effective method of integrating multiple chat AIs. Our empirical evidence suggests that when specific smaller models are synergistically blended, they can potentially outperform or match the capabilities of much larger counterparts. For instance, integrating just three models of moderate size (6B/13B paramaeters) can rival or even surpass the performance metrics of a substantially larger model like ChatGPT (175B+ paramaters). This hypothesis is rigorously tested using A/B testing methodologies with a large user base on the Chai research platform over a span of thirty days. The findings underscore the potential of the *Blended* strategy as a viable approach for enhancing chat AI efficacy without a corresponding surge in computational demands. [1]

1 Introduction

Due to the remarkable capabilities of current generative AI technology, pre-trained large language models (LLMs) have found extensive utilization across a diverse array of applications. One such application is in chat AI, where automatic systems are deployed as conversational assistants that keep users engaged in entertaining conversations. A common finding is that as one scales up the number of model parameters and the training data size, the quality and ability of the LLMs dramatically increases. This has led to the current trend of scaling up models to tremendous sizes, with current state-of-the-art systems having hundreds of billions of parameters. Although this has enabled highly capable chat AI with extraordinary emergent abilities, this comes at a practical cost of large inference overheads, with specialized infrastructure required, and access to these systems restricted through public APIs. It is therefore highly desirable to overcome these dramatic practical limitations, and have smaller and efficient chat AIs, while keeping users engaged and maintaining the conversational quality that current 100B+ parameters LLMs have achieved.

Although a single small model is unlikely to compete against the current behemoth state-of-the-art LLMs, one may question whether a group of moderately-sized LLMs can together form a chat AI of equivalent or perhaps better ability. In this work, we introduce Blended, an innovative and simple approach where we demonstrate that, surprisingly, if responses are selected randomly from a group of base chat AIs, the resulting combined chat AI is highly capable and engaging, and can outperform systems with orders of magnitude more parameters. We interestingly observe that the blended model appears to take characteristics that are the "best of all", and that by conditioning a response on the conversational history, a single model with particular properties learns abilities from other systems. This leads to more captivating and diverse responses, and a more engaging user experience. We demonstrate the effectiveness of Blended over large-scale A/B tests on real users on the CHAI platform, where our results show that a Blended ensemble with three 6-13B parameter LLMs, outcompetes OpenAI's 175B+ parameter ChatGPT. We observe significantly higher user retention for blended ensembles than for ChatGPT-based chat AIs, illustrating that users find Blended chat AIs to be more engaging, entertaining and useful, despite

[1]All trained models are provided at https://huggingface.co/ChaiML.

Blended only requiring a fraction of the inference cost and memory overhead.

2 Related Work

2.1 Chat AI approaches

Chat AIs have been developed for a variety of applications, from user assistance to casual interactions (for *chitchat*) (Chen et al., 2017). Early designs were based on rule-based algorithms (Weizenbaum, 1966) which later progressed to generative retrieval-based models (Papangelis et al., 2021). The emergence of pre-trained transformer language models marked a significant change in chat AI development (Zhu, 2022; Vaswani et al., 2017; Zaib et al., 2020), where scaling-up trends led to increasingly larger Transformer-based models finetuned to conversational datasets for the development of chat AIs (Adiwardana et al., 2020; Roller et al., 2021; Bao et al., 2020; Choudhary and Kawahara, 2022; Yan et al., 2022).

Traditionally, chat AIs have been trained with self-supervised methods on conversational datasets. However, more recent approaches highlight the importance of human feedback in training to align better with human expectations of an engaging conversation (Leike et al., 2018; Askell et al., 2021; Gabriel, 2020). This is typically achieved through either reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020) or by using the reward model on its own to select or filter out responses (Dathathri et al., 2019; Irvine et al., 2023)

In our work, our *Blended* approach does not consider how one can train better conversational LLMs, and instead demonstrates that one can leverage a group of existing small conversational LLMs and encourage them to collaborate over a conversation to form a single chat AI that generates more engaging and diverse responses.

2.2 Generative system combination

Systems combination has been well-explored for deep-learning systems, with approaches such as stacking (Wolpert, 1992), negative correlation learning (Liu and Yao, 1999), max-voter schemes (Ju et al., 2018; Simonyan and Zisserman, 2014) or probability averaging (He et al., 2016; Raina et al., 2020; Szegedy et al., 2015) employed for a range of regression and classification tasks. With these ensembling methods, it has further been shown that increasing the diversity of the individual members can lead to better-performing combined systems (Kilimci et al., 2018; Seijo-Pardo et al., 2017).

However, for generative language tasks where the outputs are a sequence of tokens, most ensembling approaches become inapplicable and ineffective. Sequence-level ensembling approaches, though, get around this by often averaging conditional token level probabilities of multiple systems (Sennrich et al., 2015; Freitag et al., 2017; Malinin and Gales, 2021; Fathullah et al., 2021). This approach, however, often requires identical member architectures and access to the output probabilities of the tokens. With an increasing trend of limited black box access to LLMs (e.g. ChatGPT (Liu et al., 2023) and BARD (Nyberg et al., 2021)), ensembling methods that only use output sequences may have practical benefit. Minimum Bayes' Risk (MBR) decoding (Kumar and Byrne, 2004) enables this by using system outputs to select the predicted 'best' system output. Though this approach has traditionally been used for Automatic Speech Recognition (ASR), it has also been successfully applied to NLP tasks (Rosti et al., 2007; Freitag et al., 2022; Manakul et al., 2023; Raina and Gales, 2023). With a growing number of (API-access only) deployed large language models, performing well at different tasks, (Jiang et al., 2023) also observed the need for a method to combine outputs in a blackbox setting. They propose *LLM-Blender* to *blend* the outputs from different language models by first ranking the outputs as per a *PairRanker* and then *fuse* the top-K outputs using a separate deep sequence-to-sequence system (termed *GenFuser*).

As with MBR and LLM-Blender, in this work we also propose an ensembling approach that is able to combine outputs from blackbox language models. However, by designing our method for the specific nature of a multi-turn task (such as dialogue agents) our Blended approach does not require all component systems to generate outputs but instead stochastically selects the system that generates the next response, allowing for model blending at the level of a multi-turn conversation.

3 Blended

3.1 Chat AI

The objective of a chat AI is to design an automatic system that can produce engaging and entertaining conversations that human users can interact with. Let u_k denote the user's kth turn, where each user

turn is a sequence of words, $u_k = (w_1^{(k)} \ldots, w_{|u_k|}^{(k)})$. Similarly, let r_k denote the system's kth generated response, which is also a sequence of words $r_k = (w_1^{(k)}, \ldots, w_{|r_k|}^{(k)})$. As an implicit language model, a particular chat AI, parameterised by θ, models the probability of the next response given the previous conversational history,

$$P(r_k|u_{1:k}, r_{1:k-1}; \theta) \tag{1}$$

During training, the system implicitly learns to assign higher probability to responses that are fluent, engaging and high quality. Therefore an output can simply be sampled from its distribution, either stochastically, or through an approximate search process such as beam search.

$$r_k \sim P(r|u_{1:k}, r_{1:k-1}; \theta) \tag{2}$$

Inspired by InstructGPT (Ouyang et al., 2022) and outlined in (Irvine et al., 2023), state-of-the-art chat AIs tends to follow a three-stage-pipeline. First, a pre-trained language model (PrLM) is fine-tuned on a relevant textual domain, e.g. entertaining literature for the design of an *engaging* chatbot. Second, a reward model is trained using explicit human feedback, for example, by using user engagement as a proxy for response quality (Irvine et al., 2023). Then finally, the reward model is used to improve the original PrLM, either by Proximal Policy Optimisation (Ouyang et al., 2022) or by following a simple rejection sampling strategy.

In developing a particular chat AI, there are many design choices such as the base PrLM, the conversational data used in fine-tuning, and the nature of human feedback used to update the system. One may expect that different recipes and training seeds may lead to highly diverse systems that each demonstrate unique strengths and characteristics. One can then consider how a set of chat AIs can be combined for a system with overall better characteristics.

3.2 Ensembling

In accordance with Bayesian statistical principles, the probability assigned to a particular response can be conceptualized as the marginal expectation taken over all plausible chat AI parameters,

$$P(r_k|u_{1:k}, r_{1:k-1}) \tag{3}$$
$$= \mathbb{E}_{\theta \sim P_\Theta} \left[P(r_k|u_{1:k}, r_{1:k-1}; \theta) \right] \tag{4}$$
$$= \int P_\Theta(\theta) P(r_k|u_{1:k}, r_{1:k-1}; \theta) d\theta \tag{5}$$

In practice, where we only have access to a finite set of chat AI systems $\{\theta_1, \theta_2 ... \theta_N\}$, one can approximate the continuous integral as a discrete summation. Further, one can assume that $P_\Theta(\theta)$ is distributed uniformly over the systems such that $P_\Theta(\theta_n) = \frac{1}{N}$, which may be a valid assumption if the set consists of similarly performing models. This yields the approximation,

$$P(r_k|u_{1:k}, r_{1:k-1}) \tag{6}$$
$$\approx \sum_\theta P_\Theta(\theta) P(r_k|u_{1:k}, r_{1:k-1}; \theta) \tag{7}$$
$$= \frac{1}{N} \sum_{n=1}^{N} P(r_k|u_{1:k}, r_{1:k-1}; \theta_n) \tag{8}$$

3.3 Blended

The objective of our approach is to approximately draw samples from the true ensemble distribution (equation 8). To achieve this approximation, each turn Blended randomly (and uniformly) selects the chat AI θ that generates the current response. This process is illustrated in Algorithm 1. It can be noted that during a conversation, the response generated by a specific chat AI is conditional on all previous responses generated by the previously selected chat AIs. This means that the different chat AIs are able to implicitly influence the output of the current response. As a result, the current response is a *blending* of individual chat AI strengths, as they *collaborate* to create an overall more engaging conversation.

Algorithm 1 Blended Algorithm

1: $k \leftarrow 1$
2: **while** true **do**
3: $u_k \leftarrow$ user's current input turn
4: Sample model parameter $\theta_n \sim P_\Theta$
5: Generate response r_k according to:

$$r_k \sim P(r|u_{1:k}, r_{1:k-1}; \theta_n)$$

6: $k = k + 1$
7: **end while**

4 Evaluating Chat AIs

Evaluating the quality of NLG outputs is a notoriously challenging task (Fabbri et al., 2021; Liusie et al., 2023), where traditional gold-standard approaches use human evaluators that score the quality of generated responses, which can be costly.

However, since chat AIs are by definition deployed in social environments with humans, one can leverage statistics of users interaction as a meaningful and aligned measure of chat AI engagingness and quality. To assess the 'quality' of a chat AI, we consider two main proxy functions: the industry standard *user retention* and the main objective function, *user engagement*.

4.1 User Retention

User retention is a standard industrial measure of a platform's success by measuring the fraction of users that return to the platform k days after joining. Let the control group $\mathcal{G}_n$ be a randomly selected group of new users, where each user in this group will only be served chat AI θ_n. Let $S_n(k)$ be the number of users from $\mathcal{G}_n$ that use the platform and interact with the chat AI on day k. Therefore, the k-day user retention rate, $R(k)$, is simply given by the fraction,

$$R(k) = \frac{S_n(k)}{|\mathcal{G}_n|}. \tag{9}$$

Retention rates from different models can be compared throughout the A/B testing period, where one can compare the immediate and long-term engagement of different chat AIs. Hence, for a considered group $\mathcal{G}_n$ and control group $\mathcal{G}_c$, one can define the test to control **retention ratio**, $q_n(k)$ as

$$q_n(k) = \frac{R_n(k)}{R_c(k)}. \tag{10}$$

Beyond comparing models, it is useful to extract retention curve statistics that can summarize a chat AI's performance with interpretable metrics. Empirical evidence suggests that the retention rate can be modelled well as,

$$R^*(k) = \frac{R(1)}{k^{-\beta}}, \tag{11}$$

where the parameter β indicates the rate of user retention decay days, k. Taking the log of both sides yields;

$$\log(q^*(k)) = \Delta\zeta + \Delta\beta \log k, \tag{12}$$

where $\Delta\zeta = (\log(R_w(1)) - \log(R_c(1))$ and $\Delta\beta = (\beta_w - \beta_c)$. One can therefore use the gradient and intercept of the log-log linear best-fit line to estimate the parameters $\Delta\beta$ and $\Delta\zeta$, which gives a useful comparison of the initial retention ratio and retention ratio decay rate relative to the control chat AI.

4.2 User Engagement

User retention is a useful industry metric, however, it may not perfectly align with the metrics that are of true interest. High-quality, engaging conversations are likely to keep users captivated for longer; therefore we directly define a proxy user engagement metric as the average time spent per visiting user. Let $E^{(u)}(t)$ represent whether a user is *engaged* at a time t,

$$E^{(u)}(t) = \begin{cases} 1, & \text{user interacts in } t-\Delta \text{ to } t+\Delta, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

Then we can define $E_n(t)$, the engagement at time t for all users in cohort $\mathcal{G}_n$, as

$$E_n(t) = \frac{1}{|\mathcal{G}_n|} \sum_{u \in \mathcal{G}_n} E^{(u)}(t). \tag{14}$$

As with user retention, the A/B setting allows for direct comparison of the engagement between different chat AIs. Hence we define the test to control **engagement ratio**, $r_n(t)$ as

$$r_n(t) = \frac{E_n(t)}{E_c(t)}. \tag{15}$$

It is also useful to have an overall single metric for the engagement score of a chat AI over time t. Hence, to obtain this, it is empirically observed that a sensible approximation for a chat AI engagement's decay is [2],

$$E^*(t) = \alpha t^{\gamma}, \tag{16}$$

This then gives a model for the test to control engagement ratio as

$$\log(r^*(t)) = \Delta\alpha + \Delta\gamma \log t, \tag{17}$$

where $\Delta\alpha = (\log(\alpha^{(w)}) - \log(\alpha^{(c)}))$ and $\Delta\gamma = (\gamma^{(w)} - \gamma^{(c)}))$. By plotting $r(t)$ against t, a linear line of best fit can be found, with the parameters $\Delta\alpha$ and $\Delta\gamma$ being the intercept and gradient respectively. This gives the summarising metrics $\Delta\alpha$ and $\Delta\gamma$ to compare the engagement quality of different test chat AIs.

5 Experiments

5.1 Experimental Set Up

Base chat AI systems: In our experiments we consider four different base chat AI systems. We

[2] Periodic oscillations are not modeled here.

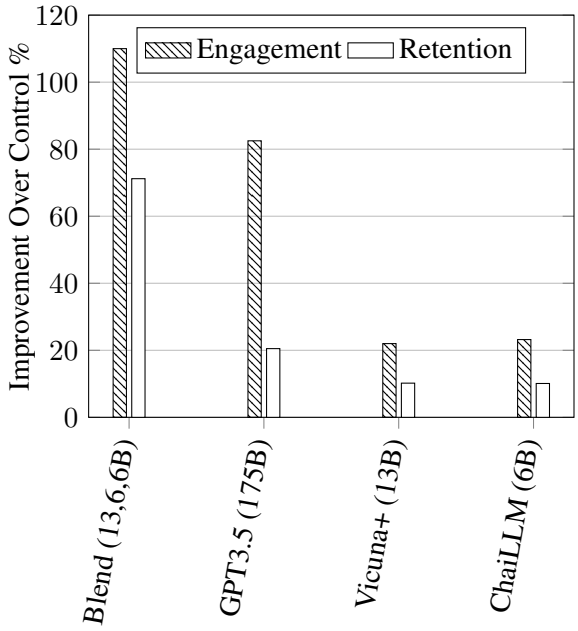

Figure 1: Model performance comparisons, setting the baseline as Pygmalion 6B. Each model is assigned to 5,000 unique new users, graphs report the day 30 retention and engagement improvement with respect to the baseline.

first have 3 moderately sized open-sourced LLMs: Pygmillion 6B[3], Chai Model 6B[4] and Vicuna 13B[5]. Each base LLM has been further finetuned on conversational data, and uses rejection sampling from a trained reward model (detailed in (Irvine et al., 2023)). We finally also consider the state of art chat AI, OpenAI's Davinci (GPT3.5), which has 175B parameters and is only available through a closed API call.

Methodology: Each of the base chat AI systems are deployed with A/B tests on independent user groups, as discussed in Section 3.3, where the groups are of real users engaging with the Chai Research Platform. We conduct a large-scale evaluation with at least 10000 users in each group, and we monitor the user engagement on the platform over a 30-day period. Further, we deploy our blended system (Blended), encompassing Pygmillion, Chai Model and Vicuna. Since there can be external factors that may influence users' retention and engagement (e.g. platform popularity, holidays etc.), systems are only compared using relative engagement and relative retention, which are the metrics normalised to the selected baseline group.

[3] https://huggingface.co/PygmalionAI/pygmalion-6b

[4] https://huggingface.co/ChaiML/edit_sft_pyg_v2e_cp_17515

[5] https://huggingface.co/lmsys/vicuna-13b-v1.3

5.2 Experimental Results

For each chat AI deployed on the Chai Research platform, we compute the user engagement for each day k, as per Equation 15 in an A/B test setting. By considering the 20th day ($k = 20$), Figure 1a shows the engagement ratio of Blended, its constituent chat AIs and Open AI's GPT-3.5. We observe that the moderate-sized chat AIs (Pygmillion, Vicuna and ChaiLLM) have significantly lower engagement than that of GPT3.5, which is expected as GPT3.5 has over an order of magnitude more parameters. However, by blending the three base chat AIs, not only does Blended have higher engagement than each of the constituent systems, but the performance gains are so significant that Blended can outperform OpenAI's GPT3.5. The success of Blended over other chat AIs can also be observed when comparing the $k = 20$ user retention ratio (Equation 10), as seen in Figure 1.

We highlight that Blended has a total of 25B parameters compared to OpenAIs 175B parameters, and further, since responses for Blended are each sampled from a single component chat AI, the inference cost is equivalent to that of a single 6B/13B system. The significant difference in inference speed (measured as the inverse of total Floating Point Operations at test time) is highlighted in Figures 2 and 2 respectively, where it can be observed that Blended offers significant performance gains with respect to engagement and user retention, with speeds similar to that of small chat AIs. Implications of this are strong: instead of scaling up systems to improve quality, one can simply blend multiple smaller open-source systems, and without increasing any inference costs can drastically improve a user's conversational experience. This demonstrates the importance of model collaboration over simple model parameter scaling when designing engaging and successful chat AIs.

As an objective comparison, Table 1 reports the single metric summaries (proposed in Section 3.3). With Pygmillion as the control, we report the test-to-control engagement ratio metrics $\Delta\alpha$ and $\Delta\gamma$, as well as the test-to-control retention ratio metrics $\Delta\zeta$ and $\Delta\beta$. Blended has the highest relative initial engagement, $\Delta\alpha$ and the best engagement ratio decay rate, $\Delta\gamma$. Although the *retention* ratio decay

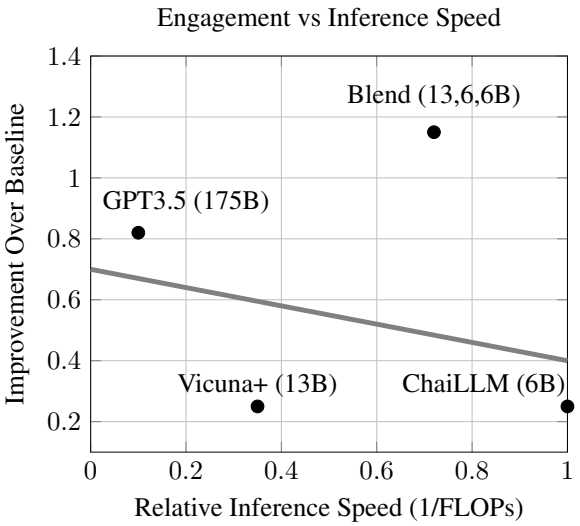

Figure 2: User Engagement

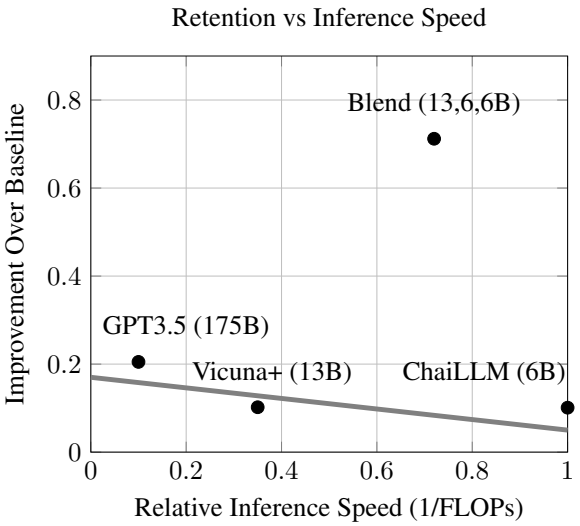

Figure 3: User Retention

rate, $\Delta\beta$ is better for Vicuna than Blended, Vicuna has a significantly lower initial retention ratio, $\Delta\zeta$, demonstrating that Vicuna would require an extended period of time to reach Blended's retention score [6], as can be seen from Figure **??**. Overall it is clear that Blended, using a collaboration of smaller chat AIs, is effective in offering higher quality conversations than a single, much larger chat AI (OpenAI's GPT3.5).

chat AI	$\Delta\zeta$	$\Delta\beta$	$\Delta\gamma$	$\Delta\alpha$	FLOP
Chai	0.1	0.0	0.3	0.2	1.0
Vicuna	-0.4	0.9	0.0	0.1	2.2
Pygmillion (**ctrl**)	0.0	0.0	0.0	0.0	1.0
Blended	**0.2**	0.5	**2.1**	**1.7**	1.4
GPT3.5	0.0	0.3	1.4	0.5	29.2

Table 1: Test to Control Retention and Engagement summary statistics and inference time (total Floating Point Operations / control) for component chat AIs (ChaiModel, Vicuna, Pygmillion (**control**); Blended and OpenAI's Davinci GPT3.5.

6 Future Work

The work demonstrated that Blended, a collaboration of multiple small chat AIs, performs better than a single large-scale chat AI, such as OpenAI's Davinci (ChatGPT). In this section we offer methods by which the Blended model can be further improved to create even more engaging user conversations.

[6]This period of time is estimated to be around one year.

Selection set scaling: Experiments in this work have demonstrated that with even a selection set of three component chat AIs (Chai model, Vicuna and Pygmillion), Blended is able to perform better than the much larger Davinci GPT3.5 model. This performance gain is attributed to the individual expertise of each individual component model that creates a conversation with a diverse set of qualities as the component systems collaborate. Hence, one simple approach to further increase the diversity and thus richness in the conversation is to scale to more than three component systems. Increasing the number of component systems has no computational cost, as inference is always only run through a single system for each response in Blended's methodology. Therefore, future work will explore the impact of increasing the selection set of component chat AIs on the overall quality of conversations.

Optimal Selection Distribution: As demonstrated in Equation 6, Blended in this work adopts a simple approximation for model selection, $P_\Theta(\theta_n) = \frac{1}{N}$. However, although each component chat AI, θ_n, may have some value to add to an overall conversation, an equal contribution from each chat AI may not be the optimal setup. Hence, to combat this, a better approximation for the model selection distribution can be made with,

$$P_\Theta(\theta_n) = \mathcal{F}(u_{1:k}, r_{1:k-1})_n, \qquad (18)$$

where $\mathcal{F}$ is a deep-learning classifier trained to predict the probability distribution over the chat AI selection set for identifying the θ_n to give the

next most *engaging* response r_k. This classifier can be trained using standard signals from Human-Feedback to identify effective and ineffective responses generated in conversations, e.g. if the user regenerated the response it is indicative of being an undesirable response. Future work will explore methodologies to design and train such a classifier, $\mathcal{F}$ to allow for a more optimal (aligned with user engagement) distribution, P_Θ to select the component chat AI for each response, r_k. A further advantage of this approach is that we can now add new chat AIs to the selection set, without the risk of damaging the performance of Blended, as the classifier learns to de-weigh the contribution from bad quality chat AIs.

7 Conclusions

This paper introduced Blended, a simple approach of combining multiple chat AIs by stochastically selecting responses from the different systems. Though simple, the approach is surprisingly powerful and enables a group of three 6-13B parameter models to achieve retention and engagement that is superior to that of the 175B ChatGPT. We demonstrate findings over large scale user A/B tests, which highlights that blending might be a promising solution to improve the quality of chat AIs, all while maintaining inference costs of smaller systems.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. PLATO-2: towards building an open-domain chatbot via curriculum learning. *CoRR*, abs/2006.16779.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Ritvik Choudhary and Daisuke Kawahara. 2022. Grounding in social media: An approach to building a chit-chat dialogue model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 9–15, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *CoRR*, abs/1912.02164.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

Yassir Fathullah, Mark J.F. Gales, and Andrey Malinin. 2021. Ensemble distillation approaches for grammatical error correction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2745–2749.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Iason Gabriel. 2020. Artificial intelligence, values and alignment. *CoRR*, abs/2001.09768.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, Xiaoding Lu, Thomas Rialan, and William Beauchamp. 2023. Rewarding chatbots for real-world engagement with millions of users.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion.

Cheng Ju, Aurélien Bibaut, and Mark van der Laan. 2018. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818.

Zeynep H Kilimci, Selim Akyokus, et al. 2018. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, 2018.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models.

Yong Liu and Xin Yao. 1999. Ensemble learning via negative correlation. *Neural networks*, 12(10):1399–1404.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2023. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Yassir Fathullah, Adian Liusie, Vyas Raina, Vatsal Raina, and Mark Gales. 2023. Cued at probsum 2023: Hierarchical ensemble of summarization models.

Erik P. Nyberg, Ann E. Nicholson, Kevin B. Korb, Michael Wybrow, Ingrid Zukerman, Steven Mascaro, Shreshth Thakur, Abraham Oshni Alvandi, Jeff Riley, Ross Pearson, Shane Morris, Matthieu Herrmann, A.K.M. Azad, Fergus Bolger, Ulrike Hahn, and David Lagnado. 2021. BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, 42(6):1155–1178.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors. 2021. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online.

Vyas Raina and Mark Gales. 2023. Minimum bayes' risk decoding for system combination of grammatical error correction systems.

Vyas Raina, Mark J.F. Gales, and Kate M. Knill. 2020. Universal adversarial attacks on spoken language assessment systems. In *Interspeech 2020*. ISCA.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York. Association for Computational Linguistics.

Borja Seijo-Pardo, Iago Porto-Díaz, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos. 2017. Ensemble feature selection: homogeneous and heterogeneous approaches. *Knowledge-Based Systems*, 118:124–139.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

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

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Rui Yan, Juntao Li, Zhou Yu, et al. 2022. Deep learning for dialogue systems: Chit-chat and beyond. *Foundations and Trends® in Information Retrieval*, 15(5):417–589.

Munazza Zaib, Quan Z. Sheng, and Wei Emma Zhang. 2020. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '20, New York, NY, USA. Association for Computing Machinery.

Zhenyi Zhu. 2022. A simple survey of pre-trained language models. *Preprints.org 202208.0238*.