HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230103

Towards Simple Hybrid Language Model Reasoning Through Human Explanations Enhanced Prompts

Benjamin CLAVIÉ [a], Guillaume SOULIÉ [a] Frederick NAYLOR [a] and Thomas BRIGHTWELL [a]

[a] *Bright Network, Scotland*

Abstract. Large Pre-Trained Models (LLMs) have reached state-of-the-art performance in various Ntural Language Processing (NLP) application tasks. However, an issue remains these models may confidently output incorrect answers, flawed reasoning, or even entirely hallucinate answers. Truly integrating human feedback and corrections is difficult for LLMs, as the traditional approach of fine-tuning is challenging and compute-intensive for LLMs, and the weights for the best models are often not publicly available. However, the ability to interact with these models in natural language opens up new possibilities for Hybrid AI. In this work, we present a very early exploration of Human-Explanations-Enhanced Prompting (HEEP), an approach that aims to help LLMs learn from human annotators' input by storing corrected reasonings and retrieving them on the fly to integrate them into prompts given to the model. Our preliminary results support the idea that HEEP could represent an initial step towards cheap alternatives to fine-tuning and developing human-in-the-loop classification methods at scale, encouraging more efficient interactions between human annotators and LLMs.

Keywords. Large Language Models, Text Classification, Prompt Engineering, Human-feedback Enhanced Prompts, Human-in-the-Loop

Large Pre-Trained Models have increasingly assumed a central role in Natural Language Processing (NLP). Following the state-of-the-art results for numerous tasks achieved primarily with models based on BERT and its variants [1], there has been a growing focus on Generative Large Language Models (LLMs).

In 2022, the development of large models fine-tuned through Instruction Tuning, such as Flan-PaLM and later GPT-3.5 variants based on InstructGPT [2,3], along with the public release of ChatGPT, has intensified the spotlight on LLMs and their applications. GPT-4, the current top-performing LLM, claims state-of-the-art performance on a considerable number of tasks across multiple languages [4].

The remarkable performance of language models is achieved through *prompting* in natural language, which requires careful crafting via *prompt engineering* to elicit stronger model "reasoning" [5]. However, this natural language interaction style has its drawbacks. Notably, as text generators, LLMs are generally unaware of their limitations and can confidently output incorrect answers, flawed reasoning, or even entirely hallucinate answers instead of "admitting" they do not know the answer or lack information [6].

In a recent study, we explored Graduate Job Classification, a task to identify whether a job posting is suitable for a recent graduate or not, using Large Language Models [7].

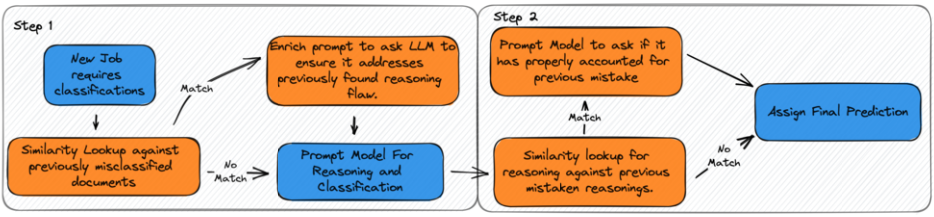

Figure 1. High-level view of the full process. Pre-existing steps in blue, explanation-injection in orange.

We demonstrated that GPT-3.5 significantly outperformed existing state-of-the-art text classification approaches and highlighted the critical importance of prompt engineering.

To ensure that graduates have access to the best possible curation and breadth of jobs, the model's predictions are validated by human annotators. All jobs labeled as "Graduate-suitable" and between 5 and 10% of rejected jobs are sent for review.

This situation, while necessary to maintain quality standards, is less than ideal. Although the LLM classifier reduces the workload for humans involved in classification tasks, it does not genuinely work **with** them. This is not a case of *hybrid* intelligence, as humans review the model's output without any means of influencing it.

Moreso, LLMs often repeat the same reasoning mistakes. Certain concepts or tricky cases consistently result in the same flawed reasoning, which must be corrected manually.

Historically, human feedback and corrections could be integrated into smaller models through fine-tuning. However, this process is challenging with LLMs due to the unavailability of best-performing model weights, which are now only accessible through provider APIs, making it impossible for end-users to fine-tune them. Even for publicly available LLMs like the Flan-* family of models [2], fine-tuning LLMs remains a complex and compute-intensive task, making it prohibitive for smaller organizations.

HEEP. To address this issue and enable "learning" from human annotators, we propose **Human-Explanations-Enhanced Prompting (HEEP)**, a simple approach to retrieve potentially relevant corrections and integrate them into the prompts given to the model. This allows the model reasoning to be directly influenced by human-written annotations relevant to the case at hand.

The HEEP approach is straightforward. For all documents (in our case, job postings) classified by the LLM, we store the model's output, referred to as its *reasoning*, and provide it to human reviewers. In cases of misclassification, reviewers are asked to supply a concise explanation of the reasoning flaws that led to the misclassification. We then store a vectorized representation of the document and the model reasoning, both generated through an embedding model like e5 [8], along with the reasoning correction.

For each new document requiring classification, we perform two vector similarity searches using cosine distance. The first search is conducted on the document itself to find similar documents for which the model's output contained reasoning flaws. If a high-similarity match is found, we enrich the prompt by asking the model to explicitly address the potential mistake before assigning a label. We then perform a second similarity search against previously corrected reasonings. In case of a match, we prompt the model again, asking whether it believes its reasoning contains the same flaw and whether its prediction should be updated accordingly.Finally, we send the model prediction to human validators for review. A high-level overview of the process is presented in Figure 1.

We are currently in the early stages of implementing and deploying HEEP on a larger scale. Preliminary results have shown a significant reduction in repeated mistakes and an increase in precision thanks to step 1, which employs document similarity and is integrated in the original prompt. Step 2, however, seems to negatively affect performance, as the LLM appears hesitant to commit to its answers when confronted about potential flaws in its reasoning in a subsequent prompt, even if its initial response was accurate. Further experiments are efforts are necessary to attempt to mitigate this issue.

We believe HEEP to be a promising approach, providing an efficient alternative to fine-tuning when either the required compute resources or the base model weights are unavailable. Further refinements could assist in enabling human-in-the-loop classification methods at scale, where human annotators' input is immediately leveraged. This represents an improvement over the previous "annotate then train" paradigm, and holds potential for further development, facilitating interactions between human annotators and LLMs.

References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of NAACL-HLT* (2019).

[2] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma et al., Scaling instruction-finetuned language models, *arXiv preprint arXiv:2210.11416* (2022).

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* **35** (2022), 27730–27744.

[4] OpenAI, GPT-4 Technical Report, 2023.

[5] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, in: *Advances in Neural Information Processing Systems*.

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto and P. Fung, Survey of Hallucination in Natural Language Generation, *ACM Computing Surveys* **55**(12) (2023), 1–38. doi:10.1145/3571730.

[7] B. Clavié, A. Ciceu, F. Naylor, G. Soulié and T. Brightwell, Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification, *Proceedings of the 28th Conference on Natural Language Processing and Information Systems (NLDB)* (2023).

[8] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder and F. Wei, Text Embeddings by Weakly-Supervised Contrastive Pre-training, *arXiv preprint arXiv:2212.03533* (2022).