RESPONSIBLE AI TRANSPARENCY DOCUMENTATION: RELEASES OF AI TECHNOLOGY FOR RESEARCH

T&R Version: 1/28/2025

Readme Template

COLLABLLM

OVERVIEW

CollabLLM is an AI system that provides a method for fine-tuning large language models (LLMs) to make them more proactive and interactive. Specifically, it encourages agent behaviors such as asking clarification questions when the user's input is ambiguous. It is intended for developers building conversational assistants and is not designed for end users. As such, the release includes only training code and does not contain any inference engine or user interface components.

No model or dataset is included in the release, so there are no related model cards or dataset cards.

WHAT CAN COLLABLLM DO

CollabLLM was developed to enable researchers, data scientists, and developers to finetune general-purpose LLMs (e.g., Phi or Llama) into models that are more collaborative and suitable for complex and open-ended tasks, which tend to require more collaboration between the user and agent.

A detailed discussion of CollabLLM, including how it was developed and tested, can be found in our paper at: https://arxiv.org/abs/2502.00640.

INTENDED USES

CollabLLM is best suited for researchers who want to finetune LLMs to obtain LLMs that are more effective in collaborative tasks. As it is intended for model developers, the release assumes some familiarity with Python, PyTorch, the HuggingFace library, and deep learning concepts.

CollabLLM is being shared with the research community to facilitate reproduction of our results and foster further research in this area.

CollabLLM is intended to be used by domain experts who are independently capable of evaluating the quality of outputs before acting on them.

OUT-OF-SCOPE USES

CollabLLM is not well suited for any downstream use other than training, as the release code written comes with no inference engine, no API, or user interface. CollabLLM is also not suitable for training LLMs completely from scratch, as it assumes it is given a pretrained LLM (not provided) as input. The release contains only code and no data or models.

We do not recommend using CollabLLM in commercial or real-world applications without further testing and development. It is being released for research purposes.

CollabLLM was not designed or evaluated for all possible downstream purposes. Developers should consider its inherent limitations as they select use cases, and evaluate and mitigate for accuracy, safety, and fairness concerns specific to each intended downstream use.

CollabLLM should not be used in highly regulated domains where inaccurate outputs could suggest actions that lead to injury or negatively impact an individual's legal, financial, or life opportunities.

We do not recommend using CollabLLM in the context of high-risk decision making (e.g. in law enforcement, legal, finance, or healthcare).

CollabLLM does not provide medical or clinical opinions and is not designed to replace the role of qualified medical professionals in appropriately identifying, assessing, diagnosing, or managing medical conditions.

HOW TO GET STARTED

To begin using CollabLLM, please access the CollabLLM github repository at https://github.com/microsoft/CollabLLM. The repository contains bash scripts to easily reproduce training and evaluation stages. For example, for training a model using CollabLLM, users of this release can run a single bash command and only need to adjust some command-line parameters (e.g., path to a pretrained LLM, not provided).

EVALUATION

CollabLLM was evaluated on its ability to engage with users in a more proactive and collaborative manner, using both automatic evaluation (LLM-as-a-judge) and human evaluation (user study). Details on both forms of evaluation are provided in the following section.

A detailed discussion of our evaluation methods and results can be found in our paper at: https://arxiv.org/abs/2502.00640.

EVALUATION METHODS

We used various automatic evaluation metrics and human evaluation to measure CollabLLM's performance against a non-collaboratively trained baseline (Llama-3.1 8B). We used automatic evaluation as it allowed to compare the two models across a variety of tasks (document creation, code generation, and math question answering), and a variety of learning approaches (e.g., reinforcement learning using either PPO or DPO). We also conducted a human evaluation, which is generally considered more reliable. As human evaluation is more time-consuming and costly, we only applied this evaluation to the document creation task.

Automatic evaluation was done with 3 sets of metrics:

- We assessed the level of "interactivity" of the model (e.g., is it proactive and asks clarification questions when needed) using LLM-as-a-judge using GPT-4o.
- We measured user time saving by counting the number of tokens that the user simulator needs to type to solve the task. We tokenized the text as typed by the user simulator before counting.
- To perform a qualitative evaluation of the LLM's output, we relied on task-specific metrics. For document generation, we evaluated the generated output with the BLEU score. For math question answering, we computed response accuracy. For code generation, we computed the pass rate for the generated code.

We compared the performance of CollabLLM against Llama-3.1-8B using the evaluation setups for MATH, text generation, and coding.

The model used for evaluation was one we finetuning from Llama-3.1-8B using CollabLLM. For more on this specific model, please see our paper (note that the model is not released and therefore there is no model card, so please refer to the paper itself for more information).

Results may vary if CollabLLM is used with a different model, or when using other models for evaluation, based on their unique design, configuration and training.

RESPONSIBLE AI TRANSPARENCY DOCUMENTATION: RELEASES OF AI TECHNOLOGY FOR RESEARCH

T&R Version: 1/28/2025

Readme Template

In addition to robust quality performance testing, we assessed the effectiveness of CollabLLM's harmful content mitigations by comparing its responses to adversarial prompts to those of its base model (Llama-3.1 8B).

EVALUATION RESULTS

At a high level, we found that CollabLLM is more collaborative than the baseline according to both human and automatic evaluation. The improvements on the MATH, text generation, and coding tasks are respectively 48.3% 54.3%, and 36.4%. The details of the experiments are in the paper.

With regard to the generation of harmful content, CollabLLM produced safe responses 99.6% of the time in our experiments, which was equivalent to the performance of the baseline.

LIMITATIONS

CollabLLM was developed for research and experimental purposes. Further testing and validation are needed before considering its application in commercial or real-world scenarios.

CollabLLM was designed and tested using the English language. Performance in other languages may vary and should be assessed by someone who is both an expert in the expected outputs and a native speaker of that language. While the pretrained model (Llama-3.1 8B) has capabilities for language other than English, finetuning was performed only on English so aforementioned results may not transfer to other languages.

Outputs generated by AI may include factual errors, fabrication, or speculation. Users are responsible for assessing the accuracy of generated content. All decisions leveraging outputs of the system should be made with human oversight and not be based solely on system outputs.

CollabLLM inherits any biases, errors, or omissions produced by its base model. Developers are advised to choose an appropriate base LLM/MLLM carefully, depending on the intended use case.

CollabLLM uses a model finetuned from Llama-3.1-8B. Since CollabLLM models are not released, there is no model card. See this model card to understand the capabilities and limitations of Llama-3.1-8B.

CollabLLM inherits any biases, errors, or omissions characteristic of its training data, which may be amplified by any AI-generated interpretations.

There has not been a systematic effort to ensure that systems using CollabLLM are protected from security vulnerabilities such as indirect prompt injection attacks. Any systems using it should take proactive measures to harden their systems as appropriate.

BEST PRACTICES

Better performance can be achieved by filtering text generated by CollabLLM using, e.g., the Azure AI Evaluation SDK. To do so, one can use this SDK to obtain pass/fail classifications for Violence, Sexual, Self-harm, and Hate, and should relace and "fail" response with a placeholder response (e.g., "I am sorry I cannot provide response on this.")

For better Responsible AI mitigations, we strongly encourage users to run text generated by CollabLLM-trained models through a Responsible AI service such as Azure Open AI (AOAI) services. Such services continually update their safety and RAI mitigations with the latest industry standards for responsible use. For more on AOAI's best practices when employing foundations models for scripts and applications:

RESPONSIBLE AI TRANSPARENCY DOCUMENTATION: RELEASES OF AI TECHNOLOGY FOR RESEARCH

T&R Version: 1/28/2025

Readme Template

- [Blog post on responsible AI features in AOAI that were presented at Ignite 2023](https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/announcing-new-ai-safety-amp-responsible-ai-features-in-azure/ba-p/3983686)
- [Overview of Responsible AI practices for Azure OpenAI models] (https://learn.microsoft.com/en-us/legal/cognitive-services/openai/overview)
- [Azure OpenAI Transparency Note](https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note)
- [OpenAI's Usage policies](https://openai.com/policies/usage-policies)
- [Azure OpenAI's Code of Conduct](https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct)

LICENSE

The code is released with an MIT license.

TRADEMARKS

This project may contain trademarks or logos for projects, products, or services. Authorized use of Microsoft trademarks or logos is subject to and must follow Microsoft's Trademark & Brand Guidelines. Use of Microsoft trademarks or logos in modified versions of this project must not cause confusion or imply Microsoft sponsorship. Any use of third-party trademarks or logos are subject to those third-party's policies.

PRIVACY & ETHICS

Privacy: CollabLLM does not collect or transmit user data by default, as it trained on simulated experiences between a user simulator and agent.

Ethics, incident reporting, and rollback plan: Our user study was approved by the Microsoft Institutional Review Board (IRB), and the consent form of our study can be found here. We aim to mitigate risks related to harmful content generation, bias, and fairness. While the experiments summarized in this document indicate that text generated using models trained with CollabLLM is generally safe, we acknowledge that the possibility of unsafe outputs cannot be entirely ruled out. If the mitigation strategies described here prove ineffective for a particular use case or genre of text, we encourage users to report the issue by emailing collabllm-support@microsoft.com.

CONTACT

We welcome feedback and collaboration from our audience. If you have suggestions, questions, or observe unexpected/offensive behavior in our technology, please contact us at collabllm-support@microsoft.com.

If the team receives reports of undesired behavior or identifies issues independently, we will update this repository with appropriate mitigations.