How Deep Is Your Chit-Chat? A Teacher-Oriented Evaluation of Deep Learning Models for Open-Domain Conversation

Nikolaos Avrantinis
Computer Engineering and Informatics Dpt.
University of Patras
Patras, Greece
up1059614@ac.upatras.gr

Andreas D. Andriopoulos
Computer Engineering and Informatics Dpt.
University of Patras
Patras, Greece
andriopa@ceid.upatras.gr

Dimitrios A. Koutsomitropoulos, member IEEE
Computer Engineering and Informatics Dpt.
University of Patras
Patras, Greece
koutsomi@ceid.upatras.gr

2025 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) | 979-8-3315-7024-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/INISTA68122.2025.11249589

***Abstract*—Evaluating the conversational capabilities of large language models (LLMs) in open-domain settings remains a major challenge. Unlike task-specific domains, open-domain conversations lack a definitive ground truth, making traditional evaluation metrics insufficient. In this work, we introduce a simplified, teacher-oriented evaluation framework inspired by MDD-Eval. Our approach employs a single classification model, or "teacher," to evaluate generated responses based on dialogue context. By leveraging this structure, we aim to provide a fast, interpretable, and reproducible method for assessing the conversational quality of LLMs without relying on multiple references or costly human input. Our results show that the teacher model effectively differentiates between high- and low-quality responses across several dialogue models, demonstrating the viability of this approach for large-scale open-domain dialogue evaluation.**

I. Introduction

Evaluating open-domain conversational agents remains one of the most persistent challenges in natural language processing (NLP) [1]. Unlike task-oriented dialogue, where responses can be benchmarked against factual or goal-specific criteria, open-domain conversation is inherently subjective and diverse. Chit-chat then makes it difficult to determine what constitutes a "good" response in the absence of a single correct answer.

Traditional evaluation methods, often based on lexical overlap with reference responses [2], struggle to capture the richness of natural dialogue. Metrics such as n-gram similarity or precision-recall balancing are often insensitive to the variability and subtlety of human conversation. In addition, BLEU [4] and ROUGE [3] rely heavily on lexical overlap with reference responses and fail to account for the diversity and context sensitivity of human-like dialogue. A response can be highly appropriate, coherent, and engaging while sharing little lexical similarity with the ground truth, leading to misleadingly low scores under these metrics.

In response to these limitations, there has been a growing effort in the scientific community to establish standardized benchmarks for comparing [5] and evaluating models on open-domain dialogue tasks. These benchmarks aim to provide consistency and objectivity in measuring dialogue quality, while also encouraging progress toward more conversationally capable models.

Recent works such as MDD-Eval [6], USR [7], RUBER [8], GRADE [9] and FED [10], propose various combinations of reference-based, unreferenced, and human-judgment proxies. While effective to an extent, many of these approaches are either computationally intensive, require human annotations, or are difficult to interpret.

Motivated by MDD-Eval, this research presents a simplified and scalable alternative that leverages a single teacher model for response classification. Implementing MDD-Eval with only the teacher model, omitting the student model, is a practical and methodologically valid approach—particularly in the context of evaluating dialogue systems where efficiency, interpretability, and focus are key concerns. The core strength of MDD-Eval lies in its teacher model, which is trained to assign fine-grained quality labels to dialogue-response pairs. By leveraging a well-structured classification framework, the teacher alone can provide valuable insights into the contextual appropriateness and coherence of a response without requiring a generative student model to mimic scoring behavior. This simplification significantly reduces computational overhead and allows researchers to focus directly on the evaluation task, rather than the added complexity of aligning two models. Moreover, this approach retains the core benefit of MDD-Eval—context-aware, learned evaluation—while making the method more accessible and easier to deploy in practical scenarios. As such, a teacher-only implementation represents a

strong, lightweight alternative to the full MDD-Eval pipeline, especially for comparative benchmarking of open-domain chatbots. Our source code and results are openly available at: https://github.com/nikolasavra/OpenDomainChatEval

In the following, we briefly review related work in the field of dialogue evaluation (Section II). Then, in Section III, we introduce our methodology for the teacher-oriented evaluation framework and describe the finetuning process, evaluation datasets and LLMs tested. Section IV presents the evaluation outcomes and provides insights to the results obtained. Finally, Section V includes our conclusions and future work.

II. Related Work

Recent advances in dialogue evaluation have focused on reference-free, learning-based methods that aim to better align with human judgments across multiple dimensions. FineD-Eval [11] introduces a multi-dimensional, reference-free metric that evaluates open-domain dialogues at the dialogue level. It combines coherence, flexibility, and topic depth through a self-supervised training process, leveraging metric ensembling and multitask learning. FineD-Eval demonstrates strong correlation with human annotations across several benchmarks but faces limitations with long or noisy dialogues.

Another promising approach is SLIDE [12], a hybrid framework that integrates small language models (SLMs) with large language models (LLMs) to address the inherent one-to-many nature of open-domain dialogue. SLIDE employs contrastive learning to train SLMs to differentiate between positive and adversarial responses, combining their judgments with LLM scores for enhanced accuracy. It achieves high alignment with human judgments and outperforms traditional lexical metrics like BLEU and ROUGE on datasets such as DailyDialog++ and PersonaChat.

A broader evaluation of LLM-based evaluators is provided in [13], which systematically benchmarks their performance on 12 meta-evaluation datasets targeting coherence, engagingness, and relevance. The study introduces adversarial perturbation strategies to test robustness and finds that instruction-tuned models such as GPT-4 consistently outperform open-source alternatives. However, it also highlights that no current model is fully robust to adversarial variations. The authors contribute enriched datasets that facilitate more rigorous benchmarking of dialogue evaluation metrics.

These works underscore the growing interest in learning-based, reference-free approaches to dialogue evaluation, motivating our implementation and extension of MDD-Eval—a framework designed to combine human-annotated supervision with scalable self-training for robust, multi-domain chatbot evaluation.

III. Methodology

This study proposes a teacher-oriented evaluation framework that utilizes a fine-tuned transformer-based classifier to assess the quality of open-domain responses generated by various large language models (LLMs). The core of this framework is a distilled version of BERT [14]—specifically, the "distilbert-base-uncased-finetuned-sst-2-english" model—selected for its compact size, generalizability, and robust performance in sentiment classification tasks.

A. Teacher Model Fine-Tuning

To adapt the base classifier to the dialogue evaluation task, we fine-tuned it on a dataset of 1,000,000 dialogue samples. Each sample consists of a dialogue context and a corresponding response, labeled into one of three categories:

- Original: Human-like responses that are coherent, contextually appropriate, and informative.
- Adversarial: Responses that are misleading, subtly incorrect, or contextually inappropriate.
- Random: Irrelevant or nonsensical replies with no meaningful link to the given context.

B. Training Dataset

The dataset used for the fine-tuning can be found in the git repository of the MDD-Eval paper.[1]

Label	Dialogue	Response
Adversarial	That is the kind of thing you think has been around forever. Maybe that is why we only have two major parties... So have you ever played fantasy congress?	Fantasy sound like is the whole fantasy sports t see people yelling cspan because politician the move
Original	Yes Tom was busy in his younger days. He was in many iconic films, such as Top Gun and The Color of Money...	Yes he received the golden globe for the film. He has won several awards in his career.
Random	Hello there, do you like Netflix? Hi! I love Netflix! I wonder if they have that new TV show called Grownish...	Was blockbuster mad that to founder, yeah had late pay at he fee a the

TABLE I
Training Dataset Sample

This tri-label classification framework enables the model to distinguish nuanced response quality, analogous to a teacher grading student answers. Table 1 includes three examples, one for each label, in order to better understand the structure of the dataset.

C. Fine-Tuning Strategies

In order to understand the influence of different fine-tuning strategies on classification performance, we explored a variety of training configurations:

- Epoch-based Tuning: The model was trained for 4, 6, 8, and 12 epochs to evaluate learning progression over time.

[1]https://github.com/e0397123/MDD-Eval

- Baseline Comparison: We evaluated the pre-trained model without any additional fine-tuning. As expected, it performed poorly on this task, confirming that specialized supervision is essential for response classification.

D. Evaluation Dataset and LLM Response Collection

To test the trained teacher model, we created an evaluation set of 2,000 open-domain chit-chat prompts. Each prompt was fed into five distinct LLMs: BlenderBot, Mistral, GPT-2, GPT-Neo, and DialoGPT. These models were selected to represent a diverse range of architectures, training objectives, and capacities.

GPT-2: Autoregressive model with 117 million parameters, designed for general-purpose text generation, with no fine-tuning for dialogue.

DialoGPT: Fine-tuned GPT-2 model with 345 million parameters, explicitly trained on conversational datasets [15].

BlenderBot: Dialogue-specific model with 3 billion parameters, designed for open-domain and task-oriented conversations [16].

GPT-Neo: Autoregressive model with 1.3 billion parameters, trained on diverse text data without explicit dialogue fine-tuning.

Mistral: Autoregressive model with 7 billion parameters, representing the largest model in the tests, without specific fine-tuning for dialogue [17].

For each model, the generated response was paired with the original prompt and submitted to the teacher classifier for labeling. This yielded a labeled set of model outputs for comparative analysis.

E. Evaluation Metrics

Two main metrics were used to assess the conversational competence of the LLMs:

1) **Original Label Percentage:** The proportion of responses classified as "original" by the teacher model. This metric serves as a proxy for response quality, assuming that a higher percentage of original labels correlates with more human-like and contextually appropriate outputs.
2) **Relevance Score:** In our study, this score is defined as the probability the teacher model assigns to the "original" class, regardless of which label is ultimately predicted. This probabilistic measure captures the model's confidence in the response being human-like, even in borderline or ambiguous cases. It complements the categorical accuracy by offering a graded view of contextual relevance and quality.

Together, these metrics offer a dual-view of performance: categorical classification for interpretability and semantic similarity for depth. This methodology facilitates a systematic, scalable, and interpretable way to evaluate LLMs on open-domain dialogue tasks—an area where traditional metrics fall short.

IV. Experiments and Results

A. Configuration

All experiments were conducted on a machine equipped with an NVIDIA Tesla V100 GPU with 32 GB of memory. This computational setup ensured that both training and evaluation processes could be performed efficiently at scale. The implementation leveraged a suite of widely used Python libraries. For model development and training, we used PyTorch, a flexible deep learning framework that enabled rapid experimentation and GPU acceleration. The Transformers library by Hugging Face was employed for loading and fine-tuning pretrained language models, including DistilBERT and the conversational LLMs evaluated in this study. Evaluation and classification metrics were implemented using scikit-learn, which provided tools for precision, recall, and confidence score analysis. Additionally, Hugging Face's datasets and tokenizers libraries facilitated data preprocessing, batching, and token-level operations essential to model training. When applicable, we also used the Hugging Face inference API to query hosted LLMs (e.g., Mistral) to ensure reproducibility across environments. This software and hardware environment provided a robust foundation for training the teacher model, performing inference on multiple LLMs, and efficiently handling large-scale experimental data.

B. Results

Tables II-IV present evaluation scores of the 5 models for teacher finetuning with 4, 6 and 8 epochs respectively. Fig. 1 outlines overall relevance scores including the baseline case of zero finetuning for the teacher model (DistilBERT).

Model	Parameters (B)	Architecture	Original Labels %	Relevance Score
GPT-2	0.117	Autoregressive	51%	0.51
DialoGPT	0.345	Fine-tuned for dialogue	72%	0.72
BlenderBot	3	Fine-tuned for dialogue	**88%**	**0.88**
GPT-Neo	1.3	Autoregressive	42%	0.42
Mistral	7	Autoregressive	38.57%	0.38

TABLE II
Model Evaluation for 4 epochs

C. Analysis

These results reveal several important trends:

- BlenderBot consistently achieved the highest percentage of original labels and relevance scores across all epochs. This is expected due to its architecture and training: BlenderBot is a transformer-based, autoregressive model fine-tuned on multiple large-scale conversational datasets such as ConvAI2 and Blended Skill Talk. It has approximately 3 billion parameters and is explicitly designed to handle dialogue coherence, empathy, and knowledge

Model	Parameters (B)	Architecture	Original Labels %	Relevance Score
GPT-2	0.117	Autoregressive	48%	0.47
DialoGPT	0.345	Fine-tuned for dialogue	71%	0.72
BlenderBot	3	Fine-tuned for dialogue	**87%**	**0.87**
GPT-Neo	1.3	Autoregressive	39.61%	0.39
Mistral	7	Autoregressive	33%	0.33

TABLE III
MODEL EVALUATION FOR 6 EPOCHS

Model	Parameters (B)	Architecture	Original Labels %	Relevance Score
GPT-2	0.117	Autoregressive	45%	0.45
DialoGPT	0.345	Fine-tuned for dialogue	71%	0.71
BlenderBot	3	Fine-tuned for dialogue	**87%**	**0.87**
GPT-Neo	1.3	Autoregressive	31%	0.30
Mistral	7	Autoregressive	30%	0.30

TABLE IV
MODEL EVALUATION FOR 8 EPOCHS

Model	Parameters (B)	Architecture	Original Labels %	Relevance Score
GPT-2	0.117	Autoregressive	43.83%	0.43
DialoGPT	0.345	Fine-tuned for dialogue	71%	0.71
BlenderBot	3	Fine-tuned for dialogue	**83%**	**0.83**
GPT-Neo	1.3	Autoregressive	35%	0.35
Mistral	7	Autoregressive	28%	0.28

TABLE V
MODEL EVALUATION FOR 12 EPOCHS

grounding. Its performance confirms that models purpose-built and optimized for open-domain conversation will be recognized as superior by our teacher classifier.

- DialoGPT also performed well, although slightly behind BlenderBot. DialoGPT is a dialogue-specialized variant of GPT-2, trained on 147M Reddit conversations and leveraging the autoregressive architecture of GPT-2. With a parameter count of roughly 345 million in the medium version, it maintains fluency and contextual relevance in responses. The consistent results across epochs validate the teacher model's ability to identify responses that resemble real human dialogue.
- GPT-2 showed moderate performance. While it shares the autoregressive architecture of DialoGPT, it lacks any dialogue-specific fine-tuning, having been pretrained on a general web corpus (WebText). It has 117M parameters in the base version, which may further limit its ability to handle extended dialogue turns with contextual depth. The results align with the expected gap between general-purpose and dialogue-specialized models.
- GPT-Neo performed poorly despite being larger than GPT-2 (1.3B parameters). GPTNeo replicates GPT-2's autoregressive transformer architecture and was trained on The Pile, a diverse dataset that contains structured and unstructured text, but with limited conversational data. Its results suggest that sheer parameter count does not compensate for the lack of domain alignment in pretraining.
- Mistral, though a cutting-edge general-purpose model with impressive results in reasoning and general NLP tasks, scored lowest in this dialogue evaluation. This could be attributed to two key factors: (1) a lack of dialogue-specific fine-tuning, and (2) a potential mismatch between its pre-training distribution and informal, human-like chit-chat. While its architecture and training size are state-of-the-art, they do not compensate for the absence of conversational grounding.

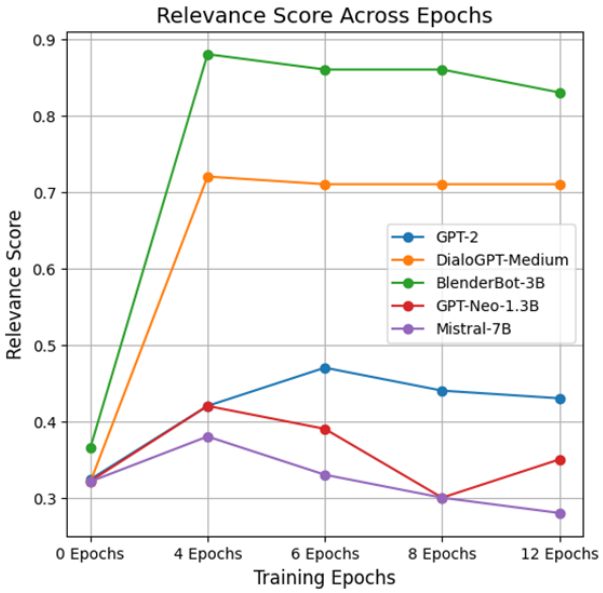

Fig. 1. Relevance scores for 0-4-8-12 teacher finetuning epochs.

It's also worth considering whether the teacher model exhibits architectural or domain bias. Since it is based on DistilBERT—a transformer model more similar in style and language exposure to GPT-2, DialoGPT, and BlenderBot—it is possible that it is more attuned to language patterns common in these models. However, the sharp performance gradient between BlenderBot/DialoGPT and Mistral suggests that the issue is more likely due to model training objectives than structural favoritism.

It is also worth noting that the percentage of "original" labels and the corresponding relevance scores often appear very close in value. This is due to the fact that when the

teacher model classifies a response as "original," it typically does so with high confidence. In such cases, the model assigns a softmax probability close to 1.0 to the "original" class, which directly impacts both the classification outcome and the relevance score. As a result, the relevance score—although continuous—closely mirrors the discrete classification percentage, especially in high-confidence predictions.

V. Discussion

Our results demonstrate the effectiveness of the teacher-oriented framework in differentiating between dialogue-specialized and general-purpose LLMs. BlenderBot and DialoGPT outperformed others, highlighting the importance of dialogue-focused fine-tuning over model size alone. GPT-Neo and Mistral, despite their larger parameter counts and strong general NLP performance, scored lower due to the lack of task-specific tuning. A core strength of this framework is its ability to compare LLMs on a shared task using interpretable, consistent metrics. However, it does not serve as an absolute measure of a single model's quality and can be misled by adversarial or atypical responses, particularly due to reliance on synthetic training labels. Unexpectedly, Mistral underperformed despite its scale and architecture, emphasizing that large size does not guarantee conversational skill.

The observed decline in performance as the number of training epochs increased further suggests that the model may have begun to overfit to the training distribution. When overfitting occurs, the model captures spurious patterns or noise specific to the training data, which reduces its ability to generalize to unseen inputs. This phenomenon often results in rising training accuracy but diminishing evaluation performance after a certain point. In addition, reliance on synthetic labels may amplify this effect, as prolonged training could reinforce misleading associations. These observations highlight the importance of early stopping, careful monitoring of validation trends, and potentially employing regularization or more diverse training data to mitigate overfitting risks.

Our results indicate that dialogue-oriented models such as DialoGPT and BlenderBot outperform larger, general-purpose LLMs on conversational tasks, suggesting that model size alone is not sufficient for strong dialogue performance. Instead, task-specific fine-tuning and domain adaptation appear to play a critical role. The proposed teacher-oriented evaluation framework highlights these differences across models using consistent and interpretable metrics, providing a unified approach for comparing conversational abilities.

VI. Conclusion and Future Work

We introduced a scalable, teacher-based evaluation method for benchmarking dialogue performance in LLMs. The framework provides interpretable insights into how well models handle open-domain conversation, supporting model comparison without the need for human annotation. The evaluation confirmed that conversational alignment, not parameter count, is key to dialogue performance. Best results were obtained when the teacher model was fine-tuned for 4–6 epochs. Future work includes extending the framework to multi-turn and multilingual dialogues, incorporating human evaluation for alignment, and experimenting with ensemble classifiers to enhance robustness.

References

[1] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics. https://aclanthology.org/P17-1103/

[2] Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren, and Zhaochun Ren. 2023. RADE: Reference-Assisted Dialogue Evaluation for Open-Domain Dialogue. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12856–12875, Toronto, Canada. Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.719/

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Jul. 2002, pp. 311–318.

[4] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. https://aclanthology.org/W04-1013/

[5] E. Kamalloo, N. Dziri, C. L. A. Clarke, and D. Rafiei, "Evaluating Open-Domain Question Answering in the Era of Large Language Models," *arXiv preprint arXiv:2305.06984*, 2023. [Online]. Available: https://arxiv.org/abs/2305.06984

[6] C. Zhang, L. F. D'Haro, T. Friedrichs, and H. Li, "MDD-Eval: Self-Training on Augmented Data for Multi-Domain Dialogue Evaluation," *arXiv preprint arXiv:2112.07194*, 2022. [Online]. Available: https://arxiv.org/abs/2112.07194

[7] S. Mehri and M. Eskenazi, "USR: An Unsupervised and Reference-Free Evaluation Metric for Dialog Generation," *arXiv preprint arXiv:2005.00456*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00456

[8] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 89, 722–729.

[9] L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang, "GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 9230–9240. [Online]. Available: https://aclanthology.org/2020.emnlp-main.742/

[10] S. Mehri and M. Eskenazi, "Unsupervised Evaluation of Interactive Dialog with DialoGPT," in *Proc. of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 1st virtual meeting, Jul. 2020, pp. 225–235. [Online]. Available: https://aclanthology.org/2020.sigdial-1.28/

[11] C. Zhang, L. F. D'Haro, Q. Zhang, T. Friedrichs, and H. Li, "FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation," *arXiv preprint arXiv:2210.13832*, 2022. [Online]. Available: https://arxiv.org/abs/2210.13832

[12] K. Zhao, B. Yang, C. Tang, C. Lin, and L. Zhan, "SLIDE: A Framework Integrating Small and Large Language Models for Open-Domain Dialogues Evaluation," *arXiv preprint arXiv:2405.15924*, 2024. [Online]. Available: https://arxiv.org/abs/2405.15924

[13] C. Zhang, L. F. D'Haro, Y. Chen, M. Zhang, and H. Li, "A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators," *arXiv preprint arXiv:2312.15407*, 2023. [Online]. Available: https://arxiv.org/abs/2312.15407

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2020. [Online]. Available: https://arxiv.org/abs/1910.01108

[15] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 270–278, Online. Association for Computational Linguistics. https://aclanthology.org/2020.acl-demos.30/

[16] K. Shuster, "BlenderBot 3: A Deployed Conversational Agent That Continually Learns to Responsibly Engage," *arXiv preprint arXiv:2208.03188*, 2022. [Online]. Available: https://arxiv.org/abs/2208.03188

[17] A. Q. Jiang, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023. [Online]. Available: https://arxiv.org/abs/2310.06825