Supervised Fine-Tuning with an Empathic Dataset: A Study on LLAMA and GEMMA

Matheus Muniz Damasco
Department of Computational Modeling
Federal University of Juiz de Fora
Juiz de Fora, Brazil
matheus.damasco@estudante.ufjf.br

***Abstract*—This study investigates the potential of large language models (LLMs), such as LLAMA and GEMMA, to improve their base models through SFT (Supervised Fine-Tuning) using an empathy-focused dataset. Empathy is a crucial but challenging aspect for AI dialogue systems due to the scarcity of suitable training datasets. To address this, we applied SFT (Supervised Fine-Tuning) using the Huggingface TRL library, fine-tuning the 'llama-3-8b-bnb-4bit' and 'gemma-7b-bnb-4bit' models from the Unsloth library on the EmpatheticDialogues dataset. Our findings demonstrate that fine-tuning with empathy-focused datasets significantly improves the models' performance as measured by METEOR and BERTScore, highlighting the potential of task-specific fine-tuning to enhance model adaptability.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Empathy is commonly defined as the ability to understand and share another individual's state of mind or emotions. With the growing integration of chatbots across various domains, such as children seeking help with homework, individuals consulting for medical advice, or people using chatbots for companionship, the role of empathy in human-computer interaction has become increasingly significant [1].

Over recent years, chatbots have emerged as a prominent tool in daily life [2]. These systems are capable of simulating human-like conversations [3], offering practical assistance [4], delivering information [5], and providing emotional support [6]. Empathic responses are particularly important for dialogue systems designed for general or informal conversations, as everyday communication is often driven by people sharing their emotions or personal circumstances [7].

Given this, I chose to apply fine-tuning techniques using empathetic data. Specifically, I used the Empathic Dialogues dataset [8] available on Kaggle [1]. Fine-tuning data is defined as the data applied during the fine-tuning phase of large models, the scope of which depends on the delimitation of the fine-tuning. We were inspired by the fine-tuning steps described in OpenAI's InstructGPT paper [9], covering SFT (Supervised Fine-Tuning). [10]. Due to limited resources, I utilized Google Colab Pro with an NVIDIA T4 GPU along with the Unsloth library [11], which offers several models and methods, including LoRA (Low-Rank Adaptation) [12] and QLoRA (Quantized LoRA) [13], making fine-tuning large language models more practical, simple and cost-effective.

[1]https://www.kaggle.com/datasets/atharvjairath/empathetic-dialogues-facebook-ai

II. Background

A. Empathy in AI

What is empathy, and how can it improve AI systems? Empathy is the ability to understand and relate to emotions and experiences of others and to effectively communicate that understanding [14]. Empathy, widely seen as an essential human trait, signifies one's response to another's experiences, encapsulating empathetic concern and understanding another's perspective [15]. Empathetic support is one of the critical factors (along with ability to listen, concrete problem solving, motivational interviewing, etc.) that contributes to successful conversations in mental health support, showing strong correlations with symptom improvement [16] and the formation of alliance and rapport [10], [14], [17], [18]. In addition to mental health support, empathy can improve AI systems in various other fields. In customer service, for instance, empathetic AI systems can recognize frustration and offer personalized solutions, enhancing customer satisfaction [15], [19]. In education, AI that can recognize and respond to students' emotional states can help foster better learning environments, increasing engagement and motivation [20]. In healthcare, AI-assisted diagnostic tools that incorporate empathy can improve patient communication, making interactions feel more supportive and enhancing patient compliance [21]. Empathetic AI systems are increasingly being developed to facilitate more human-like, emotionally intelligent interactions across a variety of industries, creating more meaningful and productive engagements. Given the potential of LLMs to enhance empathetic interactions, the next step is to explore how these models are fine-tuned to handle such tasks effectively.

B. Large Language Model (LLM)

Large Language Models (LLMs) represent a remarkable advancement in NLP (Natural Language Processing) and artificial intelligence research [22]. These models have significantly enhanced the capabilities of machines to understand and generate human-like language [23]. Moreover, LLMs are new and essential part of computerized language processing, having the ability to understand complex verbal patterns and generate coherent and appropriate replies in a given context. Though

this success of LLMs has prompted a substantial increase in research contributions, rapid growth has made it difficult to understand the overall impact of these improvements [24]. Despite their impressive capabilities, LLMs are not without challenges. They face significant issues that may undermine the trust and reliability of their applications. One of the most notable safety concerns is hallucination [25], where the model generates plausible but incorrect or nonsensical information. This issue can lead to the dissemination of false information, which poses a particular risk in critical fields such as healthcare or legal advice. Moreover, the growing interest in LLMs has led to substantial investments from both private and public sectors. Major technology companies, including OpenAI, Google, Microsoft and others are heavily investing in the development and refinement of LLMs, recognizing their potential to revolutionize industries ranging from healthcare to finance. These investments are shaping the future of AI, with expectations that LLMs will continue to transform industries and create new opportunities across diverse sectors. Below we can see a chronological timeline showcasing the evolution of Large Language Models (LLMs):

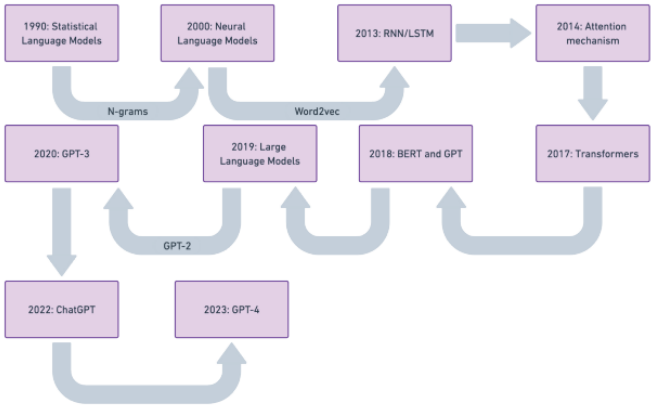

Fig. 1. A chronological timeline showcasing the evolution of Large Language Models (LLMs) from 1990 to 2023. This progression begins with early statistical models such as N-grams, transitions through neural language models like Word2Vec and RNN/LSTM, and advances into the era of pre-trained models with the introduction of transformers and attention mechanisms. The figure highlights significant milestones, including the development of BERT, GPT series, and recent innovations such as GPT-4 and ChatGPT, demonstrating the rapid advancements in LLM technology over time [26]

C. Fine-Tuning

Fine-tuning uses a pre-trained model, such as OpenAI's GPT series, as a foundation. The process involves further training on a smaller, domain-specific dataset. This approach builds upon the model's pre-existing knowledge, enhancing performance on specific tasks with reduced data and computational requirements. Fine-tuning transfers the pre-trained model's learned patterns and features to new tasks, improving performance and reducing training data needs. It has become popular in NLP for tasks like text classification, sentiment analysis, and question-answering [27]. Although SFT (Supervised Fine-Tuning) has emerged as an essential technique to align large language models with humans, it is considered superficial, with style learning being its nature [28]. SFT involves providing the LLM with labelled data tailored to the target task. For example, fine-tuning an LLM for text classification in a business context uses a dataset of text snippets with class labels. While effective, this method requires substantial labelled data, which can be costly and time-consuming to obtain [27]. "Although fine-tuning requires fewer data than training the LLM model from scratch, it remains a costly process, particularly for individuals with limited resources. This is mainly due to the high cost of GPUs, which are often essential for the task but not always readily available. For this reason, I utilized Google Colab Pro with an NVIDIA T4 GPU, which is a more affordable option. To further reduce costs and make fine-tuning feasible, I employed the Unsloth library [11], available on GitHub [2]. Unsloth accelerates the fine-tuning process and reduces memory usage. It utilizes the Hugging Face TRL [3] for model fine-tuning and incorporates several techniques, such as 4-bit and 16-bit QLoRA/LoRA, which help further reduce the computational costs of fine-tuning. Specifically, I employed the LoRA (Low-Rank Adaptation) [12] technique available in Unsloth to achieve additional cost reductions in the fine-tuning process.

III. Materials and Methods

A. Database

One challenge for dialogue agents is recognizing feelings in the conversation partner and replying accordingly, a key communicative skill. While it is straightforward for humans to recognize and acknowledge others' feelings in a conversation, this is a significant challenge for AI systems due to the paucity of suitable publicly-available datasets for training and evaluation. The dataset I chose is the EmpatheticDialogues, available on Kaggle[4], which consists of 25k conversations grounded in emotional situations. These conversations are based on a wide range of emotions, with participants selecting one of 32 emotion labels to describe the situation. Each conversation is composed of an exchange between two participants: the speaker describes the emotional situation and the listener responds based on cues from the conversation, without prior knowledge of the emotion label or the situation description. Previous experiments indicate that models trained on this dataset are perceived as more empathetic by human evaluators, compared to models trained solely on large-scale Internet conversation data. The dataset ensures a balanced distribution of emotions, including less frequent emotions, which helps make the training process more robust and diverse. The 32 emotion labels considered are listed in the paper [8].

B. Large Language Models (LLM)

Both models used in the project are available in the Unsloth [11]. In this library, you can query multiple notebooks, add

[2]https://github.com/unslothai/unsloth
[3]https://huggingface.co/docs/trl/sft_trainer
[4]https://www.kaggle.com/datasets/atharvjairath/empathetic-dialogues-facebook-ai

your dataset, and receive a fine-tuned model that is twice as fast as traditional methods. These fine-tuned models can be exported to formats such as GGUF, Ollama, vLLM, or uploaded to Hugging Face for later use and deployment. This makes the fine-tuning process more accessible and efficient for various applications. The `Llama-3-8b-bnb-4bit` and `Gemma-7b-bnb-4bit` are models 4-bit quantized versions, optimized by Unsloth for efficient inference and reduced memory usage. This models represents a significant advancement in making large language models more accessible and resource-efficient. Utilizing `bitsandbytes` [5] quantization, this model achieves 4-bit precision, resulting in approximately 70% reduced memory usage compared to the original version. The model also supports multiple tensor precisions, including F32, BF16, and U8, allowing for flexible deployment across various hardware setups.

1) Llama-3-8b-bnb-4bit: Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pre-trained and instruction-tuned generative text models available in 8B and 70B parameter sizes. The Llama 3 instruction-tuned models are optimized for dialogue use cases and have demonstrated superior performance compared to many open-source chat models across industry-standard benchmarks. Special attention was given during development to enhance both helpfulness and safety.

Model Architecture: Architecture in Llama 3 has a relatively standard decoder-only transformer architecture. Compared to Llama 2, we made several key improvements. Llama 3 uses a tokenizer with a vocabulary of 128K tokens that encodes language much more efficiently, which leads to substantially improved model performance. To improve the inference efficiency of Llama 3 models, we've adopted grouped query attention (GQA) across both the 8B and 70B sizes. We trained the models on sequences of 8,192 tokens, using a mask to ensure self-attention does not cross document boundaries [29].

2) Gemma-7b-bnb-4bit: The `Gemma-7b-bnb-4bit` model is another key language model optimized by Unsloth. Gemma is an open weight LLM. It comes in both instruction-tuned and raw, pretrained variants at various parameter sizes. It is based on the LLM architecture introduced by Google Research in the Attention Is All You Need paper [30]. Its primary function is to generate text token by token, based on a prompt provided by a user. In tasks like translation, Gemma takes a sentence from one language as input and outputs its equivalent in another language.

Model Architecture: The Gemma model architecture is based on the transformer decoder [30]. You can consult the core parameters of the architecture in the paper [31]. The model is trained on a context length of 8192 tokens. Several improvements proposed after the original transformer article were also used and are listed below:

- **Multi-Query Attention** [32]: Notably, the 7B model uses multi-head attention, based on ablations that showed that multi-query attention works well at small scales.
- **RoPE Embeddings** [33]: Rather than using absolute positional embeddings, we use rotary positional embeddings in each layer; we also share embeddings across our inputs and outputs to reduce model size.
- **GeGLU Activations** [34]: The standard ReLU nonlinearity is replaced by the approximated version of the GeGLU activation function.
- **RMSNorm** [35]: We normalize the input of each transformer sub-layer, the attention layer and the feedforward layer, with RMSNorm (Zhang and Sennrich, 2019) to stabilize the training.

[5]https://github.com/bitsandbytes-foundation/bitsandbytes

C. ChatML

ChatML[6] is a format designed for structuring dialogues in a way that facilitates interaction with large language models (LLMs). It organizes the conversation flow by wrapping the exchange of messages between users and the system within special tokens, helping the model understand the roles of the participants and the context of the conversation. In the context of this project, we applied a ChatML-based format to the EmpatheticDialogues dataset, which contains conversations grounded in emotional situations. The conversation data is structured using special tokens to demarcate the start and end of each interaction. Each conversation is framed between the `<|im_start|>` and `<|im_end|>` tokens, with the model receiving input in this structured format. The key advantage of ChatML is its ability to clearly define the roles of the participants in a conversation (e.g., the system, user) and provide the necessary context for each message, such as the emotional tone of the conversation. This structure allows LLMs to respond more accurately and empathetically by maintaining an understanding of the conversation flow and the emotional context.

D. Supervised Fine-Tuning (SFT) Parameters

The Supervised Fine-Tuning (SFT) architecture for both models (Llama-3-8b-bnb-4bit and Gemma-7b-bnb-4bit) is configured with the following parameters:

- Epochs: 3
- Batch size: 2
- Gradient Accumulation: 4
- Learning Rate: 2e-4
- FP16: True
- Optimizer: AdamW-8bit
- Weight Decay: 0.01
- LR Scheduler Type: Linear
- Gradient Checkpointing: unsloth
- LoRA (Low-Rank Adaptation): True
- Seed: 42
- GPU: T4

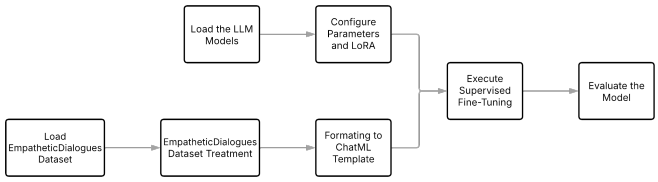

Fig. 2. Flowchart explanation of the project

E. Project Flowchart

IV. Results

This section presents a comprehensive analysis of the METEOR and BERTScore results for both base and fine-tuned models. The goal is to understand how fine-tuning has impacted the performance of the models and to discuss the differences between the models.

A. METEOR

METEOR, an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings; furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference [36].

As shown in Tables I and II below, the METEOR scores for both models improve significantly after fine-tuning. `Llama-3-8b-bnb-4bit` improves from 0.30 to 0.51, while `Gemma-7b-bnb-4bit` improves from 0.48 to 0.54. This indicates that fine-tuning enhances the alignment between machine-generated outputs and human references. Despite both models improving, Gemma maintains a slight performance advantage.

Metric	Llama-3-8b-bnb-4bit	Gemma-7b-bnb-4bit
METEOR	0.30	0.48

TABLE I
METEOR SCORES FOR BASE MODELS: PERFORMANCE OF LLAMA-3-8B-BNB-4BIT AND GEMMA-7B-BNB-4BIT BEFORE FINE-TUNING.

Metric	Llama-3-8b-bnb-4bit	Gemma-7b-bnb-4bit
METEOR	0.51	0.54

TABLE II
METEOR SCORES FOR FINE-TUNED MODELS: PERFORMANCE OF LLAMA-3-8B-BNB-4BIT AND GEMMA-7B-BNB-4BIT AFTER FINE-TUNING.

[6]https://learn.microsoft.com/pt-br/azure/ai-services/openai/how-to/chat-markup-language

B. BERTScore

BERTScore leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Moreover, BERTScore computes precision, recall, and F1 measure, which can be useful for evaluating different language generation tasks [37]. For improved correlation with human scores, we chose to use '`microsoft/deberta-xlarge-mnli`' model type because we want the scores to better correlate with human scores.

As shown in Tables III and IV below, the BERTScore metrics reveal important differences between the models in terms of precision, recall, and F1-score. For `Llama-3-8b-bnb-4bit`, fine-tuning results in a slight improvement in recall, but we end up losing in precision or F1-score. In contrast, `Gemma-7b-bnb-4bit` shows improvements in precision and F1-score, with a small decline in recall.

BERTScore	Llama-3-8b-bnb-4bit	Gemma-7b-bnb-4bit
Precision	[0.425, 0.384]	[0.272, 0.255]
Recall	[0.806, 0.649]	[0.835, 0.732]
F1-Score	[0.556, 0.482]	[0.411, 0.378]
Model Type	microsoft/deberta-xlarge-mnli	

TABLE III
BERTSCORE FOR BASE MODELS: PERFORMANCE OF LLAMA-3-8B-BNB-4BIT AND GEMMA-7B-BNB-4BIT BEFORE FINE-TUNING.

BERTScore	Llama-3-8b-bnb-4bit	Gemma-7b-bnb-4bit
Precision	[0.320, 0.294]	[0.372, 0.326]
Recall	[0.815, 0.730]	[0.811, 0.732]
F1-Score	[0.460, 0.419]	[0.510, 0.451]
Model Type	microsoft/deberta-xlarge-mnli	

TABLE IV
BERTSCORE FOR FINE-TUNED MODELS: PERFORMANCE OF LLAMA-3-8B-BNB-4BIT AND GEMMA-7B-BNB-4BIT AFTER FINE-TUNING.

V. Conclusions and Future Work

A. Conclusions

This study explored the application of Supervised Fine-Tuning (SFT) to enhance the performance of large language models (LLMs) such as `Llama-3-8b-bnb-4bit` and `Gemma-7b-bnb-4bit`, utilizing a dataset focused on empathy. The models used in this project are available on the HuggingFace profile [7] and in the Google Colab notebooks available on GitHub [8].

The results for the METEOR and BERTScore metrics demonstrate that fine-tuning significantly improved both base models. Specifically, the METEOR scores highlight substantial improvements, with `Llama-3-8b-bnb-4bit` increasing from 0.30 to 0.51 and `Gemma-7b-bnb-4bit` from 0.48 to 0.54. For the BERTScore metrics, the effects of fine-tuning were mixed: while `Gemma-7b-bnb-4bit` showed improvements in precision and F1-score, it experienced a slight

[7]https://huggingface.co/MathMuniz
[8]https://github.com/math-muniz

decrease in recall. In contrast, `Llama-3-8b-bnb-4bit` exhibited a small improvement in recall but experienced losses in precision and F1-score. This suggests that fine-tuning benefits Gemma more effectively across both metrics, while Llama shows substantial gains in METEOR but declines in BERTScore.

The SFT (Supervised Fine-Tuning) implemented in this study using the EmpatheticDialogues dataset proved to be an effective approach for improving LLM base models. The use of the Unsloth library, combined with techniques such as LoRA (Low-Rank Adaptation), allowed for significant reductions in memory usage, making these models more resource-efficient and accessible for multiple applications, even in resource-constrained environments.

B. Future Work

In future work, I plan to evaluate the models using empathy-focused benchmarks such as GIEBENCH [38]. Additionally, if resources become available, I plan to train larger models such as Llama 3.3 (70B) and other models to further investigate how these models perform when fine-tuned using empathy-focused datasets.

REFERENCES

[1] K. Schaaff, C. Reinig, and T. Schlippe, "Exploring chatgpt's empathic abilities," *IU International University of Applied Sciences*, 2024, email: kristina.schaaff@iu.org; caroline.reinig@gmail.com; tim.schlippe@iu.org.

[2] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior*, vol. 122, p. 106855, 2021.

[3] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *ArXiv Preprint*, vol. ArXiv:2001.09977, 2020.

[4] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a campus environment: Design of lisa, a virtual assistant to help students in their university life," in *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III*. Springer, 2018, pp. 103–116.

[5] D. Arteaga, J. Arenas, F. Paz, M. Tupia, and M. Bruzza, "Design of information system architecture for the recommendation of tourist sites in the city of manta, ecuador through a chatbot," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2019, pp. 1–6.

[6] C. Falala-Séchet, L. Antoine, I. Thiriez, and C. Bungener, "Owlie: A chatbot that provides emotional support for coping with psychological difficulties," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 236–237.

[7] Y. Chen, H. Wang, S. Yan, S. Liu, Y. Li, Y. Zhao, and Y. Xiao, "Emotionqueen: A benchmark for evaluating empathy of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2409.13359

[8] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: a new benchmark and dataset," 2019. [Online]. Available: https://arxiv.org/abs/1811.00207

[9] 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*, vol. 35, pp. 27 730–27 744, 2022.

[10] R. Ma, W. Li, and F. Shang, "Investigating public fine-tuning datasets: A complex review of current practices from a construction perspective," *arXiv preprint arXiv:2407.08475*, 2024. [Online]. Available: https://arxiv.org/abs/2407.08475

[11] M. H. Daniel Han and U. team, "Unsloth," 2023. [Online]. Available: http://github.com/unslothai/unsloth

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[13] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[14] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy," *Psychotherapy*, vol. 48, pp. 43–49, 2011.

[15] Y. Liu-Thompkins, S. Okazaki, and H. Li, "Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience," *Journal of the Academy of Marketing Science*, vol. 50, pp. 1198–1218, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s11747-022-00892-5

[16] R. Elliott, A. C. Bohart, J. C. Watson, and D. Murphy, "Therapist empathy and client outcome: An updated meta-analysis," *Psychotherapy*, vol. 55, pp. 399–410, 2018.

[17] A. C. Bohart, R. Elliott, L. S. Greenberg, and J. C. Watson, *Empathy*. New York, NY, US: Oxford University Press, 2002, vol. 452, pp. 89–108.

[18] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[19] M. Bastani and J. Smith, "Empathetic ai in customer service: Enhancing customer satisfaction through emotional recognition," *Journal of Business and Technology*, vol. 58, pp. 221–234, 2022.

[20] M. Rodrigues and R. Pereira, "Empathetic ai in education: Enhancing student engagement and motivation," *Educational Technology and Society*, vol. 24, pp. 12–24, 2021.

[21] J. Mayer and A. Lewis, "The role of empathy in ai-assisted healthcare communication," *Journal of Health Communication*, vol. 25, pp. 1010–1023, 2020.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022. [Online]. Available: https://arxiv.org/abs/2212.10403

[24] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26 839–26 874, 2024.

[25] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, Nov. 2024, just Accepted. [Online]. Available: https://doi.org/10.1145/3703155

[26] Z. Wang, Z. Chu, T. Doan, S. Ni, M. Yang, and W. Zhang, "History, development, and principles of large language models: an introductory survey," *AI and Ethics*, pp. 1–17, 10 2024.

[27] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," 2024. [Online]. Available: https://arxiv.org/abs/2408.13296

[28] M. Shen, "Rethinking data selection for supervised fine-tuning," 2024. [Online]. Available: https://arxiv.org/abs/2402.06094

[29] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton,

J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[31] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024. [Online]. Available: https://arxiv.org/abs/2403.08295

[32] N. Shazeer, "Multi-query attention," *arXiv preprint arXiv:1905.09309*, 2019. [Online]. Available: https://arxiv.org/abs/1905.09309

[33] Y. Su, L. Gu, J. Wang, Y. Chen, F. Wei, and W. Lu, "Rope: Rotary position embedding for efficient transformers," *arXiv preprint arXiv:2104.09864*, 2021. [Online]. Available: https://arxiv.org/abs/2104.09864

[34] N. Shazeer, "Geglu: A novel activation function for transformers," *arXiv preprint arXiv:2002.05202*, 2020. [Online]. Available: https://arxiv.org/abs/2002.05202

[35] W. Zhang and R. Sennrich, "Rmsnorm: A simple and efficient normalization method for transformer models," *arXiv preprint arXiv:1910.03751*, 2019. [Online]. Available: https://arxiv.org/abs/1910.03751

[36] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[37] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[38] L. Wang, Y. Jin, T. Shen, T. Zheng, X. Du, C. Zhang, W. Huang, J. Liu, S. Wang, G. Zhang, L. Xiang, and Z. He, "Giebench: Towards holistic evaluation of group identity-based empathy for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.14903