Exploring the Impact of Temperature on Large Language Models: A Case Study for Classification Task based on Word Sense Disambiguation

Deshan Sumanathilaka
Department of Computer Science
Swansea University
Swansea, United Kingdom
deshankoshala@gmail.com

Nicholas Micallef
Department of Computer Science
Swansea University
Swansea, United Kingdom
nicholas.micallef@swansea.ac.uk

Julian Hough
Department of Computer Science
Swansea University
Swansea, United Kingdom
julian.hough@swansea.ac.uk

***Abstract*—With the advent of Large Language Models (LLMs), Natural Language (NL) related tasks have been evaluated and explored. While the impact of temperature on text generation in LLMs has been explored, its influence on classification tasks remains unexamined despite temperature being a key parameter for controlling response randomness and creativity. In this study, we investigated the effect of the model's temperature on sense classification tasks for Word Sense Disambiguation (WSD). A carefully crafted Few-shot Chain of Thought (COT) prompt was used to conduct the study, and FEWS lexical knowledge was shared for the gloss identification task. GPT-3.5 and 4, LlaMa-3-70B and 3.1-70B, and Mixtral 8x22B have been used as the base models for the study, while evaluations are conducted with 0.2 intervals between the 0 to 1 range. The results demonstrate that temperature significantly affects the performance of LLMs in classification tasks, emphasizing the importance of conducting a preliminary study to select the optimal temperature for a task. The results show that GPT-3.5-Turbo and Llama-3.1-70B models have a clear performance shift, the Mixtral 8x22B model with minor deviations, while GPT-4-Turbo and LlaMa-3-70B models produce consistent results at different temperatures.**

***Index Terms*—Large Language Models, Word Sense Disambiguation, Temperature Parameter, Few-shot Prompting, Classification Tasks**

I. Introduction

With recent advancements in language models, Large Language Models (LLMs) are increasingly recognized as transformative technologies for Natural Language Processing (NLP) tasks. Their well-documented capabilities, including translation, text generation, and question answering, have led to their widespread commercial adoption across various industries, such as finance and healthcare [1]. Researchers have been exploring different computational techniques to fit LLMs into a specific domain by prompt tuning, prompt augmentation, and using different fine-tuning techniques [2]. However, to our knowledge, the optimal hyperparameters that need to be pre-assigned before the inference process remain unexplored, and the behaviour of LLM's responses based on these crucial parameters has not yet been examined.

The temperature is considered one of the critical factors that influence the randomness and creativity of the response generation by an LLM [3]. It adjusts the probability distribution over the possible output, making output more creative and random. The model tends to be more deterministic in a low temperature (T) setup (T close to 0), ensuring low response variability leads to coherent outputs. However, in a high temperature (T > 1) setup, the likelihood of selecting less probable outcomes is increased by generating less coherent, varied responses. In a practical environment, the temperature can be kept at zero (0) to increase the reproducibility of the work, as it always generates the same response if the context is the same [4]. The models behave as designed for average temperatures (0.5 to 1), balancing creativity with coherence.

Mathematically, the T variable in the softmax function describes the temperature behaviours in stochastic models. The Equation 1 indicates the softmax function with temperature (T) parameter.

$$\text{softmax}(z_i, T) = \frac{e^{z_i/T}}{\sum_{j=1}^{n} e^{z_j/T}} \tag{1}$$

To evaluate the behaviour of the temperature factor in a classification task such as WSD [5] in a multi-class setup, we have assessed the responses of different models on six temperature intervals [4]. The contribution of this study can be highlighted as follows.

- Evaluating flagship Large Language Models' behaviour on temperature changes in a *classification task* with few-shot chain thought prompting.
- Showing that GPT 3.5 and Llama 3.1 results are varied with temperature values, while GPT 4 and Llama 3 remain consistent for the WSD task.
- Demonstrating that temperatures above 1 disrupt the precision needed for multi-class classification tasks of WSD.

The following sections will describe similar studies, explain our chosen methodology, present our findings, and discuss the limitations of the proposed methods. We conclude by suggesting areas for potential future research.

II. Related Works

In recent years, with the introduction of commercial and open-source LLMs, various text-based related tasks have been performed using language models due to their capabilities in creative text generation. Various functions such as chatbot response generation, information retrieval, summarization, and many more tasks have been explored and evaluated [1]. The banking industry, education, and business industries have employed different ranges of LLMs to perform various classification tasks [6]–[8]. However, the effect of hyperparameters on response generation has barely been explored. Our findings reveal that the impact of temperature factors on multi-class classification tasks has neither been thoroughly investigated nor evaluated. To our knowledge, this is the first study dedicated to this topic. A few recent works associated with the domain are summarized below.

Patel has explored the effects of temperature on LLMs across clinical tasks with the primary intention of evaluating text generation tasks [3]. This work assesses three LLMs for different temperature settings to predict the mortality, length of stay and medical coding. They have observed that the temperature factor does not directly affect the classification and clinical reasoning tasks, gaining stable accuracies. Rense and the team have conducted the multiple choice question answering exam with five prompt engineering techniques with multiple temperature setups [4]. The sampling temperature has been evaluated between 0 and 1, showing no statistically significant impact on performance for problem-solving tasks. The study conducted by Wang proposes a contextual temperature for the training language model, which learns an optimal temperature for each class. This approach has shown that temperature schedules change by the vocabulary, and uncertainties can be controlled using proper temperature scheduling [9]. T2oT, proposed by Cai et al., enhances the reasoning and decision-making capabilities of LLLMs by dynamically adjusting search parameters like temperature through Temperature Tree prompting and Particle Swarm Optimization, demonstrating improved accuracy, solution diversity, and text quality in tasks like the Game of 24 and Creative Writing [10].

Zang's team proposed an effective entropy-based dynamic sampling method capable of dynamically selecting the temperature parameter at a decoding step to produce a balanced performance. This approach has been evaluated for text summarization, question answering, and machine translation and has shown promising results compared to the benchmark studies [11]. The study by Zue has introduced an adaptive temperature (AdapT), which dynamically suggests the temperature coefficient when decoding the different tokens for the code generation task. This has allowed the LLMs to use more significant temperatures for challenging token sampling while low temperatures generate confident tokens [12]. These studies have shown the importance of temperature in response generation tasks and the effect of temperature on regulating randomness and diversity outputs. While previous studies have examined the text generation process using various parameters, the performance of classification tasks, particularly concerning these parameters, remains underexplored. Our research addresses this gap by investigating the impact of temperature settings at different intervals on the performance of a multi-class classification task specifically applied to WSD. By conducting an empirical study, this research contributes to the literature by evaluating how temperature variations affect classification performance.

TABLE I
Input and output sequence.

Input Sentence:	The first author meticulously cross-checked the manuscript against various <WSD> dictionaries </WSD>, striving to ensure both word choice and proper usage.
Possible sense IDs	dictionary.noun.0, dictionary.noun.1, dictionary.noun.2, dictionary.noun.3
Output	dictionary. noun.0
Correct Gloss	A reference work with a list of words from one or more languages, normally ordered alphabetically, explaining each word's meaning, and sometimes containing information on its etymology, pronunciation, usage, translations, and other data.

III. Methodology

This study employs an empirical study on how temperature affects response generation in a multi-class classification problem concerning WSD task. Different glosses from FEWS senseNet have been shared with the model to predict the exact outcome suitable for sense [13]. A few shot chain-of-thought prompts are utilized for the study with Knowledge Base (KB) as a retriever to provide the few shots that drive the correct predictions. The 0.2 temperature intervals from 0 to 1 have been selected for this study, where the selected intervals provide a sufficiently fine-grained range to capture the changes in output while ensuring that the evaluation remains computationally manageable [14].

A. Dataset

This study utilized the publicly accessible FEWS dataset [13], which includes a list of sense tags and training and test data. We assessed each proposed method by examining how well the models could accurately assign sense tags to ambiguous words placed between <WSD> tokens in sentences. From the data parameters in lexical knowledge, we have incorporated senseID, word, gloss, and synonyms as major parameters to drive the inference process effectively with LLMs. TABLE I summarizes the input data and outputs in the inference process. The study was conducted using 1050 unique instances from the test set of the FEWS, where the 4:3:3 ratio was used for nouns, verbs, and adjectives. Fifty adverbs are considered to maintain diversity in all the part of speech (POS) tags. The data selection was performed to ensure that each record uniquely represented an ambiguous word, avoiding the repetition of the same sense tags in the testing data.

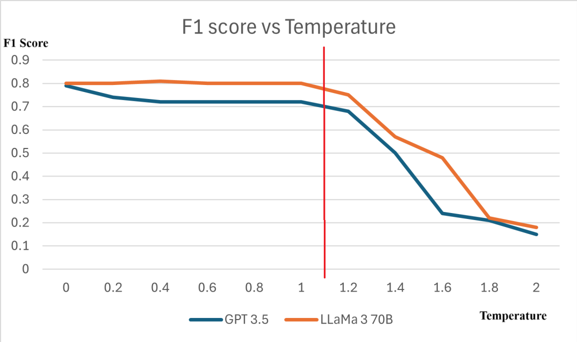

Fig. 1. F1 vs Temperature for primary study

B. Procedure

The study was mainly conducted using five commonly used LLMs of different vendors. Previous studies were reviewed to identify the most suitable candidates for the study who are known to perform well in classification tasks [15]. In our initial study, we assessed the behaviour of LLMs (GPT-3.5 Turbo and LLaMa 3 70B) across a temperature range from 0 to 2, with intervals of 0.2, to understand their performance patterns. The results revealed that temperatures above 1 led to a clear decline in performance, indicating that higher temperatures negatively impact the models' ability to perform the task accurately. This observation has been validated through additional experiments at higher temperatures, further supporting our decision to limit the temperature range for this study [4], [16]. According to Fig. 1, it is clear that performance drastically drops after 1. Therefore, we selected 0-1 as the range for evaluating other LLMs for the study's second phase.

The optimal prompt required for the evaluation was derived from the authors' previous studies, which continuously evaluated different prompting strategies [17]. The optimal prompt for the study consists of a few-shot COT prompt where the few shots required are retrieved from the KB and implemented using the FEWS training data. The training data has been arranged in a graph by keeping ambiguous words as the root node, the POS tag as a first-level parent and related examples with corresponding sense IDs in the leaf node. This graph-based structure provides the samples for in-context learning during the inference process of LLMs. The selection of graph-based KB was mainly motivated by its ability to extract data in a constant amount of time. The FEWS senseNet was used to extract the possible senses of the ambiguous word, along with their glosses and synonyms required for the classification task. NLTK python library filters the sense space by identifying the targeted senses based on the POS tag related to the ambiguous word. Filtered definitions from sense knowledge are appended to prompt the inference process. The evaluation flow can be found in Fig. 2, and the prompt for the study is presented in TABLE II.

TABLE II
PROMPT USED FOR THE STUDY

Prompting with Knowledge Base
You are going to identify the corresponding sense tag of an ambiguous word in English sentences. Do the following tasks. 1. {word} has different meanings. Below are possible meanings. Comprehend the sensetags and meanings. Synonyms are provided if available.{filtered_definitions} 2. You can learn more on the usage of each word and the meaning through below Examples. Examples are {examples}. 3. Now Examine the sentence below. You are going to identify the most suitable meaning for an ambiguous word. {sentence} 4. Analyze the sentence using the following techniques and identify the meaning of the ambiguous word. -Focus on keywords in the sentence surrounding the ambiguous word. -Think about the overall topic and intent of the sentence. Decide on the sense of the word that makes the most logical sense within the context. 5. Based on the identified meaning, try to find the most appropriate senseIDs from the above sense tag list. 6. If you have more than one senseIDs identified after above steps, you can return the senseIDs in order of confidence level. 7. Return JSON object that contains the ambiguity word and the finalized senseIDs.

For instance, consider the sentence: "The <WSD> bark </WSD> of the oak tree was rough to the touch." This sentence contains a polysemous word, "bark". Here, the word refers to "The exterior covering of the trunk and branches of a tree.". During the inference process,

- {Examples} holds few instances which are retrieved from the knowledge base, followed by the sense tag.
  - The dog's bark echoed through the quiet forest as it chased after the fox. bark.noun.0
  - The coach's bark of commands kept the players in line during practice. bark.noun.1
  - The bark of the birch tree is smooth and often used for decorative crafts. bark.noun.2
  - In the 18th century, Peruvian bark was widely used as a treatment for malaria. bark.noun.3
- {sentence} holds the sentence The <WSD> bark </WSD> of the oak tree was rough to the touch.
- {filtered_definitions} represent the meanings of each sense which has been filtered based on POS tag.
  - bark.noun.0: The short, loud, explosive sound uttered by a dog, a fox, and some other animals. Synonyms: [yelp, howl, growl]
  - bark.noun.1: An abrupt loud vocal utterance.
  - bark.noun.2: (tree) The exterior covering of the trunk and branches of a tree. Synonyms:[rind, husk, cortex (scientific), outer layer]
  - bark.noun.3: Peruvian bark or Jesuit's bark, the bark of the cinchona from which quinine is produced.

This example is selected to demonstrate the prompt. In the actual inference process, "bark" has 9 different glosses, which will be shared along with the prompt. We ensure that a maximum of 3 instances for each gloss, based on data retrieved from the knowledge base (KB), are provided to the model to guide the few-shot COT process.

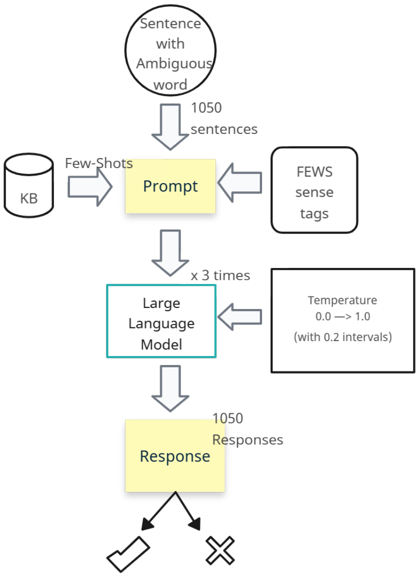

Fig. 2. Dataflow of the Evaluation process

C. Study Setup

In this study, our goal was to evaluate the performance of different temperatures as a hyperparameter. Given their recognised proficiency in semantics, we specifically targeted the GPT-3.5 Turbo, GPT-4, LlaMa-3-70B, LlaMa-3.1-70B and Mistral 8x22B [18]. We utilized an OpenAI API key associated with a tier-one OpenAI account to access OpenAI's models, along with Llama and Mistral models accessed via the Together.ai API. The models were configured using temperature settings ranging from 0 to 1, incremented in intervals of 0.2, to explore their capabilities systematically. Each temperature was simulated three times, except for temperatures at 0 (Produce deterministic response with highest probable output) and the average of the F1 score was reported. The maximum output limit is set up at 500 tokens, maintaining the primary task assigned to all large language models (LLMs) as word sense identification, with their role defined as a "helpful assistant for identifying word senses." Each model was evaluated for six distinct temperature values with 1050 instances. We then analyzed the models' performance by examining the number of correct predictions after the post-processing on the response to extract the sense IDs. The F1 scores were produced for each temperature interval.

IV. Results and Discussion

We conducted a preliminary study to identify the temperature range to evaluate in this study by generating F1 scores for the Llama-3-70B model and the GPT-3.5 Turbo model. While the differences in F1 scores between temperature intervals were modest for values between 0 and 1, a significant deviation in results was observed when the temperature exceeded 1, as shown in Fig. 1. Based on these findings, we concluded that temperatures above 1 are not suitable candidates for this study. Additionally, the results support the idea that high-temperature values (above 1) introduce too much uncertainty, making them unsuitable for classification tasks like WSD.

TABLE III
F1 score for different temperatures

Temperature	**0**	**0.2**	**0.4**	**0.6**	**0.8**	**1**
GPT-3.5 [19]	0.79	0.74	0.72	0.72	0.72	0.72
GPT-4 [20]	0.81	0.82	0.82	0.82	0.82	0.83
Llama-3-70B [21]	0.80	0.80	0.81	0.80	0.80	0.80
Llama-3.1-70B [21]	0.84	0.84	0.83	0.82	0.81	0.78
Mixtral-8x22B [22]	0.80	0.80	0.81	0.77	0.77	0.77

TABLE III shows the average F1 score of the model performance concerning different temperatures for 1050 instances on WSD with three-time simulations of each experiment. Results depict that the GPT-3.5-turbo significantly declined the score from temperature 0 to 0.2 but remained constant after 0.4 to 1. This conveys that the temperature value at 0 (almost deterministic state) produces better results than the other temperature values. The GPT-4-turbo model exhibits a marginal improvement in performance with an increase in temperature, achieving optimal results at a temperature value of 1. Llama-3.1 gradually decreased performance with the increase in temperature, showing the worst results with temperature 1. However, Llama-3 and Mixtral demonstrate consistent performance across the various temperature settings, indicating consistency in classification tasks regardless of the temperature value. This variation in performance across different LLMs underscores the importance of pre-studying the temperature factor in classification tasks. These findings open new avenues for further investigation into temperature as a hyperparameter, which can significantly impact final performance.

Fig. 3 presents the distribution of performance based on temperature, highlighting the key points of the performance thresholds for Llama-3 and 3.1 and Mixtral models.

However, previous studies on text generation tasks show that temperature does not significantly impact performance except for the creativity factor of the text generation process. In contrast, temperature plays a crucial role in classification tasks. This is because classification requires a sharper probability distribution to accurately identify the correct class, minimizing randomness in the decision-making process. Therefore, for optimal classification performance, selecting the appropriate temperature is essential.

V. Conclusion and Future Directions

This research investigates the impact of temperature on classification tasks, with a specific focus on WSD. The findings reveal that temperature significantly influences the performance of classification tasks in certain LLMs, such as GPT-3.5-Turbo while demonstrating consistency across others. These results highlight the critical importance of exploring and optimizing temperature settings to achieve optimal outcomes in linguistic tasks. The study employs a Few-shot COT prompt, extracted from prior work by the authors, as the foundational methodology. While the use of a publicly available

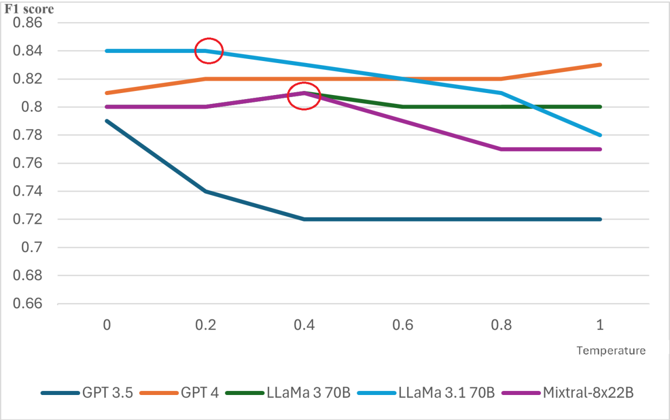

Fig. 3. F1 score vs temperature

dataset ensures accessibility and reproducibility, limitations inherent in the dataset creation process of the initial studies inevitably influence this research. Although these limitations do not directly affect the study's validity, they underscore the constraints of the prompt used. To address this, future studies should explore different prompt engineering techniques, which could provide a broader understanding of temperature's role in model performance. Furthermore, while this research focuses primarily on WSD as the classification task, future investigations could extend to other linguistic classification tasks, such as sentiment analysis and emotion classification. Such studies would not only deepen the understanding of temperature as a hyperparameter but also offer insights into its broader applicability across various tasks. Overall, the study underscores the importance of refining both temperature settings and prompt engineering techniques to enhance performance in classification tasks, paving the way for more robust applications of large language models in linguistic domains.

Acknowledgment

We would like to extend our sincere appreciation to the OpenAI Researcher Access Program for generously providing the credits that made the development of this project possible. Additionally, we are grateful to Swansea University for its financial backing and ongoing commitment to the success of this project.

References

[1] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, 2023.

[2] M. Abeysiriwardana and D. Sumanathilaka, "A survey on lexical ambiguity detection and word sense disambiguation," *arXiv preprint arXiv:2403.16129*, 2024.

[3] D. Patel, P. Timsina, G. Raut, R. Freeman, M. Levin, G. Nadkarni, B. S. Glicksberg, and E. Klang, "Exploring temperature effects on large language models across various clinical tasks," *medRxiv*, pp. 2024–07, 2024.

[4] M. Renze and E. Guven, "The effect of sampling temperature on problem solving in large language models," *arXiv preprint arXiv:2402.05201*, 2024.

[5] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent trends in word sense disambiguation: A survey," in *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, Inc, 2021, pp. 4330–4338.

[6] L. Loukas, I. Stogiannidis, O. Diamantopoulos, P. Malakasiotis, and S. Vassos, "Making llms worth every penny: Resource-limited text classification in banking," in *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 392–400.

[7] S. Al Faraby, A. Romadhony *et al.*, "Analysis of llms for educational question classification and generation," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100298, 2024.

[8] S. Gholamian, G. Romani, B. Rudnikowicz, and S. Skylaki, "Llm-based robust product classification in commerce and compliance," *arXiv preprint arXiv:2408.05874*, 2024.

[9] P.-H. Wang, S.-I. Hsieh, S.-C. Chang, Y.-T. Chen, J.-Y. Pan, W. Wei, and D.-C. Juan, "Contextual temperature for language modeling," *arXiv preprint arXiv:2012.13575*, 2020.

[10] C. Cai, X. Zhao, Y. Du, H. Liu, and L. Li, "ℓ2 of thoughts: Temperature tree elicits reasoning in large language models," *arXiv preprint arXiv:2405.14075*, 2024.

[11] S. Zhang, Y. Bao, and S. Huang, "Edt: Improving large language models' generation by entropy-based dynamic temperature sampling," *arXiv preprint arXiv:2403.14541*, 2024.

[12] Y. Zhu, J. Li, G. Li, Y. Zhao, Z. Jin, and H. Mei, "Hot or cold? adaptive temperature sampling for code generation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 437–445.

[13] T. Blevins, M. Joshi, and L. Zettlemoyer, "Fews: Large-scale, low-shot word sense disambiguation with the dictionary," *arXiv preprint arXiv:2102.07983*, 2021.

[14] F. Gloeckle, B. Roziere, A. Hayat, and G. Synnaeve, "Temperature-scaled large language models for lean proofstep prediction," in *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS*, vol. 23, 2023.

[15] T. Sumanathilaka, N. Micallef, and J. Hough, "Can llms assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation," *arXiv preprint arXiv:2411.18337*, 2024.

[16] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?" *arXiv preprint arXiv:2405.00492*, 2024.

[17] T. Sumanathilaka, N. Micallef, and J. Hough, "Assessing GPT's Potential for Word Sense Disambiguation: A Quantitative Evaluation on Prompt Engineering Techniques," in *Proceedings of 15th Control and System Graduate Research Colloquium (ICSGRC)*. Mardhiyyah Hotel & Suites, Shah Alam, Malaysia: IEEE, 2024.

[18] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen *et al.*, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.

[19] OpenAI. (2024) Gpt-3.5 turbo - openai api documentation. Accessed: 2024-11-17. [Online]. Available: https://platform.openai.com/docs/modelsgpt-3-5-turbo

[20] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.