NapierNLP at CheckThat! 2025: Detecting Subjectivity with LLMs and Model Fusion

Notebook for the CheckThat! Lab at CLEF 2025

Katarina Alexander[1,*], Md Zia Ullah[1] and Dimitra Gkatzia[1]

[1]School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Scotland

Abstract

This paper presents the participation of our team, NapierNLP, in the CLEF2025 CheckThat! Lab Task 1: Subjectivity. This task aimed to differentiate between subjective and objective sentences within a corpus sourced from news articles. We formulate the task as a classification problem and fine-tune three pre-trained large language models (LLMs): GPT2, Qwen, and GPT-Neo. To enhance our predictions, we combine the outputs from multiple models using a majority voting method. We conducted experiments and evaluated our approach using the CheckThat! 2025 Task 1 dataset. Our results showed that GPT-Neo outperforms the other two models. In the official competition results, we ranked 18th in the monolingual English category, but our combination method proved to be more effective than the individual models we employed.

Keywords

Subjectivity Detection, Natural Language Processing (NLP), Large Language Models (LLMs),

1. Introduction

Subjectivity detection is a difficult but important task for various applications, including fake news detection. Even human annotators can find it difficult to reach an agreement when classifying complex sentences, and the definition of 'subjective' can vary across domains. Therefore, developing a subjectivity classification system is both an interesting and significant challenge.

The CLEF-2025 CheckThat! Lab[1][2] Task 1[3][4] invites participants to develop a system that can "distinguish whether a sentence from a news article expresses the subjective view of the author behind it or presents an objective view on the covered topic instead". Datasets are available in Arabic, Bulgarian, English, German, and Italian. Subtask options were monolingual, multilingual, or zero-shot (tested on unseen languages). We participated in the monolingual English subtask only. We aimed to utilise open-source Large Language Models (LLMs) to detect subjective claims.

2. Related Work

Previous research on subjectivity detection primarily relied on rule-based templates and patterns that were bounded by the datasets used, leading to a lack of generalizability [5]. The introduction of deep-learning models enhanced this generalizability by utilising labelled data to learn models capable of recognising key features of subjective sentences. Team HYBRINFOX, which won the CLEF 2024 CheckThat! Task 2 on subjectivity detection, leveraged a BERT-based approach, specifically RoBERTa combined with VAGO, achieving a macro F1 score of 0.74 [6].

Large Language Models (LLMs) can be effectively used for language-based classification tasks [7], including subjectivity detection [8]. Shokri et al. experimented with GPT-3.5, GPT-4, and Gemini on three datasets: MPQA [9], News-1 [10], and News-2 [11]. They found that even a zero-shot approach

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

katarina.alexander@napier.ac.uk (K. Alexander); m.ullah@napier.ac.uk (M. Z. Ullah); d.gkatzia@napier.ac.uk (D. Gkatzia)

https://www.napier.ac.uk/people/katarina-alexander (K. Alexander); https://www.napier.ac.uk/people/md-zia-ullah (M. Z. Ullah); https://www.napier.ac.uk/people/dimitra-gkatzia (D. Gkatzia)

0000-0002-4022-7344 (M. Z. Ullah); 0000-0001-8568-7806 (D. Gkatzia)

performed well at the task, with an average of 0.71, 0.71, and 0.69, respectively across the datasets. LLMs were also used in the CheckThat! 2024 competition. Team SemanticCuetSync [12] used Llama-3-8b and achieved a macro F1 score of 0.72 in the Arabic test set and 0.50 in the English test set. Meanwhile, Team CLaC-2 [13] used Google's Gemini, alongside prompt engineering, to achieve a macro F1 of 0.45 on the English test set.

Using an ensemble model can yield better results than individual models [14]. Majority voting is a straightforward technique for classification problems, where individual models classify sentences independently, and the final prediction is based on the most frequently assigned label [15]. Shokri et al. [8] used this technique to enhance their model's performance, achieving a macro F1 of 0.74, an increase of 0.03 over their best individual model.

3. Methodology

The task of detecting whether a sentence is subjective or objective can be framed as a binary classification problem. By utilising human-annotated texts, a model can be developed using a supervised learning technique to classify sentences accordingly.

3.1. LLMs

LLMs are large-scaled, pre-trained neural language models that have emergent abilities that are not present in smaller models [16]. LLMs can then be fine-tuned for a variety of downstream tasks, including sentence classification.

In this study, we fine-tune three publicly available LLMs from HuggingFace: GPT2 [17], Qwen2.5 [18, 19], and GPT-Neo [20]. Details of the models can be found in Table 1.

Table 1
Details of the LLMs considered in the experiments and evaluation

Model	Parameters	Developer	Release Year	Available at:
GPT2	137M	OpenAI	2019	openai-community/gpt2
Qwen2.5 0.5B	494M	Alibaba Group	2024	Qwen/Qwen2.5-0.5B
GPT-Neo 1.3B	1.37B	EleutherAI	2021	EleutherAI/gpt-neo-1.3B

We anticipate that a fusion model could outperform individual models, so we employed majority voting as our ensemble technique. Each model provides a prediction label, and the common label across all three models is used as our final classification.

4. Experiments and Evaluations

In this section, we describe the datasets, evaluate our approach, and discuss the results.

4.1. Dataset

The English language corpus consisted of sentences sourced from news articles [21], along with an ID and a label: "OBJ" for objective and "SUBJ" for subjective. These had been annotated by humans based on the following criteria:

> "A sentence is subjective if its content is based on or influenced by personal feelings, tastes, or opinions. Otherwise, the sentence is objective." [22]

Table 2 provides a breakdown of all datasets. The training dataset and both testing datasets are imbalanced towards objective sentences. This is expected given the data source, as news articles are typically presented as an objective recounting of events. No augmentation was applied to the training set, as initial experiments demonstrated that the model performed well despite the imbalance.

Table 2
Statistics of the train, Dev, Dev-Test, and Test datasets

Dataset	Objective (%)	Subjective (%)	Total
Train	532 (64%)	298 (36%)	830
Dev	222 (48%)	240 (52%)	462
Dev-Test	362 (75%)	122 (25%)	484
Test	215 (72%)	85 (28%)	300

4.2. Experimental settings

We fine-tune the LLMs listed in Table 1 on the training data for 100 epochs using AutoModelForSequenceClassification and evaluated using the development set. Due to size constraints, both Qwen and GPT-Neo were trained using a parameter-efficient fine-tuning method called LoRA [23]. The training was conducted on an NVIDIA GeForce RTX 2080 Super with 8 GB GPU memory.

4.3. Results and Discussions

Performance on Development Test Set: Table 3 shows our results for each model, including the final combined model. The baseline was a logistic regressor trained on a multilingual SentenceBERT model, as provided by competition organisers. In the development test data, GPT2 achieved higher F1 scores for both macro (0.75) and subjective (0.62) metrics. However, the combined model demonstrated an equal accuracy of 0.81 and outperformed all other models for the objective class with an F1 of 0.88.

Table 3
Performance of models on the Dev-Test dataset showing F1, Precision, Recall and Accuracy.

Model	Macro			Subjective			Objective			Accuracy
	F1	P	R	F1	P	R	F1	P	R	
Baseline	0.63	0.63	0.63	0.45	0.45	0.45	0.81	0.81	0.81	0.72
GPT2	**0.75**	0.75	**0.74**	**0.62**	0.63	**0.61**	0.87	**0.87**	0.88	**0.81**
Qwen2.5	0.61	0.65	0.60	0.37	0.51	0.30	0.84	0.79	0.90	0.75
GPT-Neo	0.67	0.73	0.65	0.48	0.65	0.38	0.87	0.82	0.93	0.79
Combined	0.69	**0.77**	0.67	0.51	**0.72**	0.39	**0.88**	0.82	**0.95**	**0.81**

Performance on Test Set: Given GPT2's performance on the development test data, we decided to submit it as our final run for the competition. On the test data, GPT2 achieved an F1 of 0.67, placing 18th overall.

Table 4
Official performance at CLEF 2025 Subjectivity track. Macro f1 results for the test dataset comparing our approach (NapierNLP) with the top three teams.

Position	Team	Macro F1
1	msmadi	0.8052
2	kishan_g	0.7955
3	CEA-LIST	0.7739
	...	
18	**NapierNLP**	0.6724
	...	
-	Baseline	0.5370

The results of all four models on the test data can be seen in Table 5. Among the three individual models, GPT-Neo performed the best, matching the combined model with a macro F1 score of 0.72.

Notably, unlike the results on the development test data, GPT-Neo outperformed GPT2 in the subjective F1 score, achieving 0.61 compared to GPT2's 0.56. Qwen performed the worst, scoring the lowest across all tests. This lower performance is likely due to insufficient training time, as it was trained for 100 epochs, achieving a macro F1 of 0.92 on the training set, compared to F1 of 1.00 for both GPT2 and GPT-Neo.

Table 5
Performance of models on the test dataset in terms of F1, Precision, Recall, and Accuracy.

Model	Macro			Subjective			Objective			Accuracy
	F1	P	R	F1	P	R	F1	P	R	
Baseline	0.53	0.54	0.55	0.41	0.33	0.52	0.66	0.76	0.59	0.57
GPT2	0.67	0.67	0.69	0.56	0.50	0.64	0.79	0.84	0.74	0.71
Qwen2.5	0.64	0.64	0.65	0.51	0.47	0.55	0.78	0.81	0.75	0.69
GPT-Neo	**0.72**	**0.72**	**0.73**	**0.61**	**0.59**	0.64	**0.84**	0.85	**0.82**	**0.77**
Combined	**0.72**	0.71	**0.73**	**0.61**	0.57	**0.66**	0.83	**0.86**	0.80	0.76

4.4. Failure Analysis

Figure 1 shows the confusion matrices for all four models on the test data. As anticipated, all models demonstrated better performance in classifying objective sentences, which may be attributed to the imbalance in both the training and testing datasets. However, the combined model obtained the highest number of true positives for the subjective class.

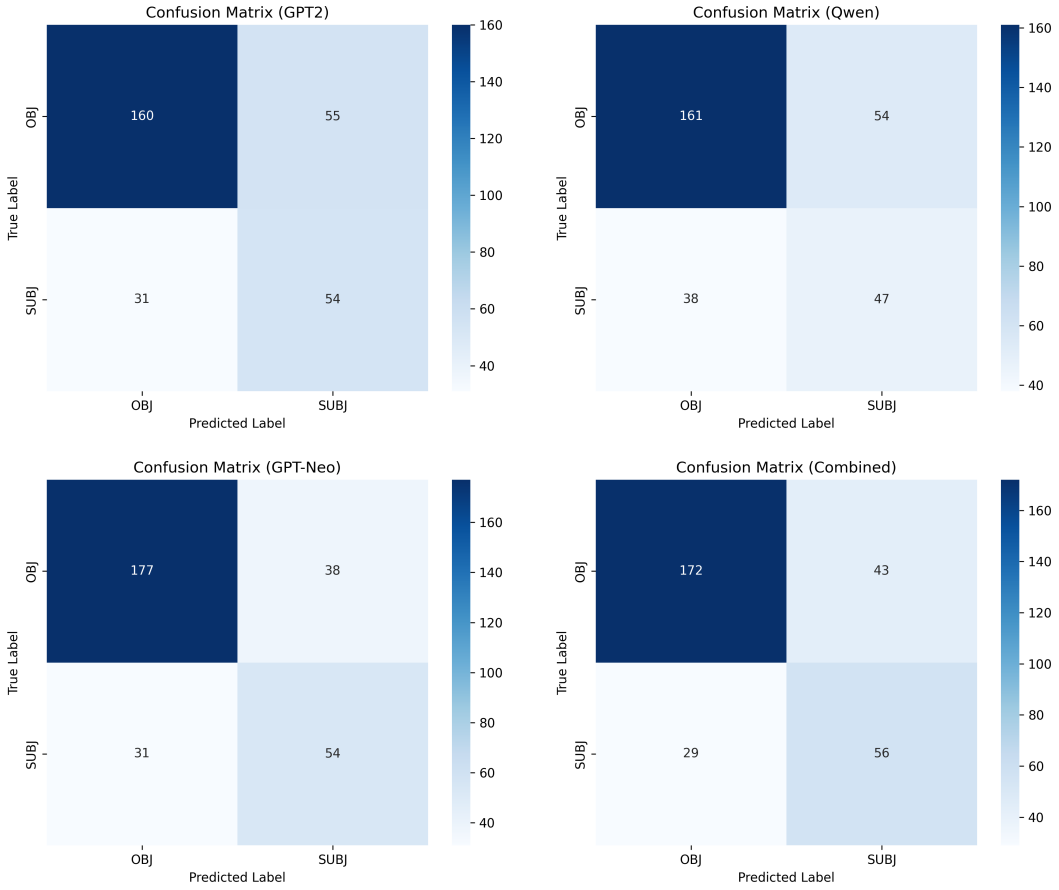

Figure 1: Confusion Matrices for each model comparing the predicted label to the gold-standard True label

Each of the three individual models correctly identified some instances that the others failed to classify accurately, ultimately contributing to incorrect final labels. Among the 72 errors in the final combined model, 24 sentences were misclassified by all three models, with an equal distribution between subjective and objective categories.

Table 6
Examples of the test sentences and their gold standard labels compared to each of the model's outputs. Objective prediction is coloured blue and subjective prediction is red.

#	Sentence	Gold	Combined	GPT2	Qwen2.5	GPT-Neo
1	Years later, in Korea, I was much happier, but I still wanted to drown out the reality of the $60,000 of student loans I felt like I was spending a lifetime paying off and my feelings of impostor syndrome.	**OBJ**	SUBJ	OBJ	SUBJ	SUBJ
2	Despite the hardships, however, revolutionary optimism was palpable.	**OBJ**	SUBJ	SUBJ	SUBJ	OBJ
3	The chancellor asserts, but can't possibly be confident, that planes will be landing on a third runway by 2035.	**SUBJ**	OBJ	OBJ	OBJ	OBJ
4	The reality is that Labour's wider agenda does not support the broad-based growth that would truly produce those improvements.	**SUBJ**	SUBJ	SUBJ	SUBJ	SUBJ
5	Speaking delegates condemned imperialists and warmongers, and shared tales of national struggles and workplace victories.	**OBJ**	OBJ	SUBJ	OBJ	OBJ
6	It must look elsewhere to make up for lost Russian gas.	**OBJ**	OBJ	OBJ	SUBJ	OBJ
7	Planning delays, likewise, are the product of austerity and the resulting privatisation and outsourcing of local government functions.	**OBJ**	OBJ	OBJ	OBJ	SUBJ
8	If investors can be found to stump up that kind of cash, they will want a return.	**OBJ**	SUBJ	SUBJ	OBJ	SUBJ
9	In a story that has been repeated countless times across Europe, the right's failed austerity agenda has abetted the rise of the far right.	**SUBJ**	OBJ	OBJ	SUBJ	OBJ
10	So they become a different person.	**OBJ**	OBJ	OBJ	OBJ	OBJ

For some sentences that were misclassified, the reasoning behind the errors can be identified. For example, Sentence 3 in Table 6 was incorrectly predicted as objective by all models; however, the phrase "but can't possibly be confident" indicates that it is actually subjective. Without this clause, the sentence could be correctly classified as objective. Similarly, sentences containing first-person pronouns, like Sentence 1, may have been misclassified because this is a common feature of many subjective sentences.

5. Conclusion

For our approach, we tested three different large language models and employed a majority voting system to classify sentences as subjective or objective. Our results indicate that even with limited GPU power, LLMs can effectively determine whether a sentence is subjective or objective. The majority voting system matched or even outperformed the individual models, suggesting that combining model outputs can yield more reliable predictions than relying on a single model.

6. Future Work

There are several ways in which this research could be enhanced in future work. Using larger versions of the models and allowing for longer training times could be feasible with improved GPU resources. This

would also enable us to explore the use of other models, such as Llama [24] or Mistral [25]. Additionally, we could investigate a prompting approach, either as an alternative to or in conjunction with our current method.

Another possible idea is to reevaluate how the final results are combined. Whilst majority voting proved effective for this project, it could become problematic with an even number of models, complicating the task of establishing a majority. Employing other ensemble methods, such as weighted voting - where the models are assigned weighted on their results- or incorporating a confidence score for each model could further improve our outcomes.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, G. Razvan, A. Mihail, Overview of the CLEF-2025 CheckThat! lab task 1 on subjectivity in news article, in: [4], 2025.

[4] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[5] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–77. URL: https://www.sciencedirect.com/science/article/pii/S1566253517303901. doi:https://doi.org/10.1016/j.inffus.2017.12.006.

[6] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier, P. Égré, Hybrinfox at checkthat! 2024 – task 2: Enriching bert models with the expert system vago for subjectivity detection, 2024. URL: https://arxiv.org/abs/2407.03770. arXiv:2407.03770.

[7] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, 2024. URL: https://arxiv.org/abs/2307.06435. arXiv:2307.06435.

[8] M. Shokri, V. Sharma, E. Filatova, S. Jain, S. Levitan, Subjectivity detection in English news using large language models, in: O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, S. Tafreshi (Eds.), Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 215–226. URL: https://aclanthology.org/2024.wassa-1.17/. doi:10.18653/v1/2024.wassa-1.17.

[9] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation (formerly Computers and the Humanities) 39 (2005) 164–210. doi:10.1007/s10579-005-7880-9.

[10] F. Antici, A. Galassi, F. Ruggeri, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on english news articles, 2024. URL: https://arxiv.org/abs/2305.18034. arXiv:2305.18034.

[11] E. Savinova, F. Moscoso Del Prado, Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements, in: J. Barnes, O. De Clercq, R. Klinger (Eds.), Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 305–314. URL: https://aclanthology.org/2023.wassa-1.27/. doi:10.18653/v1/2023.wassa-1.27.

[12] A. Paran, M. Hossain, S. Shohan, J. Hossain, S. Ahsan, M. Hoque, Semanticcuetsync at checkthat! 2024: Finding subjectivity in news articles using llama notebook for the checkthat! lab at clef 2024, 2024.

[13] S. Gruman, L. Kosseim, Clac-2 at checkthat! 2024: A zero-shot model for check-worthiness and subjectivity classification, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271822050.

[14] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, 1st ed., Chapman & Hall/CRC, 2012.

[15] Fusion of Label Outputs, John Wiley & Sons, Ltd, 2004, pp. 111–149. doi:https://doi.org/10.1002/0471660264.ch4.

[16] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. arXiv:2206.07682.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[18] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Fan, Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

[19] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[20] S. Black, G. Leo, P. Wang, C. Leahy, S. Biderman, GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021. URL: https://doi.org/10.5281/zenodo.5297715. doi:10.5281/zenodo.5297715.

[21] F. Ruggeri, F. Antici, A. Galassi, K. Korre, A. Muti, A. Barrón-Cedeño, On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection, 2023.

[22] F. Antici, F. Ruggeri, A. Galassi, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on English news articles, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 273–285. URL: https://aclanthology.org/2024.lrec-main.25/.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.