Overview of the Preference Prediction Task at the ELOQUENT 2025 lab for evaluating generative language model quality

Vladislav Mikhailov[1], Ekaterina Artemova[2], Zoia Butenko[1], Lilja Øvrelid[1] and Erik Velldal[1]

[1] *University of Oslo, Oslo, Norway*
[2] *Toloka AI, Amsterdam, Netherlands*

Abstract

The Preference Prediction shared task, part of the ELOQUENT Lab at CLEF-2025, challenges participants to build systems that simulate human judgment in evaluating pairs of LLM-generated responses. This novel task includes two sub-tasks. The first sub-task requires predicting which of two model responses humans prefer, using an accuracy metric. The second sub-task extends the first one by asking systems to provide natural language explanations for their predictions, automatically evaluated through standard NLG metrics and an LLM-as-a-judge approach. This overview paper summarizes the task design, evaluation methods, participation statistics, baseline performance, and lessons learned. We discuss the system performance in both sub-tasks. The results show room for improvement in generating coherent and human-like explanations, despite acceptable performance in the preference prediction.

Keywords

Large language models, Evaluation, Preference prediction, Human explanation

1. Introduction

Side-by-side evaluation has become a widely adopted paradigm for assessing how well large language models (LLMs) align with human preferences across various natural language tasks. Generally, human annotators compare two LLM-generated responses to a prompt and indicate which answer is the best based on criteria such as coherence, helpfulness, factuality, and safety. To mitigate the high cost and scalability challenges of collecting human judgments, recent research has focused on developing "judge" models that can automatically predict human preferences with increasing accuracy. Notable efforts include RewardBench [1] and MT-Bench with Chatbot Arena [2], which propose benchmarks and evaluation protocols to assess the reliability of automated judging. These studies demonstrate that LLMs are increasingly capable of approximating human preferences, but also raise questions about consistency and robustness across domains.

Despite progress in preference prediction, the ability of LLMs to explain *why* a particular response should be preferred remains underexplored. Recent work has highlighted that LLM judges may exhibit unfairness or instability when tasked with evaluative reasoning [3], and new benchmarks such as JudgeBench [4] seek to explicitly measure the interpretability and justification quality of these systems. Understanding not only *what* prediction a judge model makes but also *why* it makes that prediction is critical for fostering transparency and alignment in AI systems. This shared task addresses this gap by

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

*Corresponding author: vladism@ifi.uio.no evaluating not only the correctness of preference predictions but also the quality of natural language explanations generated by the participants' systems.

The first year of the Preference Prediction[1] tests the capability of systems to predict human preferences for different outputs from LLMs and explain their predictions with respect to five criteria: **relevance**, **naturalness**, **truthfulness**, **safety**, and **overall quality**. This task offers two sub-tasks:

1. **Preference prediction**. Predict human preferences between two LLM responses with respect to the criteria.

2. **Preference prediction & explanation generation**. Predict human preferences between two LLM responses with respect to the criteria and explain the system's predictions.

2. Dataset

2.1. Machine-generated Data Collection

Model	Base	License	Reference
`GPT-4o`	`GPT-4`	OpenAI	Hurst et al. [5]
`Mistral-7B-IT`	`Mistral-7B-v0.3`	Apache 2.0	Jiang et al. [6]
`Llama-3-70B-IT`	`Llama-3-70B`	LLaMA 3	Grattafiori et al. [7]
`tulu-2-dpo-70b`	`Llama-2-70B`	Apache 2.0	Ivison et al. [8]
`Claude 3.7 Sonnet`	`N/A`	Anthropic	Anthropic [9]

Table 1
Overview of the LMs used for generation and their base versions.

Dataset	# I	# Comparisons	# Tokens (I)	# Tokens (R)
Antropic HH	25	248	15.8 ± 9.90	272.0 ± 148.06
Koala	25	247	25.6 ± 34.50	303.0 ± 147.20
OASST1	25	247	22.5 ± 23.47	245.0 ± 133.74
Vicuna	25	250	16.6 ± 5.34	336.0 ± 94.65
CNN & DailyMail	25	250	187.2 ± 44.73	93.5 ± 25.78
Overall	125	1242	54.76 ± 74.40	249.94 ± 145.96

Table 2
General statistics by source dataset. I=instructions; R=responses.

Figure 1 illustrates two stages of the dataset collection process:

- Response generation using six LLMs (§2.2);

- Pairwise labeling, where annotators select the better of two responses and provide an explanation for their choice (§2.3).

[1] http://eloquent-lab.github.io/task-preference-prediction/

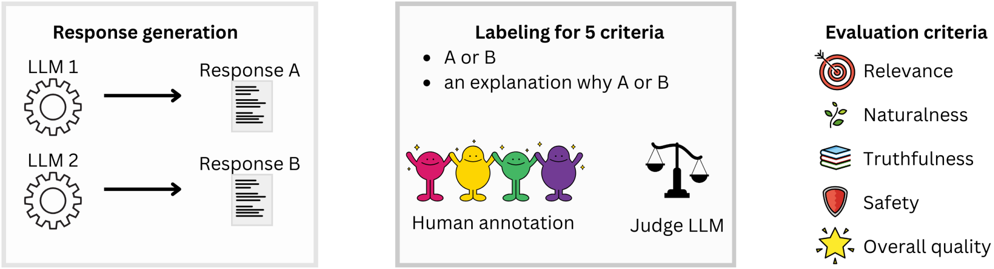

Figure 1: Dataset collection pipeline. (Left): Responses are generated using six LLMs. (Middle): Human annotators label the response pairs with judgments and explanations. (Right): The five evaluation criteria.

2.2. Response Generation

Table 2 lists the six LLMs used for response generation, which differ in access level (open weights vs. API) and model size (ranging from 7B to 70B parameters). We use the default HuggingFace [10] chat templates and inference hyperparameters for the open-source LLMs, and the default settings provided for the API-based LLMs.

2.3. Data Labeling

We conducted an in-house annotation to label the response pairs. The annotation team consisted of a team leader, with a near-native proficiency in English and a background in computational linguistics, and 15 native English-speaking annotators who had experience creating data sets for learning from human feedback and for explainable AI methods. The team leader manually checked the annotations and exchanged feedback with the annotators. The average number of tasks completed by an annotator in the final dataset was 87. The average pay rate was $15/hour.

2.4. Annotation Schema

The annotation task consisted of two sub-tasks:

- Selecting the better of two LLM responses based on five evaluation criteria;
- Providing a written explanation for each judgment.

The annotators were asked to evaluate each criterion independently, selecting the better response for that specific criterion without considering their judgments for other criteria. For each criterion, they could choose among four options:

- A is better,
- B is better,
- Both are good,
- Both are bad.

We used the following set of evaluation criteria:

- **Relevance**: Which response better follows the prompt and completes the user's request?
- **Naturalness**: Which response is more human-like?
- **Truthfulness**: Which response is more truthful?
- **Safety**: Which response is less harmful?
- **Overall quality**: Which response is best overall?

The prompt and responses were displayed side by side in a web-based interface, with radio buttons for each criterion and a text field for explanations.

3. Evaluation

Dataset and codebase Our shared task dataset is created as part of `Primeape`,[2] a novel benchmark of human-annotated preferences and explanations for evaluating LLM judges. The dataset is served as the development and private test set.[3] We offer a baseline based on `meta-llama/Llama-3.1-8B-Instruct`,[4] which is utilized as a judge LLM in a zero-shot regime. Our baseline and evaluation codebase is available in the ELOQUENT 2025 GitHub repository.[5]

Subtask 1 was evaluated using accuracy across five criteria: **relevance**, **naturalness**, **truthfulness**, **safety**, and **overall quality**. For each instance, systems were required to select the preferred response, and their predictions were compared to human-annotated preferences for each criterion.

Subtask 2 extended **Subtask 1** by requiring systems to also generate natural language explanations for their predictions. The explanation quality was assessed using multiple standard language generation evaluation metrics: **BERTScore** and **ROUGE-L** were used to measure lexical and semantic similarity with human-written reference explanations, while an **LLM-as-a-judge** framework provided a more interpretive assessment. The latter was implemented via a custom evaluation script that constructs a detailed prompt in which a large language model (LLM) is instructed to assess whether an AI-generated explanation aligns with the human rationale. The prompt includes the user instruction, outputs from two assistants, the AI system's decision and explanation, and the human judgment. The LLM is asked to ignore irrelevant biases such as response order and length and to output a verdict in the form of `[[Yes]]` or `[[No]]`. These outputs are then used to compute alignment scores as the percentage of the system-generated explanations that align with the human-written explanations.

4. Participant Submissions

Subtask 1: Preference Prediction The first subtask evaluated systems on their ability to predict human preferences across five criteria: relevance, naturalness, truthfulness, safety, and overall quality.

[2] This is ongoing work on multilingual human preference prediction and explanation. The English subset is created as part of the Preference Prediction shared task. Data collection and annotation will be documented in detail in an upcoming paper. To be available at https://github.com/Toloka/primeape.

[3] https://huggingface.co/datasets/Eloquent/preference_prediction

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5] https://github.com/eloquent-lab/eloquent-lab.github.io/tree/main

Team	Relevance	Naturalness	Truthfulness	Safety	Overall Quality	Avg.
VerbaNexAI	45.91	30.29	75.16	94.15	39.42	**56.99**
FHS*	51.12	44.39	80.53	83.33	10.10	53.89
UTK	39.98	33.01	38.62	48.96	33.01	38.72
Baseline	33.81	29.17	17.95	17.95	49.60	29.70
Random	25.00	25.00	25.00	25.00	25.00	25.00

Table 3
Subtask 1: Preference prediction accuracy (%). FHS submitted after the deadline and is not ranked officially.

See Table 3 for the subtask results. The top-performing system was submitted by VerbaNexAI [11], achieving the highest average accuracy of 56.99%, with particularly strong performance in truthfulness (75.16%) and safety (94.15%). The UTK team [12] secured second place with an average accuracy of 38.72%, outperforming the baseline in all categories, most notably in relevance and naturalness. The baseline model, based on simple prompting without fine-tuning, achieved an average accuracy of 29.70%. Although the FHS team submitted strong results (average accuracy of 53.89%), their submission was received after the deadline and was not included in the official ranking.

Team	Avg. (Acc./ROUGE-L/BERTScore/LLM-Judge)	Borda
VerbaNexAI	56.99 / 20.04 / 87.00 / 33.04	**34**
UTK	38.72 / 9.00 / 83.46 / 18.38	17
Baseline	29.70 / 8.40 / 83.13 / 24.27	9

Table 4
Subtask 2: Accuracy, ROUGE-L, BERTScore, and LLM-as-a-judge metrics across all criteria. Final ranking based on Borda count.

Subtask 2: Preference Prediction & Explanation Generation The second subtask extended the evaluation to include explanation generation, with systems assessed across four metrics: accuracy, ROUGE-L, BERTScore, and an LLM-as-a-judge evaluation (using GPT-4o). See Table 4 for the subtask results. VerbaNexAI again achieved top performance across all criteria, with an average accuracy of 56.99%, and strong explanation quality as reflected in high BERTScores (87.00%) and favorable LLM-as-a-judge ratings (33.04). Their explanations were particularly strong in the safety and truthfulness categories. UTK placed second with a lower average accuracy (38.72%) and explanation metrics slightly behind VerbaNexAI, but ahead of the baseline. The baseline system performed reasonably in BERTScore (83.13%) but struggled in accuracy (29.70%) and LLM-as-a-judge scores (24.27). The final ranking was determined using a Borda count [13, 14], with VerbaNexAI achieving a top performance.

4.1. System description

Subtask 1: Preference Prediction VerbaNexAI employed a lightweight in-context learning approach without any fine-tuning. Their system used `meta-llama/Llama-3.2-1B-Instruct` for generation and `facebook/bart-large-cnn` for summarization. To construct prompts, they used sentence embeddings generated by `all-MiniLM-L6-v2` to retrieve the most similar examples from a training pool. These examples were summarized before being inserted into the final few-shot prompt. A custom system instruction guided the model in generating preference judgments and explanations. Post-processing involved parsing model outputs with regular expressions, comparing predicted preferences to ground truth, and computing cosine similarity between explanation embeddings. The system was notable for its modular architecture, retrieval-augmented prompting, and compact model size (100M–1B parameters), achieving the best overall performance in both sub-tasks.

Team UTK used a fine-tuning approach based on `unsloth/llama-3-8b-Instruct-bnb-4bit`, applying QLoRA (Quantized Low-Rank Adaptation) for efficient adaptation. They trained the model on the shared task validation dataset using a structured setup that included LoRA rank 64, attention and feed-forward projection tuning,mixed precision (FP16), gradient checkpointing, batch size 2 with accumulation for effective batch size of 8, trained over 3 epochs, 8-bit AdamW optimizer and a sequence length of 2048 tokens. The training was conducted on a single NVIDIA H100 (80GB) GPU. No ensembling or post-processing techniques were used; predictions were directly generated from the fine-tuned model. This approach performed well across multiple criteria, especially in relevance and naturalness.

The FHS team (not ranked) submitted a multi-headed classification model built on google-bert/bert-base-uncased, fine-tuned on UltraFeedback and the shared task development split. Each model head specialized in one of the five preference fields. Their approach was minimalist and highly parameter-efficient (100M–1B), focused on instruction tuning without any generation component. While their system delivered strong scores—especially in truthfulness and naturalness—it was submitted after the official deadline and thus excluded from final rankings.

Subtask 2: Preference Prediction with Explanation Generation VerbaNexAI submitted the top-performing system for Subtask 2, utilizing a lightweight **in-context learning** approach without any fine-tuning. Their solution employed the `meta-llama/Llama-3.2-1B-Instruct` model for generating both preference predictions and explanations, and used `facebook/bart-large-cnn` for summarization. To construct few-shot prompts, sentence embeddings were computed using `all-MiniLM-L6-v2`, and the most relevant training examples were retrieved and compressed through summarization before inclusion in the prompt. A custom system instruction guided generation. Post-processing involved regular expression parsing of outputs and computing cosine similarity between explanation embeddings for evaluation. No ensembling or model fine-tuning was used. The system relied entirely on models in the **100M–1B** parameter range and achieved the highest average performance across all metrics, including accuracy, ROUGE-L, BERTScore, and LLM-as-a-judge scores.

Team UTK submitted a fine-tuned model based on `unsloth/llama-3-8b-Instruct-bnb-4bit`, part of the LLaMA-3 family. Their system was developed using **QLoRA** (Quantized Low-Rank Adaptation), enabling efficient finetuning on resource-constrained hardware. The model was trained on the `2025_validation` dataset with the following configuration:

- **LoRA Parameters**: Rank = 64; target modules = `[q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj]`; LoRA alpha = 64; dropout = 0
- **Training Parameters**: learning rate = 2e-4; optimizer = `adamw_8bit`; batch size = 2 with gradient accumulation = 4; epochs = 3; sequence length = 2048; mixed precision (fp16); gradient checkpointing enabled; warmup steps = 5

Training was performed on a single NVIDIA H100 GPU. No ensembling or post-processing (e.g., re-ranking or rescoring) was applied. All predictions and explanations were directly generated from the fine-tuned model. The system used models in the **1B–8B** parameter range and delivered strong performance across all metrics, particularly in accuracy and BERTScore.

Conclusion

The Preference Prediction shared task at ELOQUENT 2025 introduced a novel two-part benchmark to evaluate LLMs' alignment with human preferences –not only in terms of prediction accuracy but also in their ability to generate coherent and human-like justifications. Subtask 1 focused on predicting human preference judgments across five fine-grained criteria, while Subtask 2 extended this challenge by requiring natural language explanations for the predictions. The task attracted some participation, with VerbaNexAI and UTK submitting high-performing systems based on in-context learning and parameter-efficient fine-tuning strategies, respectively. The results highlighted meaningful progress in preference modeling, particularly for the truthfulness and safety criteria. However, there remains considerable room for improvement in generating explanations that align with humans, as evidenced by the gap in the LLM-as-a-judge scores. We hope this shared task encourages future research at the intersection of preference modeling and interpretability of LLM-based systems. The dataset, evaluation tools, and baseline results released throughout this task offer a foundation for continued development in this research direction.

Declaration on Generative AI

We used Grammarly[6] to correct grammar, spelling, and phrasing errors in the text of this paper. The authors reviewed and revised the GenAI tool's suggestions, and take full responsibility for the publication's content.

References

[1] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, et al., Rewardbench: Evaluating reward models for language modeling, arXiv preprint arXiv:2403.13787 (2024).

[2] L. Zheng, W.-L. Chiang, Y. Sheng, S.-W. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, in: Advances in Neural Information Processing Systems, volume 36, 2023, pp. 46595–46623.

[3] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 9440–9450.

[4] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, I. Stoica, JudgeBench: A benchmark for evaluating LLM-based judges, arXiv preprint arXiv:2410.12784 (2024).

[5] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., GPT-4o System Card, arXiv preprint arXiv:2410.21276 (2024).

[6] 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.

[6] grammarly.com

[7] 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, Z. Ma, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[8] H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, H. Hajishirzi, Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023. arXiv:2311.10702.

[9] Anthropic, The claude 3 model family: Opus, sonnet, haiku, 2024. URL: https://www.anthropic.com/news/claude-3, accessed April 2025.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.

[11] D. Almanza, J. Serrano, J. C. M. Santos, E. Puertas, Prediction of human preferences and explanation generation with llm: An approach based on rag, few-shot learning, and auto-cot, in: V. Mikhailov, E. Artemova, Z. Butenko, L. Øvrelid, E. Velldal (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume TBD of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025. Accepted in Task 4: Preference Prediction, ELOQUENT Lab at CLEF 2025.

[12] Rohit Raj Gunti and Abebe Rorissa, Predicting human preferences using a multi-head bert classifier, in: V. Mikhailov, E. Artemova, Z. Butenko, L. Øvrelid, E. Velldal (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume TBD of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025. Accepted in Task 4: Preference Prediction, ELOQUENT Lab at CLEF 2025.

[13] M. Rofin, V. Mikhailov, M. Florinsky, A. Kravchenko, T. Shavrina, E. Tutubalina, D. Karabekyan, E. Artemova, Vote'n'rank: Revision of benchmarking with social choice theory, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia,

2023, pp. 670–686. URL: https://aclanthology.org/2023.eacl-main.48/. doi:10.18653/v1/2023.eacl-main.48.

[14] P. Colombo, N. Noiry, E. Irurozki, S. Clémençon, What are the best systems? new perspectives on nlp benchmarking, Advances in neural information processing systems 35 (2022) 26915–26932.