Overview of Dialog System Evaluation Track: Dimensionality, Language, Culture and Safety at DSTC 12

John Mendonça[1,2], Lining Zhang[3], Rahul Mallidi [3], Alon Lavie [5,6], Isabel Trancoso [1,2], Luis Fernando D'Haro [4], João Sedoc [3]
[1]INESC-ID, Lisbon
[2]Instituto Superior Técnico - University of Lisbon
[3]Department of Technology, Operations, and Statistics, New York University
[4]Speech Technology and Machine Learning Group - Universidad Politécnica de Madrid
[5]Carnegie Mellon University
[6]Phrase, Pittsburgh
Correspondence: john.mendonca@inesc-id.pt

Abstract

The rapid advancement of Large Language Models (LLMs) has intensified the need for robust dialogue system evaluation, yet comprehensive assessment remains challenging. Traditional metrics often prove insufficient, and safety considerations are frequently narrowly defined or culturally biased. The DSTC12 Track 1, "Dialog System Evaluation: Dimensionality, Language, Culture and Safety," is part of the ongoing effort to address these critical gaps. The track comprised two subtasks: (1) Dialogue-level, Multi-dimensional Automatic Evaluation Metrics, and (2) Multilingual and Multicultural Safety Detection. For Task 1, focused on 10 dialogue dimensions, a Llama-3-8B baseline achieved the highest average Spearman's correlation (0.1681), indicating substantial room for improvement. In Task 2, while participating teams significantly outperformed a Llama-Guard-3-1B baseline on the multilingual safety subset (top ROC-AUC 0.9648), the baseline proved superior on the cultural subset (0.5126 ROC-AUC), highlighting critical needs in culturally-aware safety. This paper describes the datasets and baselines provided to participants, as well as submission evaluation results for each of the two proposed subtasks.

1 Introduction

The rapid advancements in Large Language Models (LLMs) have led to increasingly sophisticated conversational agents capable of engaging in complex and nuanced dialogues. As these models become more integrated into various applications, from customer service to personal assistants, ensuring their quality, reliability, and safety is paramount. However, evaluating dialogue systems comprehensively remains a significant challenge (Rodríguez-Cantelar et al., 2023; Mendonça et al., 2024a). Traditional metrics often fall short of capturing the multifaceted nature of human-like conversation, and safety considerations are frequently narrowly defined or culturally biased, failing to address the full spectrum of potential issues.

Addressing the first aspect of this challenge – the limitations of current evaluation metrics – previous challenges and works focus largely on turn-level dialogue evaluation (Zhang et al., 2021; Rodríguez-Cantelar et al., 2023; Mehri et al., 2022) and often lack further investigation of dialogue-level evaluation through automatic metrics. As LLMs advance, aspects of conversations beyond coherence, fluency, etc. should also be studied. Additionally, these aspects should provide a more fine-grained analysis of the levels of quality for the whole conversation, moving beyond simplistic turn-based scores.

Complementing the need for improved quality assessment, the safety dimension, highlighted as a critical concern from the outset, presents its own distinct set of urgent problems. Users are increasingly challenging current chatbots to generate harmful and/or unsafe answers. In addition, even without adversarial probing, generated responses may contain unhelpful and/or harmful content. Therefore, the automatic detection of this content is important in the deployment of these systems. Unfortunately, existing safety evaluation frameworks frequently narrow the notion of safety to strict definitions of bias and toxicity, discarding other safety aspects (Shuster et al., 2022; Ouyang et al., 2022). Furthermore, a significant limitation in current safety paradigms is their predominant focus on the English language. We attempt to mitigate this bias by expanding safety datasets to a diverse set of languages and cultures. Beyond facilitating the study of safety across cultures, this

Proceedings of the Twelfth Dialog System Technology Challenge (DSTC12), pages 27–35

Attribute	Description
Empathy	Do you think your conversational partner had genuine empathy?
Trust	Based on the conversation, your conversational partner seems trustworthy
Skill	Based on the conversation, your conversational partner seems skilled
Talent	Based on the conversation, your conversational partner seems talented
Capability	Based on the conversation, your conversational partner seems capable
Relevance	Responses address the given context or query well, ensuring that the information provided is pertinent and directly applicable.
Non-Repetition	How repetitive was this chatbot?
Proactivity	Responses actively and appropriately move the conversation along different topics
Curiosity	How much did the chatbot try to get to know you?
Overall	How was the conversation?

Table 1: Evaluation dimensions and definitions for Task 1.

also allows for the evaluation of the robustness of safety classifiers in terms of culture and language.

1.1 Track Details

To address these gaps, this paper presents Track 1 of DSTC12, entitled "Dialog System Evaluation: Dimensionality, Language, Culture and Safety." The shared task was divided into two tasks: Dialogue-level and Multi-dimensional Automatic Evaluation Metrics (§2), and Multilingual and Multicultural Safety Detection (§3). This year's iteration introduced two key novelties aimed at enhancing participation and streamlining the evaluation process: (1) a focus on model efficiency and (2) the adoption of an online competition platform.

Firstly, recognizing that current dialogue evaluation research (and the broader "LLM-as-a-judge" paradigm) often relies on extremely large, proprietary models such as GPT-4 or Claude accessed via APIs, we imposed a significant constraint on model size. Participants were restricted to utilizing open-source LLMs with fewer than 13 billion parameters. This decision was motivated to encourage innovative, efficient solutions that do not solely depend on prompting state-of-the-art models.

Secondly, we utilized the Codabench platform[1] for managing submissions and leaderboards. This facilitated a more dynamic and interactive participation experience. We also released the datasets via Huggingface Datasets to facilitate easy access[2]. On the one hand, it allowed participants to easily gauge their model's performance on the development set in real-time and compare their results against established baselines. On the other hand, for the test set, participants could receive immediate feedback on their system's performance upon submission. To maintain fairness and prevent over-fitting to the test set, submissions were limited to five attempts, and the test set leaderboard remained hidden until the conclusion of the competition.

2 Task 1: Dialogue-level and Multi-dimensional Automatic Evaluation Metrics

In this task, the goal was for participants to develop automatic evaluation metrics for open-domain dialogue. In particular, the submitted systems were expected to evaluate up to 10 different dimensions including previous common ones (Zhang et al., 2021; Rodríguez-Cantelar et al., 2023, i.e.), together with new ones like (Zhang et al., 2024). An overview of these dimensions are presented in Table 1. Similar to previous challenges and prior literature, we evaluated the systems using Spearman's rank correlation between human annotations and automatic metrics as our criterion.

2.1 Dataset

Our main dataset was separated into three collections: three bots (ChatGPT [2023], GPT-3, and BlenderBot-3) during Q1 2023 (TBD-Q1-2023), four bots (ChatGPT, Gemini, Claude, and Mixtral) during Q1 2024 (FBD-Q1-2024), and six bots (ChatGPT, Gemini, Claude, and through Hugging Chat (Mistral, Llama-3 instruct 70B, and Cohere)[3]

[1] We have opened the competitions as benchmarks for the broader community: Task 1; Task 2

[2] huggingface.co/dstc12

[3] The exact versions are `mistralai/Mistral-Nemo-Instruct-2407`, `meta-llama/Meta-Llama-3-70B-Instruct`, and `CohereForAI/c4ai-command-r-plus`.

during Q4 2024 (SBD-Q4-2024). The users in the conversations were undergraduate students. All conversations were read to verify that no personally identifiable information was present. For both FBD-Q1-2024 and SBD-Q4-2024 datasets, we controlled the topics present in the conversation:

T1 Talk about help for turning your homework in late.

T2 Finding an apartment.

T3 Finding something to do in the evening.

T4 Talk about something that is on your mind or bothering you.

T5 Learn about a topic that you are interested in.

T6 Talk about something silly with the chatbot.

Students were randomly assigned, without replacement, to both a chatbot and a conversation topic. They were instructed to interact for roughly 15 turns. After the conversation, they shared their conversation link and filled out the surveys. Subsequently, the conversation links were web scraped, and the conversational data were merged with the survey responses.

The dataset was split into development (TBD-Q1-2023 / FBD-Q1-2024) of 185 conversations and test set (SBD-Q4-2024) of 120 conversations. TBD-Q1-2023 included 8 participants, FBD-Q1-2024 had 4, and SBD-Q4-2024 had 6. TBD-Q1-2023 was used in the DSTC11 Track 4 challenge (Rodríguez-Cantelar et al., 2023) for both turn- and dialog-level evaluation, but only coarse-grained dimensions were used.

Following Zhang et al. (2024), we used a subset of dimensions for evaluation. Table 1 has the list of dimensions along with their definitions.

2.2 Baseline

As a baseline, we prompt Llama-3-8B-Instruct to provide an evaluation across all of the dimensions. The system prompt is presented in Table 3.

2.3 Participants

Team 1 Team 1 submitted four unique systems. System 1 employed a regression approach, training separate regression layers on top of a ModernBert encoder for each evaluation dimension using the DSTC-12, ConTurE (Ghazarian et al., 2022), and FED (Mehri and Eskenazi, 2020) datasets. System 2 utilized a prompting strategy, combining detailed dimension explanations and dialogue context with a selection of models (Deepseek Llama 8B, Deepseek Qwen 7B, Qwen 2.5 7B Instruct-1M), choosing the best-performing model per dimension based on validation set results. System 3 was a classification-based approach, training individual classifiers on an sBERT encoder for each dimension with normalized scores, also using the DSTC-12, ConTurE, and FED datasets. Finally, System 4, a hybrid model, selectively combined the outputs of System 1 (for dimensions like Talent and Relevance) and System 2 (for dimensions like Empathy and Overall) based on which system achieved the best correlation on the validation set for each specific dimension.

Team 2. This team adopted Qwen2.5-7B-Instruct as the base model and then utilized prompt engineering to enable the LLM to automatically output scores across various dimensions. Moreover, they included degree interpretations for different score levels within the prompt.

2.4 Results

The official results for Task 1 are provided in Table 2. The team score was computed as the mean absolute Spearman correlation across all dimensions. We can also see a per-dimension breakdown. Ideally, all correlations should be positive; however, in the absence of consistent definitions, some participants may have reverse-coded certain dimensions.

Quite surprisingly, the baseline outperformed the other teams, followed by Team 1 and then Team 2. This was largely due to the dimensions of relevance and non-repetition. While the baseline performed best overall, there was no clear winner across all dimensions, indicating substantial room for improvement. Most correlations were statistically significant, though relatively low for several dimensions. Unsurprisingly, all metrics correlated well with the overall conversational quality. The negative correlation for the 'skill' dimension highlights a clear area for further refinement.

We had one strange finding: on a relative basis, all automatic metrics favored Mistral, whereas human participants ranked it significantly lower. This is probably the reason for the large decrease in correlations between development and test sets.

Team	Avg.	Emp.	Trust	Skill	Talent	Capab.	Rel.	!Rep.	Proact.	Curio.	Overall
Baseline	.1681	.0647	-.1117	-.0955	.0962	.0677	.2337	*.3851*	-.0248	.2253	*.3766*
1	.1503	-.0788	.0067	-.2246	.0529	.1276	.0808	.1071	-.1457	*.3687*	*.3099*
2	.1360	-.1183	.2407	-.1094	.0873	.1717	.0738	.2220	.0185	.0064	*.3117*

Table 2: Spearman correlations for each team across all dimensions of Task 1. Correlations with $p < 0.05$ in *italic*.

3 Task 2: Multilingual and Multicultural Safety Detection

In this task, participants were tasked with developing automatic safety classifiers of responses generated by LLMs across different languages and cultures. To establish a clear and manageable benchmark, particularly given the diverse range of languages and the nuanced nature of safety, we simplified the core challenge to a binary classification problem: determining whether a given dialogue response is 'safe' or 'unsafe'. We consider a response to be "unsafe" if it violates the MLCommons AI Safety v0.5 Benchmark POC Taxonomy of Hazards (Vidgen et al., 2024)[4]. The taxonomy outlines seven key hazard categories within its scope for this version: Violent crimes, Non-violent crimes, Sex-related crimes, Child sexual exploitation, Indiscriminate weapons (CBRNE), Suicide & self-harm, and Hate. This approach, while a simplification of real-world complexities where harm can be multifaceted and context-dependent, allows for a more direct comparison of model capabilities in fundamental safety detection across varied linguistic and cultural contexts. It also provides a foundational step before tackling more granular multi-label or severity-level classifications.

3.1 Datasets

To allow participants to train and evaluate their models, we curated several datasets. These datasets were processed to conform to a shared conversational format, consisting of context and response pairs accompanied with additional metadata made available in the original datasets. These datasets were then translated to 7 languages (Arabic, German, Spanish, French, Japanese, Portuguese and Chinese) and made accessible to the participants on HuggingFace[5]. We present an overview of these datasets in Table 4.

[4] https://drive.google.com/file/d/1V8KFfk8awaAXc83nZZzDV2bHgPT8jbJY/view

[5] https://huggingface.co/dstc12

3.1.1 Development

Bot Adversarial Dialogue (Xu et al., 2021). This dataset was curated via a human-and-model-in-the-loop framework where crowdworkers were instructed to converse with various state-of-the-art dialogue models, actively probing the model to output unsafe or offensive responses. Each bot utterance within these interactions was annotated for safety, resulting in a corpus of approximately 5.8k dialogues (79k total utterances), with 40% of utterances being annotated as offensive.

Dialogue Safety (Dinan et al., 2019) was curated via a human-and-model-in-the-loop framework. Crowdworkers were presented with an existing dialogue context and were instructed to submit utterances they deemed offensive, specifically targeting those that an existing safety classifier would miss-classify as safe. This iterative process resulted in a corpus of approximately 6,000 "offensive" utterances, collected across both single-turn and multi-turn dialogue context settings. When combined with verified safe examples, these constitute a dataset totalling approximately 60,000 utterances, of which 10% are labelled offensive. For the purpose of this Task, we use the multi-turn subset.

ProsocialDialog (Kim et al., 2022) is a large-scale, multi-turn English dialogue dataset designed to teach conversational agents to respond prosocially to problematic user inputs. Generated via a human-AI collaborative framework, it contains 58,137 dialogues (331,362 utterances) covering diverse unethical, problematic, biased, and toxic situations. Prosocial responses are grounded in 160,295 commonsense social rules-of-thumb (RoTs), and dialogue turns are annotated with fine-grained safety labels accompanied by 497,043 free-form rationales.

3.1.2 Test

Soda-Eval (Mendonça et al., 2024b) is derived from the SODA dataset, and encompasses over 120,000 turn-level assessments across 10,000 dialogues. Each assessment, generated by GPT-4 and subsequently human-validated, includes identifica-

You are an impartial evaluator conducting a multidimensional assessment of text responses. Your role is to analyze and score all chatbot responses using the following criteria:

- Empathy: Based on the conversation, does the chatbot demonstrate understanding and compassion for the user's situation or emotions?
- Trust: Based on the conversation, does the chatbot seem trustworthy?
- Skill: Does the chatbot show competence in the subject matter, providing accurate and relevant information?
- Talent: Does the chatbot show talent in the subject matter, providing accurate and relevant information?
- Capability: Does the chatbot seem capable in interacting with the user?
- Relevance: Are all chatbot responses relevant given prior context?
- Non-Repetition: Does the chatbot avoid unnecessary repetition?
- Proactivity: Does the chatbot anticipate user needs?
- Curiosity: Does the chatbot demonstrate engagement by exploring the topic further or encouraging deeper discussion?
- Non-Repetition: Does the chatbot avoid unnecessary repetition?
- Overall: Overall assessment of the chatbot throughout the dialogue.

Scoring Guidelines:

- Focus only on the chatbot responses, not the user messages. - Assign a score between 1 and 5 for each relevant category based on the criteria above. - Do not output any other meta commentary or information.
Input: The input consists of a conversation between a user and a chatbot.

Output: [JSON format]

Table 3: Baseline evaluation prompt.

tion of fine-grained issues. We leverage the annotations for the quality dimensions "Anti-Social". However, since these annotations were automatically annotated using an LLM (GPT-4), we conduct a human validation step on the safety labels. From this validation step, we randomly select additional positive examples from Soda-Eval to derive a class-balanced set of size 325 examples.

CoSafe (Yu et al., 2024) is a benchmark designed to evaluate safety against multi-turn dialogue coreference attacks. The dataset was constructed by selecting 100 single-turn attack prompts for each of 14 harmful categories, originally defined by BeaverTails (Ji et al., 2023). These prompts were then automatically expanded into multi-turn dialogues using GPT-4, with the coreferentially-phrased attack query placed in the final turn to assess model vulnerabilities in resolving references within a harmful conversational context. We employed multiple LLMs to simulate diverse safety behaviors across model families and architectures. This diversity ensures that safety classifiers are not overfitted to idiosyncrasies of a single generation style and that evaluation generalizes across real-world deployment scenarios. The chosen models are aya-expanse-8b (Dang et al., 2024), EuroLLM-9B-instruct (Martins et al., 2024), LLama-3.2-Instruct (1B,3B) and LLama-3.1-8B-Instruct (Grattafiori et al., 2024), Ministral-8B-Instruct-2410 (MistralAI, 2024), and Qwen2.5-Instruct (3B,7B) (Qwen et al., 2025). We then conduct a human-validated automated annotation using GPT-4o (OpenAI et al., 2024) as an automated safety classifier. Then, all examples rated as unsafe are evaluated by a human annotator. A balanced safety-label subset is then sampled from these annotations.

Dataset	#Utterances (k)
BAD	69.3 / 7 / 2.6
Dialogue Safety	24 / 3 / 3
Prosocial Dialogue	120 / 20.4 / 25
Total	213.3 / 30.4 / 30.6
SODA-Eval	- / - / 325
CoSafe	- / - / 227
SafeWorld	- / - / 437

Table 4: Overview of datasets used in Task 2. For the development set, we provide train/validation/test sets.

SafeWorld (Yin et al., 2024) For the cultural subtask, we employ a curated version of the cultural-aware safety dataset of SafeWorld (Yin et al., 2024). SafeWorld is designed to assess alignment with geo-diverse cultural and legal safety standards by grounding queries on human-verified cultural norms and legal policies from 50 countries and 493 distinct regions/races. We focus on the "specific answer" and "comprehensive answer" query types. "Specific answer" queries (641 instances) require models to pinpoint a single, pre-defined cultural or legal guideline violated in a given scenario; "comprehensive answer" queries (577 instances) present situations where potential violations are ambiguous, tasking models to provide comprehensive responses covering relevant norms and policies across implicated regions. We prompt GPT-4o to determine if the policy or norm viola-

tion identified would elicit a safety violation. For the examples identified as unsafe, we then generate responses using several LLMs: aya-expanse-8B (Dang et al., 2024), EuroLLM-9B-instruct (Martins et al., 2024), gemma-3-9b-it (Gemma et al., 2025); gpt-4o-mini (OpenAI et al., 2024), LLama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.57B-Instruct (Qwen et al., 2025). Then, we ask GPT-4o (OpenAI et al., 2024) to determine if the the model response elicits the identified policy/norm violation.

Human validations to confirm the accuracy of the test set labels was conducted by a single annotator, a graduate student with a background of Computational Linguistics with extensive annotation experience.

3.1.3 Translations

We translate all of the context-response pairs for the development set as a whole, and for Soda-Eval and CoSafe. We selected a wide-range of languages that cover a diverse set of cultures: Arabic, German, English, Spanish, French, Japanese, Portuguese and Chinese.

Automated validation For translation model selection, we automatically validate the quality of translations in a smaller subset by employing an MT-QE model, in this case `wmt23-cometkiwi-da-xl`. We obtain a score of 0.7116 using GPT-4o and 0.7153 GPT-4o-mini. Given the small performance difference and the much lower price point for GPT-4o-mini, we opt with utilizing GPT-4o-mini for the translation of the full development set.

Human validation For the test set, we additionally conduct a human validation study to ensure the translations are of high quality and do not change the semantics of the response such that it changes the safety label. Human annotators – one per language, except for Japanese (2) – were recruited given their professional level of the target language and consisted of a mix of graduate students and post-docs with Linguistics and Computational Linguistics background with prior experience with annotations. Initial annotations showed that GPT-4o-mini was incurring in several translation errors that affect the safety of the response. As such, we re-translate the English set with GPT-4o, which, after subsequent human validation, confirmed the quality of its translations on the examples GPT-4o-mini failed.

3.2 Baseline

We adopted Llama-Guard-3-1B as the official baseline for Task 2. This model was aligned from the original Llama-3.2-1B LLM to safeguard against the MLCommons standardized hazards taxonomy. In particular, the English data used by Llama Guard was combined with additional multilingual human and synthetically generated Human-AI conversation data.

For the purposes of the task, baseline predictions for response classification are obtained by extracting the probability of the "unsafe" token appearing as the first decoded token to reflect the model's direct scoring of risk. This approach offers a deterministic, resource-efficient proxy for classification, aligning with recent work in zero-shot risk estimation and avoiding confounding artifacts from longer generation chains. Reproducible code can be found alongside the data on the HuggingFace dataset repository[6].

3.3 Participants

For this task, a total of 2 teams (Teams 2 and 3) participated with 6 submissions. Participants were asked to provide a description of their submissions. Team 2 submitted a similar system to the one presented in Task 1 (2, adapting the prompt for the safety task. Unfortunately, Team 3 did not provide an official description of their system. However, their submissions to the track platform suggest their approach consisted in the supervised finetuning of LLMs on the development data (`sft_500k_gemma-ck` and `llama3_sft_500k`) of gemma-2-9b-it and a LLama3 model respecting our model size restrictions (likely 8B).

Team	Average	Cultural	Multilingual
3	**.9046**	.4831	**.9648**
2	.8078	*.4830*	.8517
Baseline	*.7767*	**.5126**	*.8097*

Table 5: ROC-AUC results for Task 2. The first position is shown in **bold**, the second in underline and the third in *italic*.

3.4 Results

The official results for Task 2 are provided in Table 5. Team ranking is established by calculating the

[6] https://huggingface.co/datasets/dstc12/bot_adversarial_dialogue/blob/main/LlamaGuard.py

average ROC-AUC considering all languages and the cultural subset with equal weights. We also present ROC-AUC for the multilingual and cultural subsets separately. We employ ROC-AUC since it provides a threshold-independent assessment of a model's ability to distinguish between safe and unsafe content.

Team 3 ranked best in this Task, followed by Team 2. This is thanks to their strong performance on the multilingual subset, with Team 3 achieving a strong result of .9648, followed by Team 2 with .8517, which are significantly superior to the baseline results (.8097). However, when looking at the cultural subset, we note that the baseline was the best performing submission (.5126), with both Teams achieving similar results (around .4831). These results suggest that models finetuned for cultural agnostic safety concerns fail to account for cultural specificities. This behaviour may be an instance of catastrophic forgetting, since our baseline (LLama-Guard-3-1B) was able to outperform their stronger finetuned models.

4 Conclusions and Future Work

This paper presents the overview of Track 1 on "Dialog System Evaluation: Dimensionality, Language, Culture and Safety" organized as part of the 12th Dialogue System Technology Challenge (DSTC12). The track was organized in two tasks aimed at addressing two important problems of the state-of-the-art in Dialogue Systems: (1) Dialogue-level and Multi-dimensional Automatic Evaluation Metrics; (2) Multilingual and Multicultural Safety Detection.

While the track had 11 registered teams, only 3 participated. The first task drew two of these teams. We used Spearman's rank correlation coefficient absolute average value as the rank ordering for the teams. The baseline outperformed the best overall, but alone many different dimensions we see different methods performing better.

In the second task, two teams participated and comfortably outperformed the baseline on the multilingual subset, achieving very strong ROC-AUC. However, for the cultural subset, no team was able to outperform the baseline ROC-AUC, which sits at just .5126, indicating clear room for improvement.

As future work, Task 1, we plan to extend the analysis of fine-grained dimensions to understand the upper-bound of LLM-evaluation for dimensions of human quality assessment. Importantly, we plan to increase the diversity of participants to be more representative of larger populations. For Task 2, we plan on extending the safety classification task to include the full taxonomy, providing a more fine-grained assessment of risks.

Acknowledgments

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI), by Portuguese national funds through Fundação para a Ciência e Tecnologia (FCT) with references PRT/BD/152198/2021 and DOI:10.54499/UIDB/50021/2020.

This work is supported by the European Commission through Project ASTOUND (101071191 – HORIZON EIC-2021 – PATHFINDERCHALLENGES-01), and by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the European Union.

We also want to give thanks to MS Azure services (especially to Irving Kwong) for their sponsorship to continue processing new datasets that could be interesting for the dialogue community.

This research project is supported by the NYU ChatEval Team led by João Sedoc. He would like to thank NYU Stern for its funding.

References

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey

Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2022. Interactive evaluation of dialog track at DSTC9. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5731–5738, Marseille, France. European Language Resources Association.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2024a. On the benchmarking of LLMs for open-domain dialogue evaluation. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024b. Soda-eval: Open-domain dialogue evaluation in the age of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11687–11708, Miami, Florida, USA. Association for Computational Linguistics.

MistralAI. 2024. Un ministral, des ministraux | mistral ai.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander I. Rudnicky. 2023. Overview of robust and multilingual automatic evaluation metricsfor open-domain dialogue systems at DSTC 11 track 4. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 260–273, Prague, Czech Republic. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, and 81 others. 2024. Introducing v0.5 of the ai safety benchmark from mlcommons. *Preprint*, arXiv:2404.12241.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 2950–2968, Online. Association for Computational Linguistics.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Safeworld: Geo-diverse safety alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Gao Zuchen, Fei Mi, and Lanqing Hong. 2024. CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17494–17508, Miami, Florida, USA. Association for Computational Linguistics.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *Preprint*, arXiv:2111.02110.

Lining Zhang, João Sedoc, and Natalia Levina. 2024. Back to principles: Theory-driven evaluation of ai-based conversational agents. In *Forty-Fifth International Conference on Information Systems*.