Does Perspective Matter? Annotation Role Alignment in Emotion Recognition

Samuele D'Avenia[1], Timothy Meinert[2], Marta Marchiori Manerba[1], Valerio Basile[1] and Anna Koufakou[2]

[1]University of Turin
[2]Florida Gulf Coast University

Abstract

Annotated NLP datasets embed implicit communicative roles: some capture phenomena as self-reported by the text's author, others reflect third-party inferences about the author's state, and others still record the reader's own response. In each case, the annotation is grounded in the perspective of a person. Despite this variation, current practice treats these annotation roles as interchangeable, a flattening that propagates into model training and evaluation. We introduce a formal classification of three annotation roles: Speaker, Perceived Speaker, and Observer, and apply it to a benchmark of emotion recognition not originally designed with such distinctions in mind. We then conduct a systematic prompting study across five models, testing whether role-aligned prompts produce better performance and greater stability on role-matched datasets. Our results reveal that role alignment through prompting alone yields no consistent benefit: no clear performance gains emerge from matching prompt role to dataset role, and model behavior is unstable across prompt variations. These findings expose methodological gaps in existing annotation practices and suggest that perspective alignment requires more than surface-level prompt framing.

Keywords

Emotion recognition, Perspectivist NLP, Annotation role

1. Introduction

Annotated data is the backbone of NLP classification, yet the communicative role from which annotations are produced is rarely treated as a variable of interest. When considering subjective NLP tasks, some datasets capture labels that are self-reported by the authors of a text, while others rely on third-party observers, e.g., crowdworkers. Despite this methodological difference, current practice flattens these roles, without distinguishing annotation source [1]. This conflation is consequential: **intention and perception can diverge substantially** [2], and a model trained on observer-annotated data implicitly learns a different task than one trained on speaker-annotated data, whether or not this position is acknowledged.

The implications extend to **evaluation**. When a model is prompted to classify emotion, the framing of that prompt implicitly positions the model in a particular role: as the speaker inferring their own state, as an observer inferring the author's intent, or as a reader reporting their own reaction. If this role is inconsistent with the perspective embedded in the training or test data, the resulting evaluation is measuring a misaligned task. Yet this dimension has received limited systematic attention.

An early exception is Buechel and Hahn [3], who compared writer, reader, and text-centered annotation and found perspective to affect inter-annotator agreement, and released the EMOBANK corpus [4], doubly annotated for writer and reader emotion. While they distinguish a text-centered perspective, in which emotion is treated as a property of the text itself, our framework focuses on annotations tied to the perspective of a person. Additionally, these studies establish perspective as a consequential design variable, but treat it solely as a property of *human* annotation: they neither separate self-reporting authors from third parties inferring an author's state, nor ask whether models can be positioned in a

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy
0009-0000-9256-0707 (S. D'Avenia); 0000-0003-2251-1824 (M. M. Manerba); 0000-0001-8110-6832 (V. Basile); 0000-0002-0295-3759 (A. Koufakou)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper22.pdf

role at inference time. Similarly, Jang and Frassinelli [2] hypothesize an alignment between observer/speaker perspectives and encoder/decoder architectures, providing preliminary evidence on sarcasm. Building on Meinert et al. [5], who show that emotion recognition performance is influenced by dataset characteristics, annotation roles, and model architecture, we investigate whether explicitly aligning prompt roles to annotation roles produces measurable gains in performance and consistency.

To pursue this goal, we first formalize a tripartite classification of annotation roles, Speaker, Perceived Speaker, and Observer, and apply it retrospectively to existing datasets not originally designed with such distinctions in mind. We then design role-specific prompting strategies and evaluate whether role-aligned prompts yield more stable predictions on role-matched datasets. We conduct our experiments with a focus on emotion classification, a subjective task where these roles can be easily encountered given the large availability of datasets [6].

We address two research questions:

RQ1 Can models be effectively steered via prompting toward a specific annotation perspective?

RQ2 Do role-consistent prompts produce more stable predictions on role-matched datasets?

Our contributions are twofold: (i) a theoretical framework of three person-grounded annotation roles applicable to NLP tasks broadly; and (ii) a systematic prompting study evaluating role alignment across multiple models, datasets, and prompt formulations with a focus on emotion classification [1]. Our findings reveal that existing datasets contain annotation role ambiguities previously unacknowledged, that prompting alone is insufficient to reliably induce perspective alignment, and that model behavior is surprisingly unstable, e.g., with larger models sometimes being outperformed by smaller ones.

2. Framework

Our framework starts from a fundamental observation: in the construction of annotated datasets, the perspective from which labels are assigned is frequently underspecified, leading to a systematic overlap of distinct interpretive roles. We argue that making this distinction explicit is critical for ensuring proper and consistent alignment between training and evaluation, i.e., for correctly verifying that the model performs the same role at test time that it was trained to perform.

We identify three roles:

1. Speaker Perspective (S): the author of the text annotates the internal state or intent they experienced at the moment of writing;

2. Perceived Speaker Perspective (PS) : a third-party annotator is asked to infer and label the internal state or intent of the author as conveyed by the text;

3. Observer Perspective (O) : a third-party annotator labels the reaction or effect that the text produces in them as a reader.

The framework builds on Buechel and Hahn [3], refining their writer/reader distinction by separating the writer role into two: a first-person author self-reporting their own emotion (Role S), and a third-party inferring the author's emotion (Role PS). We deliberately exclude their text-centered perspective as our framework focuses on cases where the annotation can be attributed to a specific person's perspective.

To better understand the defined roles, consider the following example: *"I can't believe they cancelled the last train again. Now I'll miss the whole thing."* Under **Role 1** , the author might label the text as expressing *frustration.* Under **Role 2** , an observer inferring the author's state might identify *disappointment.* Under **Role 3** , a reader might report feeling *sympathy* or *mild annoyance.*

Having established this tripartite conceptualization, the next step is mapping datasets onto these roles, not a straightforward operation since most datasets were not designed with this fine-grained level of role-specificity. The mapping is performed by recovering the original annotation guidelines and examining the precise wording of the questions posed to annotators.

[1]The full code is publicly available https://github.com/SDavenia/emotion-roles

3. Experimental Settings

3.1. Datasets

We built upon the benchmark by Koufakou et al. [6], which unifies 11 English single-labeled datasets [2]. The identified roles for each dataset are reported in Table 1, along with the procedure or task provided to the annotators. The table also summarizes the basis for assigning each dataset to a role, making explicit the perspective represented by the resulting annotations.

Dataset	Role	Annotation task / procedure
CARER [7]	S	Distant supervision from tweets collected using emotion-related hashtags constructed by the authors, e.g., `#depressed` and `#grief` for sadness, or `#fear` and `#worried` for fear.
Covid-worry [8]	S	Participants in a UK survey were asked to indicate their own emotions and express these in text.
enISEAR [9]	S	Participants were instructed to provide a description of an event for which they had experienced a particular emotion.
GoEmotions [10]	PS	Crowdworkers annotated Reddit comments, requiring them to infer the emotional state conveyed by the author.
TweetEval [11]	PS	Based on Affect in Tweets [12]. The emotion labels were assigned by annotators based on which best described the emotional state of the tweeter.
Github-love [13]	O	GitHub comments on pull requests and issues were annotated by the dataset authors for the emotional response they elicited.
GoodNews [14]	O	English news headlines were annotated through crowd-sourcing for the emotional response they elicited.
StackOv-GS [15]	O	Stack Overflow questions, answers, and comments were annotated by volunteers for the emotional response they elicited.
EmoEvent-EN [16]	O	Tweets were selected based on affective content, and were subsequently annotated through crowdsourcing for the elicited emotion.
Universal Joy [17]	NA	The emotion set was used as search strings for Facebook posts.
WASSA-21 [18]	NA	The various text instances were auto-labeled and then human corrected.

Table 1
Datasets included in the role analysis. The Role column indicates the annotation role assigned under our taxonomy: Speaker (S), Perceived Speaker (PS), or Observer (O). The annotation task/procedure summarizes how the labels were obtained and the perspective represented by the annotations. The two datasets which were not mapped to the proposed roles are shown in gray.

The datasets vary in text source and annotation strategy: CARER uses distant supervision via hashtags, while Covid-worry relies on self-reported essays. Since original emotion labels differ across datasets, Koufakou et al. [6] mapped them to a common label set, which we adopt; some emotions (e.g., Anger, Sadness) appear across most datasets, others (e.g., Trust, Anticipation) are rare. We sampled up to 128 instances per dataset, balancing emotion labels within each, yielding 1134 records total. As not all emotions appear in every dataset, the overall distribution is uneven.

The emotions in the considered dataset are, in alphabetical order: Anger, Anticipation, Disgust, Fear, Joy, Love, Sadness, Surprise, Trust and a Neutral option.

3.2. Models

We consider instruction-tuned models from two families, `llama-3` [19] and `gemma-3` [20], across a range of sizes that fit on a single `NVIDIA A100-PCIE-40GB`, due to computational limitations. We use 5 models, namely `Llama-3.2-3B/3.1-8B-Instruct` and `gemma-3-4b/12b/27b-it`.

[2]Universal Joy and WASSA-21 datasets were removed as the annotation guidelines did not allow us to map the proposed role-classification.

Example Speaker prompt

```
(system) Answer with one emotion from [<List of emotions>].
(user) You said: "<Text>".
(model) Emotion you were feeling:
```

Example Perceived Speaker prompt

```
(system) Answer with one of the emotions from [<List of emotions>].
(user) John wrote this: "<Text>".
(model) John was feeling:
```

Example Observer prompt

```
(system) Answer with one of the emotions from [<List of emotions>].
(user) A person reads the following text: "<Text>".
(model) This text makes the reader feel:
```

Figure 1: Example of the three role-aware prompts.

3.3. Setup

In order to investigate our research questions, we experiment with prompts that reflect the different annotation roles in the various datasets considered. Adhering to the proposed framework, we name the prompts according to the formalized roles.

For each group, we consider 5 prompts positioning the model in the specific annotation role and report results averaged across those. All prompts are reported in the repository, with examples shown in Figure 1. We run each prompt 5 times with the order of emotions shuffled. The label is then selected using majority voting on these 5 runs, with ties being broken at random in a few cases.

To obtain the output, we take the emotion token assigned the highest probability among all candidate emotions, having first verified that each emotion is tokenized as a single token. To prevent the model from adding explanations or preamble, we constrain its generation by prepending a fixed prefix (e.g., `"Emotion: "`) to the model's input.

3.4. Metrics

We evaluate role-aware prompting along two axes. To assess performance (RQ1), we report the macro-averaged F1 score, which weights each emotion class equally. To assess stability (RQ2), we report the Coefficient of Variation (CV) of the F1 scores across the 5 prompts of each role type, capturing how sensitive a model's performance is to prompt phrasing within a fixed role. In all tables, we underline the best result per model and bold the best result overall. Under our hypothesis, we would expect that prompting a model with the role matching the dataset's annotation perspective should yield higher F1, lower CV, or both.

4. Results

The F1 results are shown in Table 2. For the Llama models, the role-matched prompt achieves the best F1 on only 4 of 9 datasets for the 3B model and 3 of 9 for the 8B model. For the 8B model, the Observer prompt never wins on any of the four Observer datasets, while the Speaker prompt outperforms the others regardless of the dataset's annotation role.

The Gemma models exhibit similarly weak alignment: the role-matched prompt performs best in 4 of 9 settings for the 4B and 12B models, and in 5 of 9 settings for the 27B model. The Speaker prompt tends to dominate irrespective of annotation role; this is most evident for the 12B model, where it is the

Dataset	Role	Llama-3.2-3B			Llama-3.1-8B			Gemma-3-4B-it			Gemma-3-12B-it			Gemma-3-27B-it		
		S	PS	O	S	PS	O	S	PS	O	S	PS	O	S	PS	O
CARER [7]	S	.291	**.306**	.287	.302	.294	.267	.280	.259	.254	.287	.274	.243	.257	.247	.215
Covid-worry [8]	S	.251	**.275**	.269	.248	.253	.246	.234	.224	.219	.239	.236	.216	.251	.255	.224
enISEAR [9]	S	**.539**	.514	.498	.443	.427	.417	.463	.414	.424	.521	.516	.484	.502	.499	.458
GoEmotions [10]	PS	.373	.346	.363	.388	.383	.369	.344	.390	**.424**	.391	.384	.342	.332	.335	.308
TweetEval [11]	PS	.231	**.235**	.225	.205	.222	.217	.229	.217	**.235**	.215	.214	.213	.209	.219	.205
Github-love [13]	O	.165	.169	**.182**	.158	.154	.148	.137	.114	.121	.148	.138	.127	.153	.160	.147
GoodNews [14]	O	.260	.270	.255	.258	.215	.229	.213	.219	.230	**.283**	.271	.253	.225	.219	.209
StackOv-GS [15]	O	.240	**.257**	.250	.217	.204	.203	.242	.205	.202	.232	.231	.214	.204	.215	.203
EmoEvent-EN [16]	O	.295	.332	.339	.334	.279	.299	.364	.323	.352	.370	.369	**.386**	.339	.334	.340

Table 2
Macro-F1 scores (↑) across datasets averaged across prompts of the same role.

Dataset	Role	Llama-3.2-3B			Llama-3.1-8B			Gemma-3-4B-it			Gemma-3-12B-it			Gemma-3-27B-it		
		S	PS	O	S	PS	O	S	PS	O	S	PS	O	S	PS	O
CARER [7]	S	.047	.034	.077	.065	.038	.098	.059	**.026**	.052	.085	.139	.135	.066	.137	.184
Covid-worry [8]	S	.091	.139	.134	.032	**.019**	.061	.022	.055	.058	.056	.048	.186	.042	.024	.199
enISEAR [9]	S	.166	.094	.210	.129	.096	.093	.210	**.046**	.166	.183	.122	.151	.080	.087	.159
GoEmotions [10]	PS	.074	.048	**.028**	.139	.070	.139	.155	.074	.092	.136	.099	.054	.148	.115	.172
TweetEval [11]	PS	.116	.094	.108	.063	.028	.116	.044	.057	.029	.064	.026	.091	.021	**.016**	.076
Github-love [13]	O	.115	.071	.105	.058	.108	.139	.164	.120	.205	.058	.074	.078	.066	**.054**	.144
GoodNews [14]	O	.094	.081	.084	.087	.119	.079	.127	.138	.075	.085	.062	.120	.105	**.031**	.172
StackOv-GS [15]	O	.087	.073	.113	.098	.063	.080	.063	.099	.057	.073	.024	.174	.065	**.022**	.069
EmoEvent-EN [16]	O	.087	.076	.056	.096	.246	.123	.097	.093	.041	.083	.064	.055	**.035**	.164	.037

Table 3
Coefficients of Variation (↓) across datasets averaged across prompts of the same role.

best strategy on all 9 datasets and the 27B model behaves similarly, with Speaker or Perceived Speaker prompts achieving highest scores. Across both families, the Observer prompt is rarely the best choice even on Observer-annotated data, while S and PS prompts tend to win regardless of the dataset's true role. We also note that larger models do not yield better performance: the smallest models (3B, 4B) achieve the highest F1 on a majority of datasets, and Gemma-27B is often the weakest of its family.

The CV results are shown in Table 3. Role-matched prompts do not yield more stable predictions. For the Llama models, the role-matched prompt is the least variable on only 2 of 9 datasets (3B) and 3 of 9 (8B), with no prompt type achieving consistently higher stability. The Gemma models show the same picture, with the role-matched prompt being most stable on 5, 3, and 4 of 9 datasets for the 4B, 12B, and 27B models. No prompt type emerges as consistently more stable across datasets, with the only partial exception being the Perceived Speaker prompt on the 27B model.

TAKEAWAY: It appears that regardless of the annotation type, certain roles lead to better performance for the models, with a dominance of the Speaker and Perceived Speaker roles. However, role-matched prompts do not systematically improve performance, providing no evidence that prompting alone reliably induces behavior consistent with the intended annotation role (RQ1). The stability analysis provides no evidence that role-matched prompts produce more consistent predictions (RQ2).

5. Outlook

In this work, we introduced a formal tripartite classification of annotation roles (Speaker, Perceived Speaker, and Observer) and applied it retrospectively to an existing emotion recognition benchmark not originally designed with such distinctions. Although our empirical focus is on emotion recognition, the

proposed roles are straightforwardly applicable to other NLP classification settings where the subjective perspective of the annotator impacts the nature of the label. We then conducted a systematic prompting study across five models, testing whether explicitly aligning prompt roles to dataset annotation roles produces measurable gains in performance and stability.

Under our research hypothesis, role-aligned prompts should yield better performance and more stable predictions on role-matched datasets. Our experiments across a diverse set of datasets and models reveal mixed and inconclusive results: no consistent benefit emerges from matching prompt role to dataset role, and model behavior proves surprisingly unstable across prompt variations, with larger models sometimes underperforming smaller ones. These findings suggest that perspective alignment cannot be achieved through surface-level prompt framing alone, and point to deeper methodological gaps in how annotation roles are currently treated in dataset construction and model evaluation. We hypothesize that these results may reflect a combination of models being differentially aligned to certain prompt styles through instruction tuning, and the intrinsic variability introduced by prompt formulation choices [21]. Overall, **our findings highlight the importance of explicitly accounting for annotation, model, and prompt roles when training and evaluating language models** on NLP classification tasks, a dimension that current practice largely overlooks.

Several directions remain open for future work. At the experimental level, extending this investigation to additional datasets and tasks beyond emotion recognition would help determine whether the patterns observed here generalize. At the data level, constructing a resource that systematically collects all three types of annotation for the same instances would provide a cleaner testbed for role alignment hypotheses. At the methodological level, developing preprocessing steps that assess whether annotation role differences are statistically meaningful before applying role-specific treatments would make the framework more robust and scalable. Additionally, the consistent advantage of Speaker prompts also warrants further investigation. Speaker prompts often perform best even on datasets assigned to the Observer role, raising the possibility that this advantage reflects a general prompting or instruction-tuning effect rather than perspective alignment. More specifically, first-person framing may be more natural for instruction-tuned models to follow. Given the small number of sampled instances per dataset, these differences should be interpreted cautiously, as we do not report additional uncertainty estimates or significance tests. Future work could investigate whether the advantage of Speaker prompts persists across larger samples and models. Finally, going beyond prompt-level steering toward deeper intervention techniques, such as steering vectors or representation-level control, may be necessary to reliably induce role alignment in language models.

Acknowledgments

This work was funded by the partnership with Amazon Science "Multilingual personalization through perspective-aware Language Modeling" and the project NEIKEA (Bando CSP TRAPEZIO - Linea 1 - Paving the way to research excellence and talent attraction).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to improve writing style, assist with code development, refactoring, and documentation drafting. These tools were used to accelerate implementation and improve code readability. All outputs were reviewed, verified, and edited by the authors, who take full responsibility for the content of this paper. Generative AI was not used for data analysis, result interpretation, or the formulation of scientific claims.

References

[1] S. Mohammad, Best practices in the creation and use of emotion lexicons, in: A. Vlachos, I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1825–1836. URL: https://aclanthology.org/2023.findings-eacl.136/. doi:10.18653/v1/2023.findings-eacl.136.

[2] H. Jang, D. Frassinelli, The difficult case of intended and perceived sarcasm: a challenge for humans and large language models, in: K. Evang, L. Kallmeyer, S. Pogodalla (Eds.), Proceedings of the 16th International Conference on Computational Semantics, Association for Computational Linguistics, Düsseldorf, Germany, 2025, pp. 269–281. URL: https://aclanthology.org/2025.iwcs-main.23/.

[3] S. Buechel, U. Hahn, Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation, in: N. Schneider, N. Xue (Eds.), Proceedings of the 11th Linguistic Annotation Workshop, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–12. URL: https://aclanthology.org/W17-0801/. doi:10.18653/v1/W17-0801.

[4] S. Buechel, U. Hahn, EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 578–585. URL: https://aclanthology.org/E17-2092/.

[5] T. Meinert, V. Basile, A. Koufakou, The role of emotions: Investigating communicative roles in models and data for emotion recognition, The International FLAIRS Conference Proceedings 39 (2026).

[6] A. Koufakou, E. Nieves, J. Peller, Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora, in: Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 196–206.

[7] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, Y.-S. Chen, CARER: Contextualized affect representations for emotion recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3687–3697.

[8] I. van der Vegt, B. Kleinberg, A multi-modal panel dataset to understand the psychological impact of the pandemic, Scientific Data 10 (2023) 537.

[9] E. Troiano, S. Padó, R. Klinger, Crowdsourcing and validating event-focused emotion corpora for german and english, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4005–4011.

[10] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4040–4054.

[11] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020, pp. 1644–1650.

[12] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, M. Carpuat (Eds.), Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1–17. URL: https://aclanthology.org/S18-1001/. doi:10.18653/v1/S18-1001.

[13] M. M. Imran, Y. Jain, P. Chatterjee, K. Damevski, Data augmentation for improving emotion recognition in software engineering communication, in: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, 2022, pp. 1–13.

[14] L. A. M. Oberländer, E. Kim, R. Klinger, Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 1554–1566.

[15] N. Novielli, F. Calefato, F. Lanubile, A gold standard for emotion annotation in stack overflow, in: Proceedings of the 15th international conference on mining software repositories, 2018, pp. 14–17.

[16] F. Plaza-del-Arco, C. Strapparava, L. A. Urena-Lopez, M. T. Martin-Valdivia, EmoEvent: A Multilingual Emotion Corpus based on different Events, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020.

[17] S. Lamprinidis, F. Bianchi, D. Hardt, D. Hovy, Universal joy a data set and results for classifying emotions across languages, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Online, 2021, pp. 62–75.

[18] S. Tafreshi, O. De Clercq, V. Barriere, S. Buechel, J. Sedoc, A. Balahur, Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories, in: Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[19] Llama Team, The llama 3 herd of models, 2024. URL: https://doi.org/10.48550/arXiv.2407.21783. doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[20] Gemma Team, Gemma 3 (2025). URL: https://goo.gle/Gemma3Report.

[21] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031. URL: https://aclanthology.org/2021.tacl-1.60/. doi:10.1162/tacl_a_00410.