

Leveraging Large Language Models for Automated Definition Extraction with TaxoMatic - a Case Study on Media Bias

Timo Spinde¹, Luyang Lin², Smi Hinterreiter³, Isao Echizen¹

¹National Institute of Informatics, Japan

²The Chinese University of Hong Kong, China

³University of Würzburg, Germany

t.spinde@media-bias-research.org, lylin@se.cuhk.edu.hk, smi.hinterreiter@uni-wuerzburg.de, iechizen@nii.ac.jp

Abstract

Defining complex, evolving concepts in academic research and extracting clear taxonomies from many publications is challenging. To streamline systematic reviews and capture shifts in conceptual understanding, we present our ongoing work on TaxoMatic - a framework leveraging Large Language Models (LLMs) to automate definition extraction from academic literature. The framework encompasses data collection, relevance classification to identify papers with definitions, and definition extraction using LLMs. As a first case study, we tested our relevancy evaluation component on 2,398 articles on media bias, a domain particularly rich in varying definitions and sub-concepts. Then, we evaluated our definition extraction component on manually reviewed papers, yielding 123 definitions from 113 relevant articles. Among five tested LLMs, Claude-3-sonnet achieved the highest F1 score (0.381) for relevance classification and demonstrated a median cosine similarity of 0.557 for definition extraction with role prompting. Future directions include improving relevance classification, expanding ground truth datasets, and applying this framework to other domains, potentially enhancing conceptual clarity across disciplines.

Code/Dataset —

<https://github.com/Media-Bias-Group/Taxomatic>

1 Introduction

Defining concepts and building taxonomies is a foundational research task, as it ensures methodological clarity and facilitates interdisciplinary communication (Spinde et al. 2023). However, extracting clear, systematic definitions from academic literature remains a significant challenge across domains, especially given the growing complexity of concepts and the rapid expansion of research output (Fel et al. 2024).

Recent advances in large language models (LLMs) present new opportunities to streamline definition identification in academic research (Banerjee, Chakravarthi, and McCrae 2024). Although LLMs have been explored for various information extraction and analysis tasks, their application to systematically extracting definitions and supporting conceptual clarity has not yet been largely investigated (Banerjee, Chakravarthi, and McCrae 2024).

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This study proposes and evaluates initial steps toward building TaxoMatic, a framework for automated definition extraction using LLMs. To assess its feasibility and reliability, we apply the individual parts of TaxoMatic to the domain of media bias, a concept widely studied in communication, political science, and computational linguistics but still lacking universally accepted definitions (Spinde et al. 2022; Wessel et al. 2023; Horych et al. 2024). We address the following research questions:

- **RQ1:** How accurately can LLMs evaluate the relevance of academic publications on media bias compared to human assessments?
- **RQ2:** How do LLM and human-extracted definitions compare in content and semantic similarity?

We developed a three-stage workflow encompassing relevance classification, definition extraction, and evaluation. To create a ground truth, we collected 75,151 related scientific publications and manually rated 2,398 for relevancy to media bias research. From 113 deemed relevant, we manually extracted 123 definitions (Section 3).

The key contributions of this work include:

- Presenting an LLM-based process for systematic definition extraction from academic literature.
- Demonstrating the process’s utility in extracting definitions of complex phenomena, focusing on media bias.
- Providing a dataset to evaluate future frameworks focusing on relevancy analysis and definition extraction.

To ensure the dataset adheres to FAIR principles, we make all resources used in the process available (see the link after the abstract), use persistent identifiers for future updates, and maintain open, standardized formats to enhance interoperability and reusability.

2 Related Work

Definitions are the core of academic research, providing consistent communication and interpretation. When well established, researchers can engage in shared dialogue and consistently explore the same phenomenon (Navigli, Velardi, and Ruiz-Martínez 2010). Many domains, especially those studying human behavior, culture, or communication, face challenges in agreeing on unified definitions (Spinde

et al. 2023). The phenomena they examine are often subjective, context-dependent, and influenced by multiple factors. For instance, in sociology and political science, concepts like democracy or social justice vary across cultural and ideological contexts. Similarly, media studies lack consensus on media bias (Spinde 2021), with some focusing on visible bias like partisan reporting (Milyo and Groseclose 2005), while others define media bias as a linguistic concept (Spinde et al. 2021a). The definitional fragmentation complicates comparison across studies (Spinde et al. 2023). Despite their importance, clear definitions are difficult to establish. Systematic reviews, which gather, analyze, and synthesize literature, often begin taxonomy-building but are labor-intensive and subjective. Researchers’ biases can influence the definitions, creating inconsistencies across studies (Krippendorff 2019). Additionally, manual analysis becomes impractical as datasets grow, slowing taxonomy development.

Traditional computational approaches for term extraction struggle with open-text documents due to their inability to effectively handle unstructured formats and context-dependent relationships (Bovi, Telesca, and Navigli 2015). Advances in LLMs like GPT (Brown et al. 2020) and Mistral (Jiang et al. 2023), built on transformer architectures, now enable automated definition extraction and support tasks like taxonomy building. LLMs excel at in-context learning, outperforming fine-tuned and unsupervised extractors in various domains, and their performance improves significantly with careful prompt engineering and experimenting with various prompt strategies (Banerjee, Chakravarthi, and McCrae 2024). Still, their exact quality and reliability in extracting definitions and building taxonomies are yet unclear (Banerjee, Chakravarthi, and McCrae 2024). Challenges persist, especially in domains with ambiguous or evolving terminologies and issues like hallucinations or dependency on predefined taxonomies (Banerjee, Chakravarthi, and McCrae 2024).

Various datasets exist to explore definition extraction. The WCL dataset includes 5,000 sentences of explicit definitions from Wikipedia, limited to structured content (Navigli, Velardi, and Ruiz-Martínez 2010). The DEFT corpus, for SemEval-2010, covers 4,000 annotated sentences but lacks implicit or evolving definitions (Spala et al. 2020). DefIE contains 85,000 definitions from free-text sources across diverse domains (Bovi, Telesca, and Navigli 2015). General dictionaries, like Oxford or Urban Dictionary¹, collect various kinds of definitions (Oxford University Press 2025; Urban Dictionary 2025). Despite their value, all these datasets have limitations: (1) a lack of implicit or contested definitions (DEFT) (Spala et al. 2020), (2) focus on pre-structured content (WCL) (Navigli, Velardi, and Ruiz-Martínez 2010), (3) inconsistent quality (Urban Dictionary) (Urban Dictionary 2025), (4) limited adaptability to complex domains (all), and (5) no focus on academic definitions (all). These issues hinder their use for detailed, context-rich tasks, prompting the creation of our new dataset.

¹Which also exhibit inconsistent quality and informal style.

3 Methodology

We show the general workflow of our TaxoMatic framework and the underlying case study in Section 3. Our process includes multiple steps of a systematic literature search: (1) A keyword-based search for literature, (2) an assessment of which of the search results are relevant to the topic at hand, and (3) the extraction of required information — in our case, definitions. Two intermediate steps are required to process this data with LLMs and to create our ground truth for evaluation. We describe all the steps in the following.

Step 1: Searching Articles			
1. Concepts list (4,200 keywords)	2. Keyword review (1,096 keywords)	3. Semantic Scholar: 578,447 papers	4. 75,151 open-access PDFs (no duplicates)
Intermediate step: Data Preprocessing			
1. PDF extraction via GROBID (63,038 XML files, ignoring files with errors)			
Intermediate step: Manual Ground Truth Preparation			
1. 2,398 publications manually annotated for relevancy (Only articles with 100 or more citations)		2. 123 definitions manually extracted from 113 relevant articles	
Step 2: Relevance Classification		Step 3: Definition Extraction	
Automated labeling with 5 LLMs, 8 techniques (2,389 articles)		Automated labeling with Claude 3 Sonnet, 5 techniques (113 articles)	

Table 1: Workflow for the Definition Extraction Framework

Dataset

Data Collection For our case study, we chose the media bias domain because of its large diversity and ambiguity of definitions (Spinde et al. 2023; Spinde 2021; Spinde et al. 2021b), and collected data from Semantic Scholar using a keyword-based search. As a seed for our search, we used the 21 terms from an existing but limited media bias taxonomy (Spinde et al. 2023). Aiming to cover as many media bias-related terms as possible, we expanded the list using GPT-3.5-turbo and generated 200 similar terms for each keyword². After removing duplicates, this resulted in a list of 1,096 unique keywords. For each keyword, we crawled 1,000 results and eliminated duplicates and articles with fewer than 50 citations³. Finally, we downloaded 75.151 open-access papers in PDF format.

Data Preprocessing To enable LLM processing of the PDF contents⁴, we performed PDF information extraction with GROBID (Lopez 2008–2025). After manual verification and adjustments, we successfully processed 83.8% of the PDFs (63,038 papers) into XML format for further use.

Manual Definition Extraction To evaluate LLM performance, we manually annotated a ground truth. Due to limited reviewing capacity, we filtered the dataset by citation

²4,200 keywords in total, including duplicates.

³We aim to add these in the future, but due to limited resources, initial filtering was required.

⁴We experimented with pasting entire PDFs, but most models do not allow it, and those that do showed poor performance. Using extracted text significantly improved results.

count (Bornmann and Daniel 2008), selecting 2,398 articles with 100 or more citations. Six individuals, aged 25-35, with academic backgrounds and at least 6 months of media bias experience, followed two steps: reviewing titles and abstracts for relevancy and extracting or summarizing definitions from relevant full texts. Relevance was rated by one reviewer. Definition extraction involved a first reviewer and a second approving or modifying the result. This process produced 123 definitions from 113 relevant papers. While we believe the dataset size to be reasonable for our current evaluation, we aim to extend it, as discussed in Section 5.

Publication Assessment & Extraction

LLM Selection We selected five LLMs for the relevance classification task⁵, GPT-3.5-turbo (Brown et al. 2020), Mistral 7B Instruct v0.2 (Jiang et al. 2023), Vicuna 13b v1.5 (Zheng et al. 2023), Openchat 3.6 8b(Wang et al. 2024), and Claude 3 Sonnet (Anthropic 2023).

Prompting Strategies We designed our prompting strategies based on insights from prior research on prompt engineering, such as the importance of reasoning through Chain-of-Thought (CoT)⁶ prompting (Wei et al. 2022) and the effectiveness of giving contextually relevant examples (Zhou et al. 2023). We used eight prompting strategies for the relevance classification, shown with examples in Appendix A. To select the four examples for the few-shot prompt, we applied two sampling strategies to ensure both relevance and diversity. First, for similarity sampling, we used the KATE (Knn-Augmented in-conText Example selection) strategy (Liu et al. 2021), which identifies the most semantically or lexically similar examples based on Sentence-BERT (SBERT) embeddings (Reimers and Gurevych 2019). This ensured the selected examples closely matched the input context. Second, to enhance diversity, we applied k-means clustering to group the SBERT embeddings into clusters. Then, we manually selected two "relevant" and two "not relevant" examples from distinct clusters to capture a broader range of scenarios. For the definition extraction, we only use Claude 3 Sonnet with five of the eight strategies, namely Zero-Shot, Contextual Casual, Contextual Academic, CoT, and Role, since they focus on guiding the model’s comprehension rather than relying on sampling strategies.

Experimental Setup We evaluated the relevance analysis with the 2,398 manually rated articles. Then, we analyzed the definition extraction using the 123 ground truth definitions from the 113 publications (see Section 3). Any assessment was run three times per model and prompting combination on Google Colab with the L4 GPU runtime using the Haystack library⁷.

⁵We acknowledge the models change rapidly. We selected models based on the Huggingface Leaderboard and will update them in the future; see Section 5.

⁶CoT prompting encourages the model to reason step-by-step, improving performance (Wei et al. 2022).

⁷See <https://github.com/deepset-ai/haystack>.

Evaluation

Relevance Classification We measured classification performance using Precision, Recall, F1-score, and Accuracy, and assessed label consistency with Krippendorff’s Alpha (Krippendorff 2018) per prompt.

Automated Definitions Extraction First, we used Sentence-BERT (SBERT) (Reimers and Gurevych 2019) to embed LLM-extracted and manually extracted definitions and calculated their cosine similarity. Next, we applied a similarity threshold to classify matches, enabling the computation of precision, recall, and F1-score.

4 Results

Relevance Classification

As we show in Table 2, Claude outperformed other models, achieving the highest F1-score (0.381) with balanced precision (0.440) and recall (0.482). OpenChat achieved the highest accuracy (0.803) but with less balanced precision (0.347) and recall (0.417), indicating a tendency to over-classify irrelevant items as relevant. Vicuna had the lowest accuracy (0.133), while Mistral recorded the lowest F1-score (0.100), making both unsuitable for relevance classification. ChatGPT performed average, with an F1-score of 0.216.

LLM	F1 Score	Accuracy	Precision	Recall
ChatGPT-3.5	0.216	0.485	0.353	0.399
Mistral-7B	0.100	0.200	0.283	0.302
OpenChat-3.6	0.339	0.803	0.347	0.417
Claude-3-sonnet	0.381	0.672	0.440	0.482
Vicuna-13B	0.102	0.133	0.377	0.379

Table 2: Average Relevance Classification Performance by Model

We show the results of different prompting strategies in Table 3. CoT performs best across all four metrics, followed by Role Prompting, demonstrating that step-by-step reasoning or adopting an expert role enhances performance. However, providing examples lowers performance across all metrics, especially with similar examples. Academic Contextual slightly outperforms Casual Contextual.

Prompting Strategy	F1 Score	Accuracy	Precision	Recall
Zero-shot	0.251	0.563	0.334	0.383
Contextual Similar Casual	0.156	0.248	0.391	0.419
Contextual Similar Academic	0.166	0.269	0.397	0.421
Contextual Diverse Casual	0.158	0.403	0.266	0.307
Contextual Diverse Academic	0.186	0.431	0.286	0.330
Chain-of-Thought (CoT)	0.340	0.663	0.448	0.450
Role	0.323	0.616	0.397	0.448
Emotional	0.243	0.477	0.360	0.410

Table 3: Average Relevance Classification Performance by Strategy

We find an overall Krippendorff’s Alpha of 0.162, suggesting some agreement. Across models, all five exhibit slightly negative values, indicating systematic disagreement based on the prompting technique. Among prompting techniques, CoT achieves the highest agreement (alpha = 0.702),

showing strong model alignment with step-by-step reasoning. In contrast, Contextual Diverse Casual records the lowest agreement (alpha = 0.344), reflecting greater variability in classifications due to diverse examples. More details of Krippendorff’s Alpha scores are shown in Appendix B.

Automated Definition Extraction

In our automated definition extraction experiments, we exclusively used the Claude model, as it demonstrated the strongest performance in the Relevance Classification task. The cosine similarity scores for various prompting strategies are presented in Table 4.

Role Prompting achieved the highest mean similarity score (0.540), followed closely by Zero-shot Prompting (0.527). These findings indicate that the model’s pre-trained knowledge was sufficient to grasp a broad understanding of media bias, even without additional contextual guidance. However, the relatively wide range of similarity scores reveals inconsistencies in the model’s ability to capture slight details, particularly when distinguishing between explicit and implicit definitions.

Prompting Strategy	Mean	Median	Min	Max
Zero-shot	0.527	0.548	0.084	0.940
Contextual Casual	0.508	0.525	0.138	0.895
Contextual Academic	0.519	0.516	0.091	0.876
Chain-of-Thought (CoT)	0.514	0.532	0.044	0.880
Role	0.540	0.557	0.053	0.895

Table 4: Cosine Similarity Scores for Definition Extraction by Different Prompting Strategies

We also applied a threshold-based approach using cosine similarity to provide a more intuitive evaluation. Definitions with a similarity score above a 0.5 threshold were considered a correct ground truth match. This thresholding approach highlights Role Prompting’s strengths, with the LLM achieving 70 correct definitions out of 113 (see Table 7).

Manual Error Analysis

As part of the evaluation, we performed a manual qualitative review and identified several common model errors:

Overly Broad Definitions In several cases, the LLM extracted overly general definitions that were semantically relevant but missed specific details of the bias discussed, particularly with zero-shot prompting.

Partial Definitions Some definitions, especially with CoT Prompting, were incomplete, likely due to missing information during step-by-step reasoning.

Incorrect Definitions Sometimes, the LLM extracted irrelevant or inaccurate text, misidentifying non-definitional content as definitions. This was observed more frequently in Contextual Casual Prompting, where the informal framing may have caused the model to focus on broader concepts.

5 Discussion

In this poster, we demonstrate the viability of using Large Language Models (LLMs) for extracting definitions from

academic literature, with a case study on media bias. LLMs like Claude-3-sonnet effectively identify explicit and implicit definitions. Specifically addressing **RQ1**, Claude-3 showed high agreement with human assessments in relevance classification but struggled with class imbalance and overclassification. Regarding **RQ2**, LLMs aligned well with human definitions via SBERT cosine similarity, though subtle phrasing was often missed. TaxoMatic shows potential for broader applications, and our dataset offers a valuable resource for evaluating academic definition extraction.

Limitations include a class imbalance in relevance classification, overclassification, and the small ground truth dataset, which affects evaluation robustness. Cosine similarity, while useful, may miss slight phrasing differences.

Future work will expand the dataset with expert annotations and synthetic data to address class imbalance. Improved prompting strategies (Wei et al. 2024) and techniques like ontology learning and multi-stage modeling (Ji et al. 2022) could enhance performance. Regularly updating models and integrating processes into an automated framework could make TaxoMatic a versatile tool across disciplines.

6 Conclusion

This poster introduced the initial steps and a dataset for TaxoMatic, an LLM-based framework for automating definition extraction from academic literature. Claude-3-sonnet achieved the best relevance classification performance while Chain-of-Thought and Role Prompting achieved the best definition extraction performance. Despite challenges like dataset imbalance and lexical variability, results show LLMs’ potential to enhance conceptual clarity.

Acknowledgments

The authors thank Diana Sharafeeva, Martin Spirit, Fei Wu, Dr. Lingzhi Wang, and Prof. Dr. David Garcia. This work was supported by DAAD IFI, the JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grants JP-MJCR18A6 and JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, and by JST K Program Grant JPMJKP24C2 Japan.

References

Anthropic. 2023. Claude (Oct 8 version) [Large language model]. <https://www.anthropic.com/>.

Banerjee, S.; Chakravarthi, B. R.; and McCrae, J. P. 2024. Large Language Models for Few-Shot Automatic Term Extraction. In Rapp, A.; Di Caro, L.; Meziane, F.; and Sugumaran, V., eds., *Natural Language Processing and Information Systems*, 137–150. Cham: Springer Nature Switzerland. ISBN 978-3-031-70239-6.

Bornmann, L.; and Daniel, H.-D. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1): 45–80.

Bovi, C. D.; Telesca, L.; and Navigli, R. 2015. Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis. *Transactions of the Association for Computational Linguistics*, 3: 529–543.

- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.
- Fel, T.; Boutin, V.; Moayeri, M.; Cadène, R.; Bethune, L.; Andéol, L.; Chalvidal, M.; and Serre, T. 2024. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*. Red Hook, NY, USA: Curran Associates Inc.
- Horych, T.; Mandl, C.; Ruas, T.; Greiner-Petter, A.; Gipp, B.; Aizawa, A.; and Spinde, T. 2024. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection. arXiv:2411.11081.
- Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Yu, P. S. 2022. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 494–514.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.
- Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Krippendorff, K. 2019. *Analytical constructs*. SAGE Publications, Inc., fourth edition edition.
- Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.
- Lopez, P. 2008–2025. GRO-BID. <https://github.com/kermitt2/grobid>. swlh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.
- Milyo, J.; and Groseclose, T. 2005. A Measure of Media Bias. *The Quarterly Journal of Economics*, 120: 1191–1237.
- Navigli, R.; Velardi, P.; and Ruiz-Martínez, J. M. 2010. An Annotated Dataset for Extracting Definitions and Hypernyms from the Web. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*. Valletta, Malta: European Language Resources Association (ELRA).
- Oxford University Press. 2025. Oxford English Dictionary Online. <https://www.oed.com/>. Accessed: 2025-01-10.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.
- Spala, S.; Miller, N.; Derroncourt, F.; and Dockhorn, C. 2020. SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus. In Herbelot, A.; Zhu, X.; Palmer, A.; Schneider, N.; May, J.; and Shutova, E., eds., *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 336–345. Barcelona (online): International Committee for Computational Linguistics.
- Spinde, T. 2021. An Interdisciplinary Approach for the Automated Detection and Visualization of Media Bias in News Articles. In *2021 IEEE International Conference on Data Mining Workshops (ICDMW)*.
- Spinde, T.; Hinterreiter, S.; Haak, F.; Ruas, T.; Giese, H.; Meuschke, N.; and Gipp, B. 2023. The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. *arXiv preprint*.
- Spinde, T.; Jeggle, C.; Haupt, M.; Gaissmaier, W.; and Giese, H. 2022. How do we raise media bias awareness effectively? Effects of visualizations to communicate bias. *PLOS ONE*, 17(4): 1–14. Publisher: Public Library of Science.
- Spinde, T.; Plank, M.; Krieger, J.-D.; Ruas, T.; Gipp, B.; and Aizawa, A. 2021a. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Dominican Republic.
- Spinde, T.; Rudnitskaia, L.; Mitrović, J.; Hamborg, F.; Granitzer, M.; Gipp, B.; and Donnay, K. 2021b. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3): 102505.
- Urban Dictionary. 2025. Urban Dictionary. <https://www.urbandictionary.com/>. Accessed: 2025-01-10.
- Wang, J.; Wang, C.; Tan, C.; Huang, J.; and Gao, M. 2024. Knowledgeable In-Context Tuning: Exploring and Exploiting Factual Knowledge for In-Context Learning. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3261–3280. Mexico City, Mexico: Association for Computational Linguistics.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Chain-of-thought prompting elicits reasoning in large language models.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wessel, M.; Horych, T.; Ruas, T.; Aizawa, A.; Gipp, B.; and Spinde, T. 2023. Introducing MBIB - the first Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. New York, NY, USA: ACM. ISBN 978-1-4503-9408-6.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*. Red Hook, NY, USA: Curran Associates Inc.
- Zhou, W.; Zhang, S.; Poon, H.; and Chen, M. 2023. Context-faithful Prompting for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14544–14556. Singapore: Association for Computational Linguistics.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, as it could facilitate finding and agreeing on definitions in different research domains.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, in Section 3.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **NA.** The study does not use population-specific data except for LLMs of Western origin.
 - (e) Did you describe the limitations of your work? **Yes, in Section 5.**
 - (f) Did you discuss any potential negative societal impacts of your work? The study evaluates LLM's ability to extract definitions. Hence, it could reinforce existing biases present in source materials, especially in domains with contested or politicized concepts. Additionally, reliance on LLMs may lead to overconfidence in automatically generated definitions.
 - (g) Did you discuss any potential misuse of your work? **No, besides the above mentioned bias and potential overconfidence.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **NA.** Code and data are publicly available. They do not contain personal data.
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **Yes, in Section 3 and Section 4.**
 - (b) Have you provided justifications for all theoretical results? **Yes, in Section 4.**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **Yes, in Section 5.**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **Yes.**
 - (b) Did you include complete proofs of all theoretical results? **Yes. All scripts and evaluations are available in the corresponding repository.**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, under the repository URL.**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes.**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes. See Section 3 and Section 4.**
 - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **NA.** We state the limitations of the definition extractions in Section 5.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? **Yes.**
 - (b) Did you mention the license of the assets? **This project is licensed under the Apache 2.0 License. See the LICENSE file in the repository for details.**
 - (c) Did you include any new assets in the supplemental material or as a URL? **Yes, under the repository URL.**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **NA.** Data only contains academic open source paper.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **NA.** Data does not contain personal data or offensive material.
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **Yes.**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **Yes, in Section 1.**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? **NA**

- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
- (d) Did you discuss how data is stored, shared, and de-identified? NA

A Prompting Strategies

1. **Zero-Shot:** Minimal guidance without prior examples.
2. **Contextual Similar Casual:** Similar examples in a casual context.
3. **Contextual Similar Academic:** Similar examples in an academic context.
4. **Contextual Diverse Casual:** Diverse examples in a casual context.
5. **Contextual Diverse Academic:** Diverse examples in an academic context.
6. **Chain-of-Thought Prompting (CoT):** Step-by-step reasoning.
7. **Role:** Model assumes the role of a media bias expert.
8. **Emotional:** Uses emotional stimuli for deeper engagement.

Examples for all of our prompts are published in the repository and below.

Tested Prompts

We used the following prompting strategies in our relevance assessment experiments. In each case, placeholders like [Article Title] and [Article Abstract] were replaced dynamically.

- **Zero-Shot Prompting**

Please determine if the following article is relevant to media bias research: [Article Title] - [Article Abstract]

- **Contextual Similar Casual**

Here are examples of articles relevant to media bias research: [Example 1], [Example 2]. Based on these, is the following article relevant? [Article Title] - [Article Abstract]

- **Contextual Similar Academic**

Considering the provided scholarly articles on media bias: [Example 1], [Example 2], assess the relevance of this article to media bias research: [Article Title] - [Article Abstract]

- **Contextual Diverse Casual**

We have diverse articles discussing various aspects of media studies: [Example 1], [Example 2]. Does the following article pertain to media

bias? [Article Title] - [Article Abstract]

- **Contextual Diverse Academic**

Given these diverse academic perspectives on media studies: [Example 1], [Example 2], evaluate if the following article is relevant to media bias research: [Article Title] - [Article Abstract]

- **Chain-of-Thought (CoT) Prompting**

To determine if the following article is relevant to media bias research, let's analyze it step-by-step: [Article Title] - [Article Abstract]

- **Role Prompting**

As a media bias expert, assess the relevance of this article to the field: [Article Title] - [Article Abstract]

- **Emotional Prompting**

Imagine you're passionate about uncovering media bias. Does this article excite your interest in media bias research? [Article Title] - [Article Abstract]

For the definition extraction, we used prompts as follows.

- **Zero-Shot Prompting**

Extract the definition of media bias from the following academic text: [Full Text]

- **Contextual Casual Prompting**

People often define media bias in different ways. Based on how it is discussed here, what is the definition? [Full Text]

- **Contextual Academic Prompting**

In scholarly research, definitions are often embedded in complex texts. Please extract a clear, concise definition of media bias from the following excerpt: [Full Text]

- **Chain-of-Thought (CoT) Prompting**

Let's identify the definition of media bias step by step. First, find any sentence that discusses the nature of media bias. Then, summarize that into a clear definition. Here is the article content: [Full Text]

- **Role Prompting**

You are a researcher in media studies. Based on the following academic text, please provide the clearest definition of media bias presented in the article: [Full Text]

B Additional Result Tables

LLM	Krippendorff's Alpha
ChatGPT-3.5	-0.027
Mistral-7B	-0.054
OpenChat-3.6	-0.032
Claude-3-sonnet	-0.063
Vicuna-13B	-0.108

Table 5: Krippendorff's Alpha per LLM across all prompting techniques

Prompting Strategy	Krippendorff's Alpha
Zero-shot	0.610
Contextual Similar Casual	0.575
Contextual Similar Academic	0.581
Contextual Diverse Casual	0.344
Contextual Diverse Academic	0.491
Chain-of-Thought (CoT)	0.702
Role	0.586
Emotional	0.512

Table 6: Krippendorff's Alpha per Prompting Technique across all Models

Prompting Strategy	Threshold 0.5	Threshold 0.6	Threshold 0.7
Zero-shot	62	44	25
Contextual Casual	61	37	14
Contextual Academic	59	36	19
CoT	59	44	26
Role	70	48	27

Table 7: Counts of Correct Definitions by Threshold Using Different Prompting Strategies