Frame In, Frame Out: Measuring Framing Bias in LLM-Generated News Summaries

Valeria Pastorino and Nafise Sadat Moosavi
Department of Computer Science
University of Sheffield (UK)
{vpastorino1|n.s.moosavi}@sheffield.ac.uk

Abstract

News headlines and summaries shape how events are interpreted through selective emphasis and omission, a phenomenon commonly referred to as framing. Large language models are now routinely used to generate such content, yet existing evaluation frameworks largely overlook this dimension. We introduce Frame In, Frame Out (FIFO), the first large-scale benchmark for measuring framing presence in LLM-generated news summaries, grounded in the widely used XSum dataset. FIFO combines 15,499 jury-annotated examples with 320 expert-labeled instances ($\kappa = 0.61$) to validate and calibrate model-based annotations. Using FIFO, we analyze measured framing rates across 27 summarization models. We find that LLM-generated summaries often exhibit higher calibrated framing rates than human-written references, with substantial variation across topics and training regimes, including elevated rates in scientific and public health summaries. Our results establish framing as an underexplored and consequential dimension of summarization quality.

1 Introduction

Framing is a communication strategy through which the interpretation of events can be shaped by selecting, emphasizing, omitting, and organizing information in particular ways (Entman, 1993; Goffman, 1974). From a semantic perspective, framing concerns meaning beyond propositional content: how lexical and discourse choices guide interpretation of the same underlying facts. In media and political communication, it influences attribution of responsibility (Iyengar, 1994), political tolerance (Nelson et al., 1997), and public opinion. As natural language generation systems are increasingly used to produce headlines, news briefs, and content summaries (Sadeghi and Arvanitis, 2023), they inherit a growing role in shaping public discourse.

Despite extensive study in the social sciences, framing remains largely absent from the evaluation of model-generated text. In our setting, this absence is not merely theoretical: a system-generated summary can introduce an interpretive lens that is absent from the human-written reference summary for the same article.

> **Gold summary (`Not Framed`):** "Donkey, water buffalo and goat meat have been sold as burgers and sausages in South Africa, a study says."
> **System summary (`Framed`):** "South Africa has been hit by a growing scandal over the sale of meat products that contain animals"

In these examples from our data, both statements can be compatible with the underlying facts, but the system summary foregrounds a more evaluative interpretation of the event, steering the reader toward a particular understanding of its significance. This kind of perspective shift is largely invisible to standard summarization metrics. In fact, evaluating natural language generation systems is an ongoing and challenging problem, with existing frameworks primarily focusing on factuality (Pagnoni et al., 2021; Huang et al., 2026), coherence (Fabbri et al., 2021), and coverage (Liu et al., 2023). These dimensions capture important aspects of quality, but do not account for how information is selectively emphasized or omitted in ways that may contribute to political agendas, polarization or distorted public discourse. As a result, a summary may be factually accurate and fluent, yet still convey a particular perspective in subtle and socially consequential ways (Puccetti et al., 2024).

We address this gap with **Frame In, Frame Out (FIFO)**, the first benchmark for evaluating framing bias in summarization. FIFO provides both expert-annotated gold labels and model-generated silver labels for thousands of summaries, along with a

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 378–384

calibration protocol to produce expert-calibrated framing estimates. Building on work in framing detection, we propose framing as an evaluation dimension for summarization, enabling systematic model- and topic-level comparisons of generated outputs.

Using FIFO, we analyze 27 summarization models spanning architectures and fine-tuning regimes. We find that several large language models produce higher calibrated framing rates than human-written baselines, with substantial variation by model size, training setup, and topic. Notably, elevated framing emerges even in domains such as science and public health, where neutrality is typically expected. These findings establish framing as a measurable and consequential property of LLM-generated summaries, and motivate its inclusion in future evaluation frameworks.

2 Related Work

Framing has mainly been studied as a supervised detection task using annotated corpora such as the Media Frames Corpus (Card et al., 2015) and GVFC (Liu et al., 2019). Earlier work relied on topic models and lexical cues (DiMaggio et al., 2013; Burscher et al., 2016), followed by transformer-based classifiers (Khanehzar et al., 2019; Naderi and Hirst, 2017). More recent approaches use LLMs such as GPT-4 to classify or explain frames in news text (Pastorino et al., 2026; Maab et al., 2024; Lin et al., 2024). Across these works, framing is treated as detection: the goal is to identify which frame a text expresses, if any. Our study differs by treating framing as an evaluation dimension, asking whether summarization systems introduce unintended or asymmetric framing.

Summarization evaluation has largely centered on coherence, factuality, content selection, and human preference judgments (Fabbri et al., 2021; Pagnoni et al., 2021; Liu et al., 2023; Bhandari et al., 2020). However, despite framing being central to how readers perceive neutrality, no existing benchmark evaluates framing as a dimension of generated summaries. Our work fills this gap by systematically analyzing framing in system-generated summaries, offering a new perspective on summarization quality.

3 FIFO: Framing Bias in Summarization

Our goal is to develop a scalable and reliable procedure for estimating framing in generated news summaries. This process results in **FIFO** (Frame In, Frame Out), a dataset composed of 15,499 model-generated summaries labeled via an LLM jury and 320 gold-standard annotations from human experts. We build on the best prompting strategy evaluated by Pastorino et al. (2026) to use LLMs to classify summaries as either `Framed` or `Not Framed`. We extend this approach in two key ways: (1) we introduce a human-annotated gold set to validate and estimate the reliability of model predictions, and (2) we estimate expert-calibrated reliability weights from the agreement between silver and gold labels, and use them to compute calibrated framing rates across models. This approach supports scalable, quantitatively grounded analysis anchored in expert judgment.

Summaries and Model Set Our analysis is based on the XSum dataset (Narayan et al., 2018), a single-sentence summarization corpus composed of BBC News articles and human-written summaries. Its extreme compression setting requires models to make strong content-selection and emphasis decisions, making it particularly well suited for studying framing in generated summaries. Because each output compresses a full article into a short summary, differences in what is foregrounded, backgrounded, or omitted become especially consequential. Moreover, XSum is a widely adopted benchmark, enabling systematic comparison across models and training regimes.

We analyse summaries from 27 systems sourced from Liu et al. (2024) and Panickssery et al. (2024), comprising a total of 15,499 summaries. These systems span a diverse range of architectures and training setups, including encoder-decoder models (e.g., BART (Lewis et al., 2020), T5 (Raffel et al., 2020), FLAN-T5 (Chung et al., 2024)), decoder-only models (e.g., GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), GPT-Neo (Black et al., 2021), Claude (Anthropic, 2024), Cohere (Cohere, 2024), LLaMA (Touvron et al., 2023)). Fine-tuned variants are labeled with -XSUM or -CNN, reflecting supervised training on XSum or CNN/DailyMail (Hermann et al., 2015), respectively[1].

Framing Operationalization We annotate framing as a binary property of the generated summary.

[1]All gold and silver framing annotations and model checkpoints (sourced from the original papers (Liu et al., 2024; Panickssery et al., 2024)) are available at https://github.com/vpastorino/FIFO

A summary is labeled `Framed` when it presents the article through an identifiable interpretive lens, including through selective emphasis or omission, evaluative wording, causal or moral attribution, responsibility assignment, or other discourse choices that make one interpretation more salient. A summary is labeled `Not Framed` when it conveys the core event without introducing such an interpretive lens. This binary formulation supports the benchmark's primary goal: estimating how often summarization systems produce framed outputs. Fine-grained frame taxonomies are useful for analyzing which type of frame is present, but the presence of framing is the necessary first-order question for system-level evaluation.

Silver Labeling via LLM Jury To assign framing labels at scale, we apply the best-performing prompt from Pastorino et al. (2026) to a three-member LLM jury comprising GPT-4.1-nano, GPT-4o, and GPT-3.5-Turbo. Each model independently labels a summary as either `Framed` or `Not Framed`, and the final label is assigned via majority vote. The resulting labels constitute our silver dataset, which is used for large-scale analysis of framing across models.

Expert Annotations and Jury Validation To evaluate the reliability of the silver labels, we randomly sampled 320 model-generated summaries and annotated them for framing bias. Annotation was performed by a domain expert in linguistics and manual framing analysis, using informed judgment grounded in established framing theory (Entman, 1993; Goffman, 1974). Any uncertainties were resolved with a second expert, and labels were assigned by agreement. We compared the resulting expert labels to the majority-vote predictions of the LLM jury, treating the expert annotations as the gold reference for validation. Agreement was measured using Cohen's κ (Cohen, 1960), resulting in a score of 0.616, which falls within the range of substantial agreement (Landis and Koch, 1977). This level of alignment indicates that the silver labels produced by the LLM ensemble capture framing-relevant distinctions with reasonable reliability.

Expert-Calibrated Reweighting To adjust for prediction error in the silver labels, we estimate how often each jury label aligns with expert judgments using the 320 expert-annotated examples. Among items labeled *Framed* by the jury, 77.8% were also labeled *Framed* by experts. Among items labeled *Not Framed* by the jury, 16.3% were nevertheless judged *Framed* by experts. We assign each silver headline the corresponding calibrated value and estimate framing rates as the mean of these values rather than as the mean of raw binary jury labels. For a subset of summaries S, we define the expert-calibrated framing rate $\mathrm{FR}(S)$ as:

$$\mathrm{FR}(S) = \frac{1}{|S|} \sum_{s \in S} w_s$$

where $|S|$ denotes the number of summaries in S, and w_s is the calibrated framing score assigned to summary s:

$$w_s = \begin{cases} 0.778 & \text{if } y_j(s) = F, \\ 0.163 & \text{if } y_j(s) = N. \end{cases}$$

Here, $y_j(s)$ denotes the jury label for summary s, F denotes *Framed*, and N denotes *Not Framed*.

4 Framing Behaviour Across Summarization Models

The FIFO dataset enables systematic, model-level analysis of framing in summarization outputs. Figure 1 presents expert-calibrated framing rates across 27 systems[2].

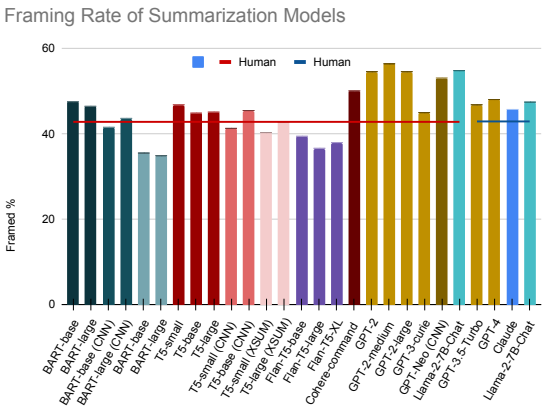

Figure 1: Expert-calibrated framing rates across 27 summarization systems. Red and blue lines indicate human-authored baselines for the two subsets.

These summaries are sourced from two prior model collections (Liu et al., 2024; Panickssery et al., 2024), each associated with a distinct portion of the XSum dataset; human-authored reference

[2]Unweighted results show similar trends but larger absolute differences and minor ranking shifts among mid-range models.

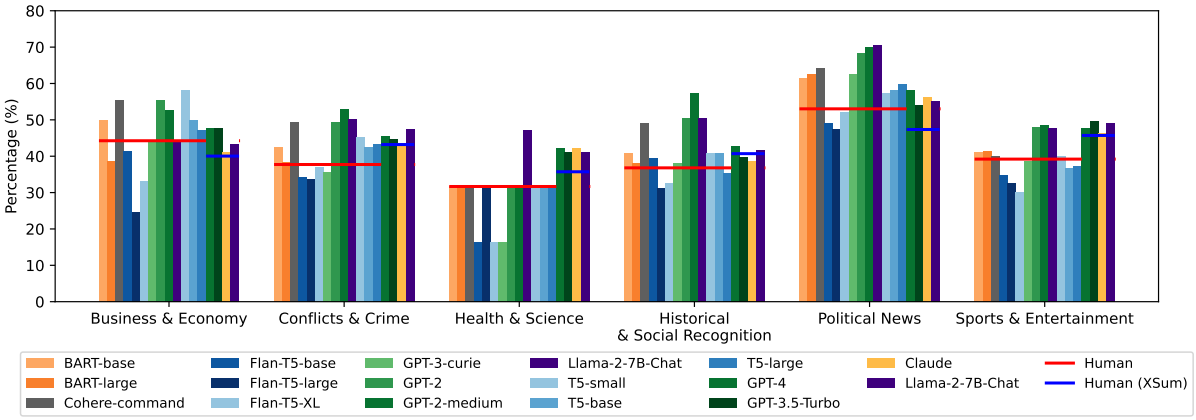

Figure 2: Framing rates by topic and model. Bars show the percentage of summaries labeled as `Framed`, grouped by topic. Each color corresponds to a model; horizontal lines indicate human framing baselines from two subsets.

summaries for each subset serve as framing baselines (indicated by red and blue horizontal lines).

Model Scale and Pretraining Scope Smaller models (e.g., BART, FLAN-T5 variants) exhibit lower framing rates, often below the human baseline. However, this often reflects lower output quality: low quality summaries may lack the structure necessary for framing to be meaningfully expressed or detected. In contrast, larger models (e.g., GPT-4, Claude, LLaMA) produce more fluent and linguistically-rich summaries and correspondingly show higher framing rates. This difference in framing rates between smaller and larger models is statistically significant (paired t-test, p = 0.0012).

Effect of Fine-Tuning Models fine-tuned on XSum exhibit lower framing rates than their base counterparts. This reduction is statistically significant (p = 0.0006; 95% CI: –19.27% to –7.78%) and suggests that task-specific finetuning is associated with lower framing rates in this setting.

Intra-Family Size Effects Within model families (e.g., BART, T5, GPT, FLAN-T5), we observe a moderate inverse correlation between model size and framing rate (Pearson $r = -0.44$), with larger variants tending to generate slightly less framed content. Together with the across-model trend above, this suggests that training regime and data may drive framing differences more strongly than parameter count alone.

These findings suggest that some high-capacity systems produce framing at rates above the corresponding human-authored baselines. This emphasizes the need for framing-aware evaluation in summarization, since standard quality metrics may miss interpretive shifts in otherwise fluent and factually compatible summaries.

5 Framing as a Function of Topic

To examine how framing varies across domains, we categorize each summary into one of six high-level topics: *Business & Economy*, *Conflicts & Crime*, *Health & Science*, *Historical & Social Recognition*, *Political News*, and *Sports & Entertainment* using a hybrid process. Framing rates are computed using the expert-calibrated reliability weights introduced in Section 3, ensuring that reported scores reflect validation against expert annotations. Figure 2 shows that framing rates vary substantially by topic and model capacity. Human-authored summaries display clear topic sensitivity, with *Health & Science* exhibiting the lowest framing rate and *Political News* the highest. Smaller models (e.g., BART, FLAN-T5 variants), which often produce lower-quality summaries, often fall below these human baselines across most topics. In contrast, larger, more capable models (e.g., GPT-4, Claude, LLaMA) frequently exceed human framing rates, particularly in *Political News*, where nearly all high-capacity models surpass the already elevated human baseline (≈53%). Even in *Health & Science*, where humans frame only 31% of instances, many larger models exceed this level. These findings show that framing in LLM-generated summaries is topic-dependent and often exceeds human norms in larger models, highlighting the need for framing-aware evaluation as such models are now widely used in large-scale content production, including news writing and headline generation (Puccetti et al., 2024).

Length vs. Framing Presence To investigate the relationship between text length and framing, we

computed a point-biserial correlation coefficient, which measures the association between a continuous variable (text length) and a dichotomous variable (framing status). Our analysis revealed a small yet statistically significant positive correlation, $r_{pb} \approx 0.1904$, indicating that longer texts are modestly associated with the presence of framing. Specifically, texts identified as `Framed` had an average length of 147 words, whereas those labeled `Not Framed` averaged 83 words. This indicates that length is associated with framing presence, but the small effect size suggests that framing is not reducible to length alone.

6 Conclusions

We introduced FIFO, a benchmark for evaluating framing in generated news summaries, combining expert-annotated gold labels with LLM-jury annotations that are reweighted using expert-calibrated reliability estimates. Across 27 systems, we found that several large language models produce higher framing rates than human-written baselines, with substantial variation by model capacity, training regime, and topic, including in domains such as health and science. These results highlight a limitation of current evaluation frameworks, which prioritize accuracy and fluency but largely ignore how information is presented and interpreted. Our findings demonstrate that framing is a measurable and socially consequential dimension of summary quality that is not captured by existing metrics. As LLMs are increasingly deployed in content production, neglecting this dimension risks reinforcing polarized or distorted narratives. By establishing framing as a tractable evaluation target, FIFO provides a foundation for incorporating perspective-sensitive analysis into future benchmarks and assessment practices.

7 Limitations

The findings of this study should be considered alongside some limitations. First, while our use of silver labels generated by a jury of large language models allows for broad coverage, these labels may reflect model-specific biases or blind spots. We mitigate this risk by validating the jury labels against a smaller expert-annotated gold set and using the observed expert–jury agreement to estimate aggregate framing rates.

Our study focuses on single-document summarization in English using the XSum dataset. This enables tight control over the input domain but leaves open questions about how framing manifests in more diverse linguistic and cultural settings.

Finally, our operationalization of framing is binary, distinguishing between `Framed` and `Not Framed` summaries. This choice matches FIFO's goal of estimating whether framing is present in generated summaries and how often systems produce framed outputs. However, it necessarily simplifies a complex, context-dependent phenomenon and does not identify fine-grained frame types.

8 Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We also acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic. Model card. Accessed: 2026-05-20.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-TensorFlow. Zenodo software release.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bjorn Burscher, Rens Vliegenthart, and Claes H. de Vreese. 2016. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5):530–545.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across

Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohere. 2024. An overview of Cohere's models. Accessed: 2026-05-20.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper colophon books. Harvard University Press.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701. Curran Associates, Inc.

Zhaoheng Huang, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2026. Evaluating the factuality of large language models using multiple plug-and-play fact sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 41607–41609.

Shanto Iyengar. 1994. *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, Chicago, IL.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050, St. Julian's, Malta. Association for Computational Linguistics.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. Media bias detection across families of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4083–4098, Mexico City, Mexico. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 536–542. Incoma Ltd. Shoumen, Bulgaria.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Thomas E. Nelson, Rosalee A. Clawson, and Zoe M. Oxley. 1997. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc.

Valeria Pastorino, Jasivan A. Sivakumar, and Nafise Sadat Moosavi. 2026. Decoding news narratives: A critical analysis of large language models in framing detection. *Preprint*, arXiv:2402.11621.

Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. Ai 'news' content farms are easy to make and hard to detect: A case study in italian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 15312–15338. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

McKenzie Sadeghi and Lorenzo Arvanitis. 2023. Rise of the newsbots: AI-generated news websites proliferating online. NewsGuard Special Report. Accessed: 2026-05-20.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.