MALTO at SemEval-2025 Task 3: Detecting Hallucinations in LLMs via Uncertainty Quantification and Larger Model Validation

Claudio Savelli and Alkis Koudounas and Flavio Giobergia
Politecnico di Torino, Italy
Turin, Italy
{firstname.lastname}@polito.it

Abstract

Large language models (LLMs) often produce *hallucinations* —factually incorrect statements that appear highly persuasive. These errors pose risks in fields like healthcare, law, and journalism. This paper presents our approach to the Mu-SHROOM shared task at SemEval 2025, which challenges researchers to detect hallucination spans in LLM outputs. We introduce a new method that combines probability-based analysis with Natural Language Inference to evaluate hallucinations at the word level. Our technique aims to better align with human judgments while working independently of the underlying model. Our experimental results demonstrate the effectiveness of this method compared to existing baselines.

1 Introduction

Large language models (LLMs) are widely used for various NLP tasks, such as information retrieval (Dai et al., 2024), medical queries (Singhal et al., 2025), and content generation (Coppolillo et al., 2024). Their ability to generate coherent and contextually relevant text has led to an increasing reliance on them as primary information sources, sometimes surpassing traditional methods like search engines, expert consultations, or structured databases (Dwivedi et al., 2023). This shift reflects the growing trust in LLMs for fast and accessible information.

However, a major challenge is their tendency to produce *hallucinations* — factually incorrect but highly persuasive outputs (Ji et al., 2023; Bertetto et al., 2024). These errors can take various forms, including false claims (D'Amico et al., 2023), fabricated references (La Quatra et al., 2021), and made-up biographies (Yuan et al., 2021), often presented in a way that makes them difficult to distinguish from accurate information. Since LLMs generate text based on patterns in their training data rather than direct verification of facts, they may confidently assert misinformation, leading to potential risks in sensitive domains such as healthcare (Bélisle-Pipon, 2024; La Quatra et al., 2025), law (Benedetto et al., 2024), and journalism (Spangher et al., 2024; Giobergia et al., 2024).

The problem is further amplified by the fact that hallucinations are often blended with accurate, truthful statements, making them harder to detect (Lewis et al., 2020; Borra et al., 2024). A model may produce a largely correct passage with only a few inaccurate details, increasing the likelihood that users will accept the entire output as trustworthy. Moreover, the possibility that models have to generate and deal with multiple languages (Huang et al., 2024; Savelli and Giobergia, 2024) can make evaluating these outputs even more complex. As LLMs become more advanced and widely deployed, addressing their tendency to hallucinate is critical to ensuring their reliability and safe integration into real-world applications.

To bring more attention to this issue, the **Mu**ltilingual **S**hared-task on **H**allucinations and **R**elated **O**bservable **O**vergeneration **M**istakes (*Mu-SHROOM*) has been introduced at SemEval 2025 (Vázquez et al., 2025). Mu-SHROOM invites researchers to detect hallucination spans in LLM outputs across multiple languages and models. The task specifically focuses on identifying which parts of a generated text contain hallucinations. Participants are provided with LLM outputs in different formats, including raw text, token lists, and logit values, and are tasked with predicting hallucination probabilities at the character level.

This work introduces a novel approach that operates at the word level to evaluate hallucinations[1]. By leveraging probability-based analysis and a Natural Language Inference (NLI) model, we compare each generated token to the most likely alternatives

[1]The code to replicate the experiments can be found at https://github.com/MAL-TO/Mu-SHROOM

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1318–1324

from a larger model, identifying potential hallucinations based on inconsistencies in predicted outputs. Our method aims to improve alignment with human annotations while remaining model-independent.

1.1 Mu-SHROOM

The objective of this task (Vázquez et al., 2025) is to identify spans of text within model-generated responses that represent hallucinations. Participants must determine which parts of a response produced by LLMs contain factual inaccuracies. The task is multilingual and multi-model, as it includes data from various languages[2] and outputs generated by different publicly available LLMs[3].

Dataset. The dataset consists of multiple fields that capture both the model-generated responses and the corresponding factuality annotations. Each data entry includes an *ID* for identification, a *language code* indicating the language of the query, and the *model input*, which represents the original question posed to the LLM. The *model output* contains the generated response, and the specific model that produced it is recorded under a *model ID*.

Two types of annotations are provided to assess the factual reliability of the model output: *soft labels* and *hard labels*. Soft labels represent a continuous evaluation of factual accuracy by assigning probability values to specific spans of text. These probabilities, ranging from 0 to 1, indicate the likelihood that a given segment is hallucinated. A lower probability suggests a higher likelihood of correctness, whereas a higher probability signals greater uncertainty or fabrication. Hard labels, on the other hand, offer a binary assessment of factual errors. They identify definitive hallucinations by marking specific spans of text that have been verified as incorrect. Each hard label is recorded as a pair of indices representing the start (inclusive) and end (exclusive) positions of the hallucinated text. For evaluation, hard labels are used to measure accuracy based on intersection-over-union (IoU), while soft labels are analyzed through Pearson correlation between system outputs and human ratings.

Each language has three different splits: an *unlabeled training set* containing raw samples without labels, a *labeled validation set* including soft and hard labels from the annotators, and the *test set* used for evaluation purposes. The three sets have 809, 50, and 154 samples, respectively.

[2]While the dataset covers 14 languages (Arabic (Modern standard), Basque, Catalan, Chinese (Mandarin), Czech, English, Farsi, Finnish, French, German, Hindi, Italian, Spanish, and Swedish), we only focus on English.

[3]For the English task, the models considered are: `TheBloke/Mistral-7B-Instruct-v0.2-GGUF`, `tiiuae/falcon-7b-instruct` (Almazrouei et al., 2023), and `togethercomputer/Pythia-Chat-Base-7B`.

2 Related Works

Fact-checking is a common approach to mitigate hallucinations in LLM outputs (Nakov et al., 2021; Guo et al., 2022). However, it typically depends on external knowledge sources such as databases, search engines, or pre-verified information repositories. These sources, while useful, are often incomplete, domain-specific, and require continuous updates to remain relevant (Cheng et al., 2024). Additionally, integrating them into real-time LLM applications introduces significant computational overhead, making the process inefficient and sometimes impractical at scale.

To overcome these limitations, this work proposes an alternative approach based on *uncertainty quantification* (UQ), which detects hallucinations directly from the model's own outputs without relying on external verification (Kotelevskii et al., 2022; Vazhentsev et al., 2022). Our methodology provides a way to assess how confident an LLM is in its generated text, offering a built-in mechanism for identifying potentially unreliable information.

Detecting hallucinations at the claim level is a challenging task (Fadeeva et al., 2024), as a single output may contain both accurate and inaccurate information, requiring finer-grained uncertainty measurement to highlight specific false or misleading claims. Our work addresses this challenge by expanding upon previous research in this direction (Fadeeva et al., 2024), which introduced a new token-level uncertainty score by aggregating token uncertainties into claim-level scores. We adapt their method for word-level detection and introduce a larger model to verify the smaller model's output, leveraging its broader knowledge. Additionally, we enhance the hallucination score by factoring in the uncertainty of both the NLI model and the LLM.

3 Methodology

To detect hallucinated content in a sentence generated by a model m, we compare it to a larger model M, which we assume has better general knowledge. The goal is to check if each word in m's output is consistent with what M would generate. Our method analyzes words individually, using

both probability scores and an NLI model[4].

Word probability from model M. For each word in m's output, we provide its preceding context to M. We then extract the most likely tokens from M's probability distribution until (i) their combined probability reaches a threshold k, or (ii) a single token has a probability lower than ρ[5]. Therefore, the number of selected tokens N varies depending on the word. Given the selected tokens, we determine the probability of the full word w by summing the probabilities of its tokens: $p(w) = \sum_{t_j \in w} p_j$, where t_j are the tokens forming w and p_j are their probabilities.

Checking semantic consistency with NLI. We assess whether each word aligns with the original sentence's meaning using an NLI model. Such model determines whether replacing one word with another changes the meaning of the sentence. If the sentence is too long, we truncate it around the word to fit the model's context window.

The NLI model assigns probabilities for three possible relationships: *Entailment* ($\mathrm{P}_+(w)$), i.e., the word fits naturally in the sentence; *Neutral* ($\mathrm{P}_=(w)$), i.e., the word has a different meaning but does not contradict the original sentence; *Contradiction* ($\mathrm{P}_-(w)$), i.e., the word changes the meaning of the sentence.

Computing the hallucination score. To measure hallucination at the word level, we start from the original approach proposed by Fadeeva et al. (2024). They compute the hallucination score (HS) at the claim level as follows:

$$HS = 1 - \frac{\sum p(c_i|e^+)}{\sum p(c_i|e^+) + \sum p(c_i|e^-)} \quad (1)$$

where $p(c_i|e^+)$ represents the probability of a claim having positive entailment, and $p(c_i|e^-)$ corresponds to the probability of a claim having negative entailment.

We extend their method by forcing it to operate at the word level and integrating NLI uncertainty:

$$HS = 1 - \frac{\sum_{i=1}^{N} (p_i(w) \cdot p_i^{nli}(w))}{\sum_{i=1}^{N} p_i(w)} \quad (2)$$

where $p_i(w)$ represents the probability assigned by the model to word w in position i, and $p^{nli}(w)$ is the sum of the probabilities for entailment and neutrality, defined as follows:

$$p^{nli}(w) = P_+(w) + P_=(w) \quad (3)$$

We incorporate neutral entailment alongside positive entailment, unlike (Fadeeva et al., 2024), as it empirically improved results on the validation set.

This formulation in Eq. 2 effectively captures the inverse relationship between word confidence and hallucination likelihood while accounting for semantic coherence through the NLI component.

The baseline methodology operates independently of human-annotated ratings, ensuring applicability in scenarios where labeled data is scarce or unavailable. However, to enhance alignment with human perception of hallucinations, we introduce a calibration mechanism through a multiplicative factor η:

$$HS^* = \eta \cdot HS \quad (4)$$

The parameter η serves as an alignment coefficient that is empirically determined using the validation dataset to maximize the correlation between our computed scores and the soft labels provided by human reviewers. This calibration process allows to fine-tune the sensitivity of the hallucination detection system to better match human judgment thresholds.

In our experimental framework, we systematically evaluate two distinct configurations: (1) a label-agnostic variant (`LAV`) where $\eta = 1$, which preserves the model's inherent hallucination detection capabilities without reliance on human feedback, and (2) a reviewer-aligned variant (`RAV`) where η is optimally selected based on validation data to maximize correlation with human annotations. The former configuration is particularly valuable in zero-shot deployment scenarios or when consistent detection criteria are required across diverse domains, while the latter configuration offers enhanced performance in applications where human perception of hallucinations is the primary evaluation metric.

4 Experimental Setup

This section outlines the various methods implemented, which will be evaluated in Section 5 using the metrics detailed below.

[4]We used `Qwen/QwQ-32B-Preview` (Yang et al., 2024) as M, and `cross-encoder/nli-deberta-v3-large` (He et al., 2020) for the NLI task.

[5]We set k to 0.9 and ρ to 0.005 in our experiments.

4.1 Methods

To evaluate the effectiveness of our proposed approach, we compare it against the baseline methods proposed by Vázquez et al. (2025).

Mark-None (-All). Trivial baselines where no (all) the words are considered as hallucinations. This serves as a lower bound for detection performance.

RoBERTA. We fine-tune a RoBERTA model (Liu et al., 2019) on the labeled validation set for all the 14 available languages to classify words as hallucinations or not. The model was trained for five epochs with a learning rate of 2e-5 and weight decay of 0.01.

Fadeeva et al. (2024) We adapt the scoring mechanism described in Eq.1 within our pipeline to evaluate hallucinations at the word level. This allows us to compare their formulation with ours, remaining consistent with the purpose of the challenge.

Ours. We evaluate our proposed hallucination detection method with two variants: (1) LAV, which applies our detection framework without any alignment to human annotations and uses only the test set, and (2) RAV, where we introduce the multiplicative factor η to adjust the hallucination scores based on the grading patterns of the human reviewers. For the latter, we use the value that maximizes the correlation with the soft labels, which is $\eta = 1.48$, as shown in Figure 1.

4.2 Evaluation Metrics

We employ two character-level evaluation metrics proposed by (Vázquez et al., 2025) to measure the performance of the different methods.

Intersection over Union (IoU) with Hard Labels to evaluate the overlap between the characters predicted as hallucinations and the hard labels.

Pearson Correlation with Soft Labels to measure how well the predicted probability of a character being part of a hallucination correlates with the empirical probabilities derived from annotator judgments.

5 Results

Table 1 compares the performance of different methods based on the two evaluation metrics described above.

As expected, the Mark-None method performs the worst. Its scores are close to zero for both metrics, showing that it fails to capture hallucinated content. On the other hand, the Mark-All method achieves a relatively high IoU. This is likely because large portions of the text in this task are hallucinated. However, its correlation score is very low ($\rho = 0.000$), meaning it does not align well with human judgments.

Method	η	ρ	IoU
Mark-None	-	.000	.032
Mark-All	-	.000	**.349**
RoBERTa	-	.119	.031
Fadeeva et al. (2024)	-	<u>.309</u>	.283
Ours-LAV	1	.300	.310
Ours-RAV	1.48	**.324**	<u>.311</u>

Table 1: Comparison of the different methods. Best results in **bold**, second-best <u>underlined</u>.

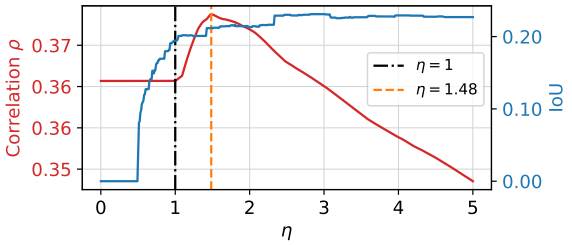

Figure 1: Analysis of ρ and IoU as η changes on the validation set. In this case, ρ is maximized for $\eta = 1.48$.

The RoBERTa-based model performs slightly better. It improves correlation ($\rho = 0.119$), meaning it captures some alignment with annotator probabilities. However, it has a low IoU, indicating that it struggles with precise localization.

Our proposed method significantly outperforms these baselines. We test it in two configurations described in Section 4. Both variants of our method achieve the highest overall performance as they provide the best balance between the two considered metrics. The label-agnostic model (LAV) reaches a correlation of $\rho = 0.300$ and an IoU of 0.310. This surpasses in IoU the scoring method proposed in (Fadeeva et al., 2024) and adapted to our scenario while maintaining a similar correlation. The reviewer-aligned version (RAV) further improves correlation ($\rho = 0.324$) while keeping a strong IoU (0.311). This result shows that our approach effectively identifies hallucinated content. When properly calibrated, it also aligns well with human judgments, thus providing a strong balance between accurate hallucination detection and agreement with human perception.

Impact of η. Figure 1 shows how IoU and correla-

tion change as η varies from 0 to 5. When $\eta = 1$, the method is label-agnostic. The highest Pearson correlation (0.324) occurs at $\eta = 1.48$ (yellow dotted line), indicating the best alignment with human soft labels. IoU remains stable across different η values, as expected. Even after applying the correction factor, hallucination failure detection (HS = 0) stays unchanged. These results confirm that tuning our hallucination scores using the validation set improves performance. The reviewer-aligned method ($\eta = 1.48$) better matches human perception of hallucinations while still performing well on the hard label detection task.

5.1 Limitation of the Proposed Method

Our method is effective at detecting hallucinations, but it has a key limitation. The model generates text step by step, using all previous tokens as context. If a hallucination appears early, it becomes part of this context. The model then builds on the false information, creating more text that fits the hallucination. This makes it difficult to spot later hallucinations that seem consistent with the first one. As a result, the model may correctly detect the initial false statement but fail to identify the ones that follow. This creates a propagation effect, where one mistake leads to more undetected errors.

For example, consider the following case from the test set:

Sentence: *What is the dry boiling point of DOT 5 brake fluid?*
`The dry boil point for DOT5 Brake Fluid is 212°F (100°C).`
Ground Truth: 212°F (100°C)
Our detection: 212°F

Here, our method correctly identifies the first hallucination (the temperature *"212°F"*) but fails to mark the Celsius conversion *"(100°C)"* as part of the hallucination. This occurs because once the model has incorporated the incorrect Fahrenheit value into its context, the corresponding Celsius conversion becomes consistent with this value despite both being wrong.

The example that follows further illustrates this limitation:

Sentence: *Which mountain range is Speichersdorf located near?*
`Speicersdorf is located in the Black Forest mountain region of Germany.`
Ground Truth: Black Forest
Our detection: Black

In this case, our method identifies only the first word of the hallucinated mountain region (*"Black"*) but misses *"Forest"*. Once the context includes the word *"Black"*, the word *"Forest"* becomes a natural and expected continuation, even though both words are factually incorrect.

This limitation shows the difficulty of detecting linked hallucinations when relying only on the model's confidence in a single word, especially if the context is already hallucinated.

6 Conclusion

This work addresses the Mu-SHROOM shared task at SemEval 2025, focusing on detecting word-level hallucinations in LLM outputs. We introduce a novel approach that uses a larger model validation without the need for external knowledge sources.

Our method achieves strong results compared to the proposed baselines, proving to be an excellent starting point for evaluating hallucinations when a ground truth or external sources are unavailable. One key limitation is handling multiple connected hallucinations. The model generates each word based on past text. If a hallucination appears, it becomes part of the context. The model may then continue building on this false information, making the hallucination harder to detect. This can lead to a chain of believable but incorrect statements. To address this, future work could use a broader context or develop mechanisms to review and correct past text when a hallucination is found.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Jean-Christophe Bélisle-Pipon. 2024. Why we need to be careful with llms in medicine. *Frontiers in Medicine*, 11:1495582.

Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, and Francesco Tarasconi. 2024. MAINDZ at SemEval-2024 task 5: CLUEDO - choosing legal oUtcome by explaining decision through oversight. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 997–1005, Mexico City, Mexico. Association for Computational Linguistics.

Lorenzo Bertetto, Francesca Bettinelli, Alessio Buda, Marco Da Mommio, Simone Di Bari, Claudio Savelli, Elena Baralis, Anna Bernasconi, Luca Cagliero, Stefano Ceri, et al. 2024. Towards an explorable conceptual map of large language models. In *International Conference on Advanced Information Systems Engineering*, pages 82–90. Springer.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. MALTO at SemEval-2024 task 6: Leveraging synthetic data for LLM hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684, Mexico City, Mexico. Association for Computational Linguistics.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, et al. 2024. Integrative decoding: Improve factuality via implicit self-consistency. *arXiv preprint arXiv:2410.01556*.

Erica Coppolillo, Marco Minici, Federico Cinus, Francesco Bonchi, and Giuseppe Manco. 2024. Engagement-driven content generation with large language models. *arXiv preprint arXiv:2411.13187*.

Sunhao Dai, Weihao Liu, Yuqi Zhou, Liang Pang, Rongju Ruan, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2024. Cocktail: A comprehensive information retrieval benchmark with LLM-generated documents integration. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7052–7074, Bangkok, Thailand. Association for Computational Linguistics.

Lorenzo D'Amico, Davide Napolitano, Lorenzo Vaiani, Luca Cagliero, et al. 2023. Polito at multi-fake-detective: Improving fnd-clip for multimodal italian fake news detection. In *CEUR WORKSHOP PROCEEDINGS*, volume 3473. CEUR-WS.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International journal of information management*, 71:102642.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Flavio Giobergia, Alkis Koudounas, and Elena Baralis. 2024. Large language models-aided literature reviews: A study on few-shot relevance classification. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323.

Moreno La Quatra, Luca Cagliero, and Elena Baralis. 2021. Leveraging full-text article exploration for citation analysis. *Scientometrics*, 126(10):8275–8293.

Moreno La Quatra, Nicole Dalia Cilia, Vincenzo Conti, Salvatore Sorce, Giovanni Garraffa, and Valerio Mario Salerno. 2025. Vision-language multimodal fusion in dermatological disease classification. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2024 International Workshops and Challenges*. Springer Nature Switzerland.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 639–649. Springer.

Claudio Savelli and Flavio Giobergia. 2024. Enhancing cross-lingual word embeddings: Aligned subword vectors for out-of-vocabulary terms in fasttext. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6. IEEE.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024. Do llms plan like human writers? comparing journalist coverage of press releases with llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in faster curation of text datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.