Multilingual Fine-Grained News Headline Hallucination Detection

Jiaming Shen[†] **Tianqi Liu**[†] **Jialu Liu**[†] **Zhen Qin**[†]
Jay Pavagadhi[‡] **Simon Baumgartner**[†] **Michael Bendersky**[†]
[†] Google Research [‡] Google
{jmshen, tianqiliu, jialu, zhenqin, jaynp, simonba, bemike}@google.com

Abstract

The popularity of automated news headline generation has surged with advancements in pre-trained language models. However, these models often suffer from the "hallucination" problem, where the generated headline is not fully supported by its source article. Efforts to address this issue have predominantly focused on English, using over-simplistic classification schemes that overlook nuanced hallucination types. In this study, we introduce the first multilingual, fine-grained news headline hallucination detection dataset that contains over 11 thousand ⟨article, headline⟩ pairs in 5 languages, each annotated with detailed hallucination types by experts. We conduct extensive experiments on this dataset under two settings. First, we implement several supervised fine-tuning approaches as preparatory solutions and demonstrate this dataset's challenges and utilities. Second, we test various large language models' in-context learning abilities and propose two novel techniques, language-dependent demonstration selection and coarse-to-fine prompting, to boost the few-shot hallucination detection performance in terms of the example-F1 metric. We release this dataset to foster further research in multilingual, fine-grained headline hallucination detection.

1 Introduction

A news headline provides a concise summary of its corresponding news article, enabling readers to quickly grasp the essence of a news story. Numerous generative models (Gu et al., 2020; Cai et al., 2023; Ding et al., 2023) have been developed to automate the process of condensing a news article into its headline, achieving generally commendable quality. However, people note that these models often encounter the hallucination problem, where the produced headline does not fully align with the source article's content. For example, as shown in Figure 1, the generation model is given an article about "UA adds new routes in Midwest" and

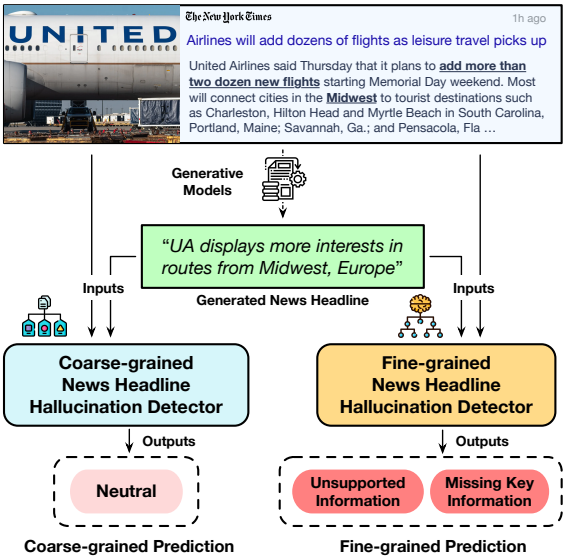

Figure 1: A comparative example of headline hallucination detection at different levels of granularity. The fine-grained hallucination detector goes beyond traditional 3-class label **Neutral** and offers more nuanced predictions like **Unsupported Information** (because the article does not references "Europe") and **Missing Key Information** (as the headline omits the crucial detail that "new routes in the Midwest are being added").

outputs the headline "*UA displays more interests in routes from Midwest, Europe*". This headline is considered as a hallucination because it references "Europe" which is absent from the article, and omits the crucial detail that "new routes in the Midwest are being added".

To mitigate these hallucinations, various studies propose to pre-process the training corpus of generation models by removing or re-weighting possibly hallucinated examples (Nan et al., 2021a; Aharoni et al., 2022; Qiu et al., 2023b). Another line of work proposes to first detect hallucinations in generated outputs and then filter them in a post-processing stage (Honovich et al., 2022; Shen et al., 2023). These approaches typically adopt a binary or three-way classification scheme and focus on

Findings of the Association for Computational Linguistics: EACL 2024, pages 7862–7875

examples written exclusively in English or translated from a single English source. Despite some promising results, it remains unclear how these approaches can capture more fine-grained hallucination error types in multilingual news articles and inform more nuanced decision making process.

In this work, we propose a new task — *fine-grained headline hallucination detection* and study it in the multilingual setting. This objective is to identify a *set* of fine-grained entailment relations between a given news article and its headline. Taking the example in Figure 1 for instance, we aim to advance beyond a simple "Neutral" classification and provide more fine-grained predictions: (1) "Unsupported Information" due to the article's lack of references to "routes from Europe", and (2) "Missing Key Information" because the headline fails to capture a core message of the article — the introduction of new routes in the Midwest.

To advance research in this area, we introduce the first Multilingual Fine-grained Headline Hallucination Detection (MFHHD) dataset, featuring 11,469 examples across 5 languages. Each example comprises a news article, a generated news headline, a coarse-grained hallucination label, and a set of fine-grained hallucination labels annotated by 2 to 4 dedicated domain experts fluent in the language of the original article. Additionally, for examples labeled as "Neutral" or "Contradict", annotators will provide a natural language justification for their fine-grained annotations.

The introduction of this new MFHHD dataset presents intriguing research challenges, such as identifying complex, nuanced types of hallucination errors and exploring whether existing hallucination detection moethods, previously English-centric, can be adapted for multilingual use. To answer these questions, we carry out extensive experiments in both supervised fine-tuning and few-shot learning scenarios. In the supervised fine-tuning context, we find that model pre-training on natural language inference datasets and incorporation of natural language explanations into seq2seq based classifier can both enhance the detection performance. In the few-shot learning domain, we evaluate various large language models (e.g., ChatGPT (OpenAI, 2022), PaLM2 variants (Anil et al., 2023)) and observe that they perform worse than the smaller fine-tuned models (e.g., mT5-XXL (Xue et al., 2020)). To improve these LLMs' in-context learning capabilities, we introduce two prompting techniques: (1) *language-dependent demonstration selection* which dynamically chooses few-shot examples in the same language as the test query example, and (2) *coarse-to-fine prompting* that guides LLMs to generate a coarse-grained prediction before making fine-grained hallucination type predictions. Both techniques can significantly enhance LLM's few-shot learning effectiveness and boost the detection performance in terms of the example-F1 metric.

Contributions. The major contributions of this paper are summarized as follows: (1) We introduce a novel task, fine-grained headline hallucination detection, aimed at identifying more nuanced hallucination error types; (2) We create a new multilingual fine-grained hallucination detection dataset MFHHD, curated by news domain experts; and (3) We conduct extensive experiments on the MFHHD dataset, delving into its complexities and offering valuable insights for improving the accuracy of fine-grained hallucination detection in both supervised and few-shot learning settings.

2 Problem Formulation

In this study, we represent both a news *article* d and a news *headline* h as a sequence of tokens. The ideal purpose of a news headline is to provide a concise summary of the news article. However, when automatically generated, the headline can hallucinate and misrepresent the original intent of the source article. The **coarse-grained headline hallucination detection** task inputs a pair of news article and headline $\langle d_i, h_i \rangle$ and outputs its coarse-grained entailment relation l_i indicating whether the headline is fully supported, directly contradicts, or remains neutral with respect to the article.

In many real-world applications, we notice this the three-way entailment classification schema is too coarse-grained and fails to pinpoint the exact hallucination reasons (e.g., the headline includes an incorrect number or reports a person's subjective opinion as a fact). Therefore, we propose the **fine-grained headline hallucination detection** task that returns *a set of* fine-grained entailment relations $\mathcal{R}_i = \{R_i^1, R_i^2, \dots\}$ for a pair of input article and headline $\langle d_i, h_i \rangle$. Each relation R_i^j specifies a detailed reason why the given headline h_i either supports, contradicts, or is neutral in relation to the corresponding article d_i.

3 MFHHD Dataset

In this work, we collect the first Multilingual Fine-grained Headline Hallucination Detection (MFHHD) dataset that contains 11,469 examples across 5 languages (English, Spanish, German, French, Portuguese). Each example includes a news article, a news headline, a coarse-grained hallucination label, a set of fine-grained hallucination labels, and an optional set of natural language annotation justifications. The dataset is currently available at: https://bit.ly/MFHHD-dataset.

3.1 Dataset Construction

We follow the same procedure as in (Shen et al., 2023) to sample a set of news articles along with their headlines generated from NHNet (Gu et al., 2020)[1]. We examine these sampled ⟨article, headline⟩ pairs and discuss with multiple news domain experts to outline the following 7 fine-grained hallucination types (see Figure 5 in Appendix A.1 for examples of each hallucination type).

(1) **Neutral (extra info)**: the headline contains unsupported additional information that cannot be verified by the given article.

(2) **Neutral (missing info)**: the headline misses important information (e.g., key dates, locators) and thus changes the scope/emphasis of the article. Those missing information will be significant to the extent that it alters reader's perception of the article's core messages. Table A.1 in the appendix lists one example.

(3) **Neutral (off topic)**: the headline and article discuss two completely different topics.

(4) **Neutral (others)**: the catch-all option for all other forms of headlines that neither fully supports nor directly contradicts the article.

(5) **Contradictory (opinion as fact)**: the article states an opinion or unconfirmed rumor while the headline interprets it as a factual statement.

(6) **Contradictory (wrong number)**: the headline includes an incorrect important number that directly contradicts the news article.

(7) **Contradictory (others)**: the catch-all option for all other forms of direct contradictions between the headline and the news article.

We prepare a detailed curation guideline that lists the definitions and representative examples for the above 7 fine-grained hallucination types plus 1 non-hallucination type (i.e., "Support") and train all the annotators for two rounds. All annotators are full-time journalist degree holders and speak the same language of the annotated article. Due to some policy constraints, we cannot report their detailed compensation here but we guarantee their pay is definitely above the corresponding local minimum wage. Given a pair of article and headline, they are instructed to first choose one coarse-grained type ("Support", "Neutral", or "Contradict") and then to select all fine-grained hallucination types of this example. Additionally, if they label one example as "Neutral" or "Contradict", we encourage them to provide an additional natural language explanation to justify their fine-grained annotations. Each example undergoes initial evaluation by two annotators and if they disagree with each other at the coarse-grained label, we engage two additional curators to thoroughly review the example. Finally, we retain all examples that receive a majority consensus at the coarse-grained label level and preserve all corresponding fine-grained labels associated with these chosen coarse-grained labels. The initial round of annotator agreement (at the coarse-grained level) is 74.3%. For the remaining 25.7% of examples that have two rounds of annotations from 4 raters, their inter-rater agreement is 0.6642 Cohen's Kappa and thus can be considered as substantial agreement.

3.2 Dataset Analysis

We analyze some properties of our MFHHD dataset and show the results in Figure 2.

First, we can see that over 65% of examples in our dataset are non-English examples and their corresponding languages are evenly distributed in German, Spanish, French, and Portuguese.

Second, we analyze the coarse-grained label distribution and observe about one-third of examples are marked as "Neutral" or "Contradict". Furthermore, each "Neutral" example has an average 1.3 fine-grained labels and about 29% of examples have more than one fine-grained label. Conversely, only 4.1% of "Contradict" examples have more than one fine-grained label. One possible explanation is that raters tend to give a single most severe contradictory reason instead of selecting multiple fine-grained ones. In a holistic view, we can see about 21% of all hallucinated examples have more than one fine-grained label, which necessitates our formulation of fine-grained hallucination detection task as a multi-label classification problem.

Finally, we analyze the textual information in our

[1] More results are discussed in Appendix A.2.

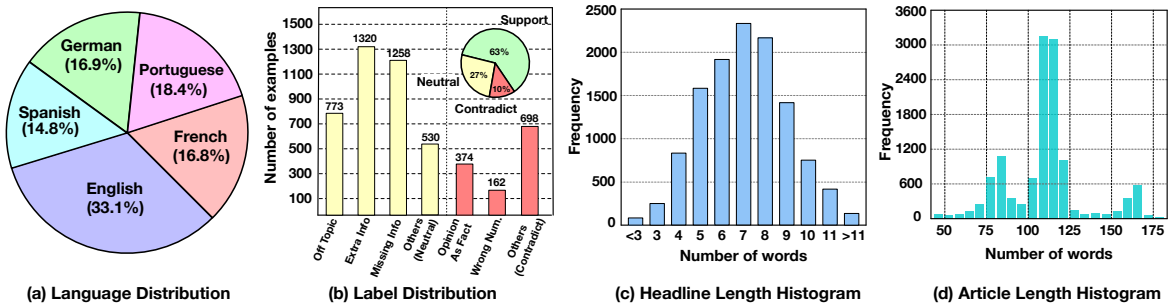

Figure 2: Analysis of our multilingual fine-grained headline hallucination detection (MFHHD) dataset.

MFHHD dataset. We draw the length histograms of news articles and headlines in Figure 2(c)(d) where we can see the median headline and article length are 7 and 106 words, respectively. Furthermore, over 96% of "Neutral" or "Contradict" examples have at least one natural language explanation and about half of them have two or more explanations. These human written explanations have 18 words in median and provide useful signals for detecting headline hallucinations.

Challenging Nature of MFHHD dataset. We want to emphasize that the MFHHD is a challenging dataset. First, as shown in the experiments below, a fully fine-tuned model with billions of parameters can only achieve around 0.74 accuracy. This number is significantly lower than most of other existing NLI datasets. Second, many fine-grained hallucination classes are very subtle. Even though those general news articles do not contain any niche topics, many domain experts need more than 5 minutes to accurately label all fine-grained hallucination classes.

4 Supervised Fine-grained Headline Hallucination Detection

In this section, we experiment a set of supervised methods for multilingual fine-grained headline hallucination detection. The goal is to better understand the characteristics and challenges of our MFHHD dataset and to share some valuable insights that can help later LLM-based few-shot method designs (c.f. Section 5).

4.1 Experiment Settings

Dataset. Given the curated MFHHD dataset, we first create the test set by randomly selecting 1000 English examples and 500 examples for each of the remaining languages (German, French, Spanish, Portuguese). Then, we use the remaining 8,469 examples for training models and test their performances on the above selected 3,000 test examples.

Compared Methods. We compare the following representative methods for the multilingual headline hallucination detection task:

- mDeBERTa$_{base}$: The multilingual version of DeBERTa (He et al., 2023) which enhances the original BERT model (Devlin et al., 2019) using replaced token detection as the pretraining task and pre-trained on CC100 multilingual data. We concatenate the headline and the article text (with a special separator token [SEP]) and feed it into the mDeBERTa$_{base}$ model for prediction.
- mDeBERTa$_{base}$ + NLI: We first adopt the above mDeBERTa$_{base}$ model and further pre-train it on various natural language inference (NLI) datasets including XNLI (Conneau et al., 2018) and the translated version of MNLI (Williams et al., 2018), ANLI (Nie et al., 2020) and WANLI (Liu et al., 2022). Then, we fine-tune the model on our MFHHD dataset.
- mT5$_{xxl}$: The multilingual version of T5 (Xue et al., 2020), an encoder-decoder model with strong representation power. We input the concatenated headline and article into the encoder and requires the decoder to output a single token indicating the final predicted coarse-grained class (or a sequence of tokens, each represents one fine-grained hallucination class).
- mT5$_{xxl}$ + Exp: We incorporate human written natural language explanations into the mT5$_{xxl}$ model by requiring its decoder to output the class token(s) followed by the explanation. See Figure 3 for a reference.
- mT5$_{xxl}$ + NLI: Similar to mDeBERTa$_{base}$ + NLI, we pre-train the mT5$_{xxl}$ model on NLI datasets and fine-tune it on MFHHD dataset.

- mT5$_{xxl}$ + NLI + Exp: The combination of mT5$_{xxl}$ + Exp and mT5$_{xxl}$ + NLI where we incorporate explanations information during the MFHHD fine-tuning stage.

For the last four encoder-decoder based models, we evaluate their abilities to detect both coarse-grained and fine-grained hallucinations. We map fine-grained predictions into their corresponding coarse-grained hallucination labels. For example, the fine-grained model output "Off-Topic Missing-Info" is mapped to the "Neutral" class.[2]

We use Huggingface library (Wolf et al., 2020) to implement mDeBERTa$_{base}$ and mDeBERTa$_{base}$ + NLI. For the remaining four mT5-based models, we develop them based on the T5X library[3](Roberts et al., 2022). Appendix A.3 provides more implementation details and hyper-parameter settings.

Evaluation Metrics. We evaluate coarse-grained hallucination detection performance using standard multi-class classification metrics including "Accuracy" and "Weighted-F1". The Weighted-F1 considers the number of true instance for each class and thus account for class imbalance. For fine-grained hallucination detection, we formulate it as a multi-label classification problem and follow previous studies (Prabhu et al., 2018; Shen et al., 2021) to use the "Example-F1" metric for evaluation. The Example-F1 is calculated as follows:

$$\text{Example-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|\mathcal{R}_i^{gt} \cap \mathcal{R}_i^{pred}|}{|\mathcal{R}_i^{gt}| + |\mathcal{R}_i^{pred}|},$$

where N is the number of test examples; $\mathcal{R}_i^{gt}$ and $\mathcal{R}_i^{pred}$ stand for the ground truth and model predicted fine-grained hallucination class set of test example $\langle d_i, h_i \rangle$, respectively.

4.2 Experiment Results

Table 1 presents our experiment results. First, we observe that pre-training on NLI datasets (before the in-domain fine-tuning) can consistently enhance the model performance. This improvement could stem from the shared characteristics between the headline hallucination detection task and the natural language inference task, both aiming to assess text grounding capability. Second, we find that

[2]Although the model can in theory output multiple incompatible fine-grained class tokens (e.g. "Off-Topic Incorrect-Number" where one token corresponds to the "Neutral" class while the other belongs to the "Contradict" class), we do not witness such a case in practice for the supervised setting.

[3]https://github.com/google-research/t5x

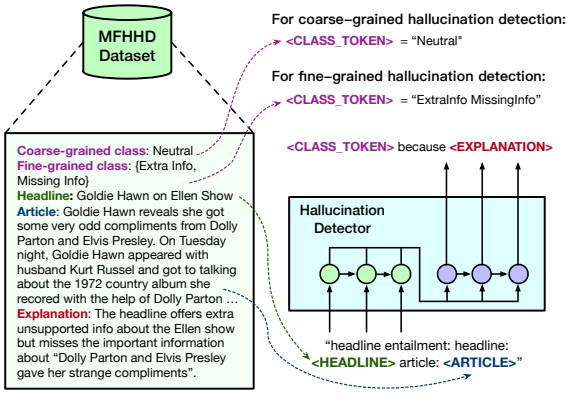

Figure 3: Detecting news headline hallucinations with models of the encoder-decoder architecture.

Methods	Accuracy	Weighted-F1	Example-F1
Coarse-grained Detection			
mDeBERTa$_{base}$	63.70	49.57	—
mDeBERTa$_{base}$ + NLI	66.80	62.30	—
mT5$_{xxl}$	71.20	69.68	—
mT5$_{xxl}$ + Exp	73.23	71.82	—
mT5$_{xxl}$ + NLI	72.60	71.52	—
mT5$_{xxl}$ + NLI + Exp	**73.97**	**73.11**	—
Fine-grained Detection			
mT5$_{xxl}$	71.80	70.71	63.89
mT5$_{xxl}$ + Exp	72.63	71.51	66.24
mT5$_{xxl}$ + NLI	73.53	72.59	66.78
mT5$_{xxl}$ + NLI + Exp	**74.27**	**73.34**	**67.52**

Table 1: The experiment results on supervised headline hallucination detection. The "Coarse-grained Detection" methods directly predict a coarse-grained label ("Support", "Neutral", "Contradict") and are evaluated by the metric "Accuracy" and "Weighted-F1". The "Fine-grained Detection" methods predict a set of fine-grained labels (evaluated by "Example-F1") and we map them back to a single coarse-grained label. Please refer to Section 4.1 for more details.

Dataset	Q2	ANLI	mT5$_{xxl}$ + NLI + Exp
MNBM	66.5	66.7	**70.44**
FRANK	**82.9**	83.5	75.85
QAGS	**78.3**	75.3	72.76
SummEval	77.3	72.9	**85.81**
FEVER	82.7	**90.2**	88.24
Vitamin-C	75.7	74.7	**84.71**
Average	77.23	77.22	**79.64**

Table 2: The experiment results on TRUE benchmark.

incorporating natural language explanations into models with the encoder-decoder architecture can significantly boost the model performance. Similar observations are found in prior studies (Narang et al., 2020; Shen et al., 2023) and here our experiments verify the same phenomenon holds for multi-label fine-grained hallucination detection.

Besides incorporating NLI based pretraining and utilizing natural language explanation, we notice that for coarse-grained detection, it is generally

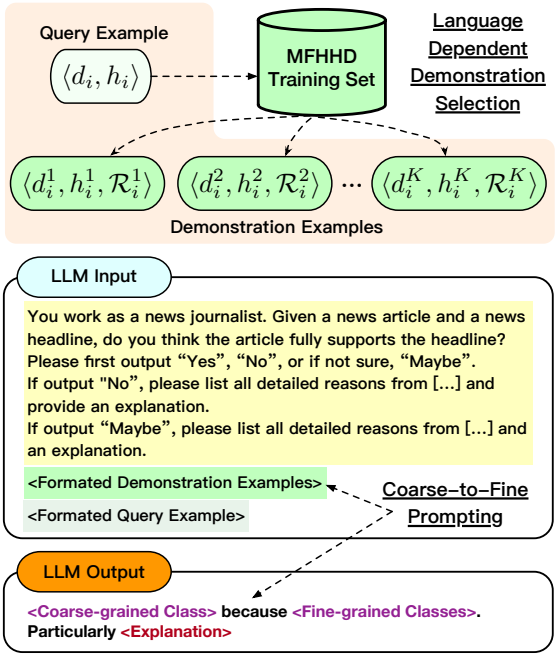

Figure 4: Detecting fine-grained headline hallucinations using LLM with language dependent demonstration selection and coarse-to-fine prompting.

preferable to initially train models for fine-grained prediction and then map fine-grained classes to coarse-grained ones. Also, we want to stress that even the best performing method, mT5$_{xxl}$ + NLI + Exp, with 13B parameters and trained with in-domain data, still only achieves about 74% detection accuracy. This indicates the challenging nature of our MFHHD dataset, leaving plenty of room for future research improvements.

Finally, we show that the model trained on our benchmark can generalize to more hallucination detection datasets. Specifically, we evaluate the model variant "fine-grained mT5$_{xxl}$ + NLI + Exp" on the TRUE benchmark (Honovich et al., 2022). Table 2 reports the experiment results. The Q2 and ANLI are two best performing methods in the original TRUE paper. We can see that the model trained on our MFHHD dataset has good zero-shot performance on the TRUE benchmark, which demonstrates the broad applicability of our dataset and the generalization ability of our method.

5 Few-shot Fine-grained Headline Hallucination Detection

We introduce various few-shot methods that leverages large language models for detecting fine-grained headline hallucinations in this section.

5.1 LLM In-Context Learning (ICL)

Recent studies have demonstrated that Large language models (LLMs) can quickly adapt to various tasks by learning only on a few demonstration examples in context (Brown et al., 2020; Wei et al., 2022; OpenAI, 2022). Specifically, given a test example $\langle d_i, h_i \rangle$ and a set of K demonstrations $\{\langle d_i^j, h_i^j, \mathcal{R}_i^j \rangle|_{j=1}^K\}$ for the hallucination detection task, we will first format them using a template and then prompt the LLM to decode an output sequence that corresponds to the final prediction (c.f. Appendix A.5 for prompt template examples).

5.2 Language Dependent Demonstration Selection and Coarse-to-Fine Prompting

In this work, we explore various LLMs to address the following research question: "How to prompt LLMs for best few-shot fine-grained hallucination detection performance in a multilingual context?". We introduce two simple yet effective techniques to achieve this objective, illustrated in Figure 4.

First, instead of using a fixed set of demonstrations for all test examples, we propose to select a dynamic set of demonstrations based on the language of the test example. Different from most previous retrieval-based ICL studies (Luo et al., 2024), this method does not rely on an external retrieval model and thus has a better application scope. Second, we note that those fine-grained hallucination classes are interrelated rather than isolated. For example, an instance cannot have both a "Neutral" and a "Contradict" subclass at the same time. Given this hierarchical organization of hallucination labels, we present a coarse-to-fine prompting approach. In this method, we prompt the LLM to produce an initial coarse-grained hallucination prediction, followed by more specific fine-grained predictions along with a natural language explanation. We conduct extensive experiments to evaluate these two techniques with various LLMs.

5.3 Experiment Settings

For main experiments, we test PaLM2-S, PaLM2-M and PaLM2-L (Anil et al., 2023) under 1-shot, 3-shot, and 5-shot settings. All demonstrations are selected from the MFHHD training set. We employ the same set of metrics as described in Section 4.1 to compare the following methods: (1) **LI-FG** uses a fixed set of demonstrations and directly prompts the LLM to output a set of fine-grained hallucination classes; (2) **LD-FG** selects language depen-

Backbone	Methods	1-shot			3-shot			5-shot		
		Accuracy	Weighted-F1	Example-F1	Accuracy	Weighted-F1	Example-F1	Accuracy	Weighted-F1	Example-F1
PaLM2-L	LI-FG	63.24	55.00	60.32	64.41	57.05	60.97	64.77	57.90	61.58
	LD-FG	65.38	55.49	61.41	64.63	57.78	61.53	64.73	57.80	61.98
	LI-C2FG	65.07	60.59	61.29	65.55	61.21	61.81	66.69	62.25	62.05
	LD-C2FG	**65.67**	**60.82**	**62.15**	**66.96**	**63.01**	**61.99**	**67.03**	**62.90**	**62.14**
PaLM2-M	LI-FG	48.34	38.85	42.87	47.76	45.81	56.96	64.99	58.75	60.01
	LD-FG	49.81	46.06	43.27	54.19	52.84	59.45	64.87	58.35	60.07
	LI-C2FG	49.32	44.61	45.65	60.67	56.64	58.40	66.65	63.88	60.44
	LD-C2FG	**56.63**	**46.73**	**47.71**	**64.88**	**62.01**	**59.65**	**67.21**	**64.18**	**61.52**
PaLM2-S	LI-FG	17.63	24.58	32.08	54.52	45.54	51.05	64.76	61.57	56.40
	LD-FG	35.73	37.52	36.86	54.33	47.86	54.59	64.96	61.39	56.25
	LI-C2FG	35.60	36.82	39.45	56.07	44.37	55.16	66.32	63.02	61.51
	LD-C2FG	**45.00**	**47.26**	**49.54**	**57.31**	**50.54**	**58.68**	**66.49**	**63.21**	**61.63**

Table 3: The experiment results on few-shot fine-grained headline hallucination detection. Prefixes "LI" and "LD" in method names stand for Language-Independent and Language-Dependent variants, respectively. We run each methods five times and report the averaged metrics.

Accuracy	EN	ES	DE	FR	PT	Avg.
EN	**80.70**	51.40	**53.40**	60.40	62.60	64.87
ES	78.40	**54.40**	50.40	60.40	63.20	64.20
DE	80.20	52.80	51.40	60.40	64.60	**65.37**
FR	80.10	52.80	51.40	**60.80**	**65.20**	65.07
PT	79.80	53.20	54.40	60.40	64.60	**65.37**
Weighted-F1	EN	ES	DE	FR	PT	Avg.
EN	**78.22**	44.36	**45.05**	52.21	56.31	**58.56**
ES	77.24	**44.79**	38.22	52.36	55.69	57.06
DE	78.78	42.57	43.26	53.46	55.39	58.14
FR	77.86	44.42	42.50	**54.24**	56.22	58.29
PT	77.99	42.94	40.07	53.22	**57.33**	57.72
Example-F1	EN	ES	DE	FR	PT	Avg.
EN	**75.63**	49.90	**49.90**	56.29	60.75	61.43
ES	75.30	**50.74**	48.05	55.41	60.92	61.07
DE	74.99	49.37	48.15	55.70	61.31	60.75
FR	75.58	49.25	49.03	**56.79**	60.91	61.21
PT	75.57	50.21	48.98	56.03	**61.64**	**61.48**

Table 4: Performance of PaLM2-L with 5-shot demonstration examples from various languages. We highlight the best demonstration example language (indicated by each row) for every test example language (indicated by each column).

dent demonstrations before prompting the LLM for direct fine-grained hallucination classes predictions; (3) **LI-C2FG** adopts the coarse-to-fine prompting technique with a fixed set of demonstrations; and (4) **LD-C2FG** combines both the coarse-to-fine prompting and language dependent demonstration selection techniques for fine-grained hallucination detection.

For all tested methods, we incorporate natural language explanations in a predict-then-explain pipeline (Lampinen et al., 2022) which outputs the explanations after the prediction and empirically works better than the Chain-of-Thought prompting (Wei et al., 2022). Furthermore, we prompt LLM to generate 4 predictions and use self-consistency (Wang et al., 2022) to aggregate them into the final prediction. Lastly, to reduce the randomness in LLM calls, we create 5 different groups of demonstrations for each k-shot setting, and report the average performance over 5 runs. Appendix A.4 provides more implementation details and hyper-parameter settings.

5.4 Experiment Results

Overall Results. Table 3 exhibits the main experiment results. First, we notice that increasing the number of demonstrations generally helps the model performance and has the most pronounced affects for small and medium sized LLMs. Second, comparing LD-FG with LI-FG and LI-C2FG with LD-C2FG reveals that language dependent demonstration selection indeed helps us to more accurately identify fine-grained hallucination classes. Third, we compare those coarse-to-fine prompting methods with the direct fine-grained prediction methods, and observe that the initial predicted coarse-grained class does guide the LLM for better fine-grained predictions. Finally, we note that even the best performing method (LD-C2FG with 5-shot PaLM2-L) still lags behind most supervised models with fewer parameters (c.f. Table 1). One reason could be the demonstrations in the prompt are not enough to fully convey the nuanced hallucination error type definitions. We encourage more future studies to fill this performance gap between the supervised and few-shot methods.

Effect of Demonstration Languages. We continue to evaluate how the language of demonstration examples affects the LLM hallucination detection performance. Specifically, we prompt the LLM with demonstrations of the same language and evaluate the prediction quality for each test example language. We report the 5-shot PaLM2-L performance (with coarse-to-fine prompting tech-

nique) in Table 4. First, we notice that forcing all demonstrations to have the same language will lead to worse performance (compared to 5-shot PaLM2-L LI-C2FG results in Table 3). Second, we observe that LLM generally performs better on those test examples that have the same language as its input demonstrations. The only exception is for German examples, it is better to prompt LLM with English demonstrations, which is somewhat understandable considering both languages share some grammatical similarities due to their common Germanic roots. This observation further explains and verifies the effectiveness of our language dependent demonstration selection strategy.

Effect of Prompting Methods. We continue to evaluate two additional variants of prompting methods: (1) Chain-of-Thought (CoT) which outputs the explanation before the final prediction and (2) Fine-to-Coarse (F2CG) that first outputs the fine-grained hallucination labels followed by a coarse-grained class. Results are shown in Table 5. First, we notice that for both language dependent and independent methods, CoT performs worse than the predict-then-explain pipeline. We hypothesize one reason could be the sparsity of explanation for non-hallucinated examples. Namely, if a headline is fully supported by the article, the raters will not provide any explanation and the CoT will have to directly make the prediction, a behavior inconsistent with the hallucination cases. Second, we find that coarse-to-fine prompting works significantly better than the fine-to-coarse prompting. This is probably because coarse-grained hallucination prediction has less mistakes then the fine-grained prediction. Therefore, it is better to condition on a more confident (i.e., coarse-grained) prediction for generating a less confident (i.e., fine-grained) prediction instead of the other way around.

Generalization to more LLMs. We test how our proposed prompting methods generalize to more LLMs including ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023)[4]. Results are exhibited in Table 6. We can see that prompting with ChatGPT generally performs worse than PaLM2-L while GPT4 outperforms the PaLM2-L on fine-grained hallucination detection. Furthermore, the experiment results still align with our prior findings that using language-dependent demonstrations and conducting coarse-to-fine prompting can consistently yield performance enhancements.

[4]More experiment details are described in Appendix A.6.

Methods	Accuracy	Weighted-F1	Example-F1
LI-FG	64.77	57.90	61.58
LI-FG + CoT	58.87	53.26	54.35
LI-C2FG	**66.69**	**62.25**	**62.05**
LI-F2CG	61.57	56.96	58.05
LD-FG	64.73	57.80	61.98
LD-FG + CoT	59.02	53.26	54.35
LD-C2FG	**67.03**	**62.90**	**62.14**
LD-F2CG	63.37	58.04	61.22

Table 5: Performance of PaLM2-L using 5-shot demonstration examples with different prompting methods.

Backbone	Accuracy	Weighted-F1	Example-F1
ChatGPT			
LI-FG	43.40	47.40	36.52
LD-FG	53.00	51.73	48.20
LD-C2FG	**54.83**	**53.53**	**50.88**
GPT4			
LI-FG	62.78	59.89	60.36
LD-FG	63.54	63.20	63.24
LD-C2FG	**65.88**	**64.05**	**64.59**

Table 6: The experiment results of ChatGPT and GPT-4 on 1-shot fine-grained headline hallucination detection with different prompting methods.

6 Related Work

Hallucination Detection. Hallucination, one long-standing issue for many natural language generation models, refers to the scenario where the generated content being nonsensical or inconsistent with the provided source content (Ji et al., 2022; Zhang et al., 2023; Chern et al., 2023; Tonmoy et al., 2024). Plenty of studies have been proposed to mitigate the hallucination issue by cleaning model training data (Nan et al., 2021a; Goyal and Durrett, 2021; Aharoni et al., 2022), modifying model learning objectives (Stiennon et al., 2020; Nan et al., 2021b), designing better decoding algorithms (Sridhar and Visser, 2022; Qiu et al., 2023a), and building specialized models to postprocess/filter generated contents (Cao et al., 2020; Chen et al., 2021; Shen et al., 2023; Manakul et al., 2023). At a high level, our study falls into the last category and focuses on multilingual fine-grained hallucination detection in the news domain.

Fine-Grained Hallucination Evaluation. A few studies are proposed to evaluate hallucination error details for different downstream applications. For text summarization, Goyal and Durrett (2020) proposes to categorize hallucination errors at the level of dependency arcs and Pagnoni et al. (2021) defines a hallucination typology based on frame semantics and linguistic discourse theory. For text

simplification, Devaraj et al. (2022) introduces a hallucination taxonomy based on the edit nature of simplification. At the same time, Dziri et al. (2022) leverages the Verbal Response Modes to define hallucination errors in knowledge-grounded conversational models. More recently, Mishra et al. (2024) proposes a hallucination taxonomy for open-ended LM generation without pre-determined grounding text. Despite some promising results, these approaches either assume one example can only have one fine-grained hallucination label or test only on English examples.

Multilingual Summarization Faithfulness. We can view the news headline generation as a special type of summarization and thus our study is also related to the research about improving faithfulness of multilingual summarization systems. Aharoni et al. (2022) proposes to train a coarse-grained entailment model based on multilingual natural language inference datasets (e.g., XNLI (Conneau et al., 2018), XTREME (Hu et al., 2020)) and adopt it to filter unfaithful summaries in the training set. Qiu et al. (2023b) introduces a multilingual faithfulness evaluation metric by aggregating four English faithfulness metrics with a machine translator. Different from these studies, our work does not rely on a translation system and focuses more on identifying fine-grained hallucination types.

7 Conclusions and Future Work

This study explores multilingual fine-grained headline hallucination detection and introduces the MFHHD dataset — a collection of over 11,000 expert-annotated examples across 5 languages with natural language explanations. Through extensive experiments, we discover that supervised models gain from pre-training on NLI datasets and the integration of explanations into their outputs. Additionally, LLM few-shot learners show improved performance when utilizing language-dependent demonstration selection and adopting a coarse-to-fine prompting strategy. Interesting future research directions include (1) employing parameter-efficient tuning techniques to directly train LLMs on the MFHHD dataset, (2) annotating some fine-grained hallucination classes at the span level, and (3) expanding the MFHHD dataset to incorporate more languages and explore multi-document hallucination detection scenarios.

Limitations

In this work, our primary goal is to identify the news headline hallucinations and thus define all fine-grained hallucination classes for the news domain applications. We recognize that some of these fine-grained definitions will be too restricted or too lenient for other domains' applications. How to effectively transfer the knowledge and signals in our MFHHD dataset to more general domain use cases would be an important research problem. Furthermore, for the few-shot detection setting, we mostly test those proprietary LLMs as they demonstrate the strongest in-context learning capability. Future work could explore whether and how various open-sourced LLMs can benefit most from our proposed prompting techniques.

Ethics Statement

This work adheres to high ethical standards in its research methodology and execution. We obtain the multilingual, fine-grained news headline hallucination detection dataset through a meticulous annotation process, ensuring the dataset quality. By addressing the issue of hallucination in automated news headline generation across multiple languages, the study contributes positively to the integrity and accuracy of news dissemination.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. In *ACL*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NeurIPS*, abs/2005.14165.

Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and

Dong Yu. 2023. Generating user-engaging news headlines. In *ACL*.

Mengyao Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. *EMNLP*, abs/2010.08712.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *NAACL*.

Ethan Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. *ACL*, 2022:7331–7345.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R. Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *EMNLP*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *NAACL*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *ACL Findings*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *NAACL*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories. *WebConf*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ICLR*, abs/2111.09543.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *NAACL*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ICML*, abs/2003.11080.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Andrew Kyle Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *EMNLP*, abs/2204.02329.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *EMNLP*.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*, abs/2303.08896.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level factual consistency of abstractive text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021b. Improving factual consistency of abstractive summarization via question answering. *arXiv preprint arXiv:2105.04623*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *ArXiv*, abs/2004.14546.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *ACL*. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT.

OpenAI. 2023. GPT-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *NAACL*, abs/2104.13346.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *NeurIPS*, abs/2105.11447.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *WebConf*, pages 993–1002.

Yifu Qiu, Varun Embar, Shay B. Cohen, and Benjamin Han. 2023a. Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. *ArXiv*, abs/2311.09467.

Yifu Qiu, Yftah Ziser, Anna Korhonen, E. Ponti, and Shay B. Cohen. 2023b. Detecting and mitigating hallucinations in multilingual summarisation. *EMNLP*, abs/2305.13632.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Jiaming Shen, Jialu Liu, Daniel Finnie, Negar Asgharipour Rahmati, Michael Bendersky, and Marc Najork. 2023. "why is this misleading?": Detecting news headline hallucinations with explanations. *WebConf*.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. Taxoclass: Hierarchical multi-label text classification using only class names. In *NAACL*.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization. *arXiv preprint arXiv:2212.02712*.

Nisan Stiennon, Ouyang Long, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Francis Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *ArXiv*, abs/2401.01313.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ICML*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *NeurIPS*, abs/2201.11903.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Huggingface's transformers: State-of-the-art natural language processing. *EMNLP (Demo)*, abs/1910.03771.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

A Appendix

A.1 Fine-grained Hallucination Types with Illustrative Examples

We show one representative example for each fine-grained hallucination class in Figure 5.

A.2 Discussions on Headline Sources

As discussion in main text, we collect our headlines from a system that utilizes NHNet for headline generation. We acknowledge that this is not a very diverse set of model generated headlines. Meanwhile, we want to emphasize that these headlines closely resemble real-world news headlines.

To further test if our model can generalize to other headline generation methods, we conduct an experiment where we first randomly select 25 examples from our MFHHD test split and use PaLM2-L and GPT4 to generate a headline that has the same fine-grained label(s) as the original headline. Then, we manually check that those generated headlines indeed have their corresponding labels (and if not, we will slightly modify their wordings to make them correctly labeled). Finally, we test our SFT model variant "fine-grained mT5$_{xxl}$ + NLI + Exp" (see Table 1 in the main text) and observe that it achieves accuracy = 0.64, example-F1 = 0.59. These results show that our model can generalize to more headline generation methods to some extent.

A.3 Experiment Details on Supervised Hallucination Detection Methods

For mDeBERTa$_{base}$[5] and mDeBERTa$_{base}$ + NLI[6], we use their corresponding pre-trained checkpoints in the Huggingface Library. We do parameter swamping on the learning rate in [5e-6, 1e-6, 5e-5, 1e-5] and perform three-fold cross validation on the training set. The final selected learning rate for mDeBERTa$_{base}$ is 1e-6 and learning rate for mDeBERTa$_{base}$ + NLI is 5e-6. Finally, we train both models on a single A100-40GB with batch size 8 for 3 epochs.

For the remaining four mT5$_{xxl}$ based models, we implement them using the T5X library[7] with pre-trained mT5 checkpoints[8]. due to computational constraints, we directly use their default hyper-parameters. Specifically, we set the batch size to be 128, constant learning rate to be 1e-3, and the maximum output tokens to be 128. If the output sequence exceeds the length limit (e.g., having a long human written explanation in mT5$_{xxl}$ + Exp), we will simply truncate the output sequences to its first 128 tokens. We train all four models on TPU v3 for 10k steps with 1k warmup steps.

A.4 Experiment Details on Few-shot Hallucination Detection Methods

Demonstration Selection. We implement all of our k-shot experiments in the true few-shot setting (Perez et al., 2021) for the multi-class classification problem. Specifically, we will sample k demonstrations for each coarse-grained hallucination label and do not assume the presence of a large labeled development set of hyper-parameter tuning. Namely, we will have 3 demonstration examples for 1-shot setting, 9 demonstrations for 3-shot setting, and 15 demonstrations for 5-shot setting. For the language-independent prompting methods, we use the same set of demonstrations for all test examples. For the language-dependent prompting methods, we will sample the corresponding number of demonstrations for each language and dynamically choose the demonstration set based on the test example language. Furthermore, to reduce the LLM call randomness, we repeat the above sampling procedure 5 times and report the averaged performance over these 5 independent runs.

Explanation Order. In our experiments, we test both the chain-of-thought (CoT) prompting (Wei et al., 2022) which generates explanations *before* making predictions and predict-then-explain prompting (Lampinen et al., 2022) which outputs explanations *after* making predictions. For our hallucination detection task, we observe that the predict-then-explain prompting consistently outperforms CoT prompting, particularly for small and medium sized LLMs. Therefore, we choose to use the predict-then-explain prompting in this work.

We hypothesize that the ineffectiveness of CoT in our task comes from two aspects. First, we do NOT manually write any CoT demonstration and directly use the expert written explanations. These explanations may not be the best rationales for CoT and thus impair its performance. Second, we think there is an intrinsic difference between the hallucination detection task and the math reasoning

[5] https://huggingface.co/microsoft/mdeberta-v3-base

[6] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

[7] https://github.com/google-research/t5x

[8] https://github.com/google-research/multilingual-t5

Fine-grained Hallucination Label	Generated Headline	News Article	Comment
Neutral (extra info)	Meghan McCain criticizes vaccine	Meghan McCain Defends Stance On Coronavirus Vaccine Following Further Backlash. Meghan McCain is drawing ire for her latest comments on “The View”. The controversy began while McCain, 36, and the other co-hosts were discussing the COVID-19 vaccine rollout on Monday's episode ...	The article doesn’t explicitly say her stance (support or not) and thus the additional information in the headline about “criticizes” is not supported
Neutral (missing info)	Comcast to increase internet fees	Comcast will charge customers more for heavy internet usage starting next year. Comcast Corp. will charge more for heavy users of home internet in Northeast states—including Pennsylvania and New Jersey—angering customers who work and study online due to the pandemic. The vast majority of Comcast's Xfinity customers won't be affected by the "data threshold" next year, company officials said this week. ...	The charge increases are only for heavy Internet users and this key information is missed in the headline
Neutral (off topic)	Meghan Markle holds Archie in South Africa	Meghan Markle Continues Africa Tour with Visit to Girls' Club to Address Sexual Violence in Schools. Meghan Markle Visits Girls' Club to Tackle Sexual Violence, as she is tackling important issues in education during her day of solo outings on Tuesday ...	Although both article and the headline are about Meghan Markle. They are focused on completely different news stories.
Neutral (others)	School closings in danville	Jackson County school closes. MARIANNA, Fla. (WJHG) - Marianna's Dayspring Christian Academy is closing their doors for the remainder of the year. Administrator Randy Ward said in a press release, ...	Based on the article information, we don’t know if Danville is related to Jackson County
Contradictory (opinion as fact)	Apple to launch with iPad Air	Apple's Fall Product Releases Rumored to Include New iPad Air and Two Apple Watches. Marking the beginning of September, the month Apple typically announces its upcoming product releases, tech insiders report that customers can expect a new iPad Air as well as two new versions of the Apple Watch ...	The article says the new iPad Air is “rumored” to be released, while the headline directly says “to launch”.
Contradictory (wrong number)	Amazon CEO Jeff Bezos sells $1.9 billion in shares	Bezos Sells \$3.1 Billion Of Amazon Shares After Wealth Jumps. The numbers are eye-popping: 1 million Amazon.com Inc. shares offloaded for more than \$3.1 billion. And yet for the seller, Jeff Bezos, it barely puts a dent in his stake in the e-commerce giant...	The article and headline disagree in the quality ($3.1 billion vs $1.9 billion) of stock sold by Jeff Bezos and thus quality is an important key number.
Contradictory (others)	Auburn loss to Tulane	What Gus Malzahn said about Auburn’s 24-6 win vs. Tulane. Auburn defeated Tulane, 24-6, in the Tigers’ home opener at Jordan-Hare Stadium on Saturday night ...	Tulane lost to Auburn, not the other way around. So the headline directly contradicts the article.

Figure 5: Fine-grained headline hallucination labels with illustrative examples.

task (e.g., GSM8K). The former one does not really require complicated multi-step reasoning and typically reaches the final decision in one or two steps. For example, if a rater witnesses an entity mismatch or “rumor as fact" statement, he/she will directly label the headline as contradictory. This effectively shrinks the CoT improvement room.

Self Consistency. Wang et al. (2022) proposes the self-consistency method which samples multiple output sequences from the LLM and aggregates them into the final prediction. In our experiments, we set the default temperature $t = 0.7$ and sample 4 decoded sequences from the LLM. For each decoded sequence, we first parse it into a set of fine-grained labels. Then, we select all fine-grained labels that appear in at least 2 decoded sequences as the final predicted hallucination labels.

A.5 Prompt Templates

Listing 1: Prompt Template for Direct Fine-Grained Headline Hallucination Prediction

```
You work as a news journalist.
Given a news article and a news
headline, you need to determine
the relation between this article
 and the headline. Possible
relations include: ["match", "
incorrect_number", "
opinion_as_fact", "
direct_contradiction", "
unsupported_additional_info", "
miss_important_info", "off_topic
", "
neither_support_nor_contradiction
"].

Please first output all possible
relations and then provide an
explanation.
See the following examples for
references.

# demonstrations
Article: [article]
```

```
Headline: [headline]
Output: [class]

Now, you are given a new news
article and a news headline.
Think step by step and then make
the prediction.

# test examples
Article: [article]
Headline: [headline]
Output: [class]
```

Listing 2: Prompt Template for Coarse-to-Fine Headline Hallucination Prediction

```
You work as a news journalist.
Given a news article and a news
headline, do you think the
article fully supports the
headline? Please first output "
Yes", "No", or if not sure, "
Maybe".
If output "No", please list all
detailed reasons from ["
opinion_as_fact", "
incorrect_number", "
direct_contradiction"] and
provide an explanation.
If output "Maybe", please list
all detailed reasons from ["
miss_important_info", "
unsupported_additional_info", "
off_topic", "
neither_support_nor_contradiction
"] and provide an explanation.
See the following examples for
references.

# demonstrations
Article: [article]
Headline: [headline]
Output: [class]

Now, you are given a new news
article and a news headline.
Think step by step and then make
the prediction.

# test examples
Article: [article]
Headline: [headline]
Output: [class]
```

A.6 Experiments on ChatGPT and GPT4

We use the `gpt-3.5-turbo-0125` model for experimenting ChatGPT and adopt the `gpt-4-0125-preview` model for GPT4. Due to the budget considerations, we only decode 1 generation from each model and remove the self consistency aggregation. When experimenting coarse-to-fine prompting using ChatGPT, we observe that the decoded sequence occasionally fails to follow the ideal output format (c.f. Figure 4) and we will re-query the model using the JSON mode.