Aligning Uncertainty: Leveraging LLMs to Analyze Uncertainty Transfer in Text Summarization

Zahra Kolagar
Fraunhofer IIS
Erlangen, Germany
zahra.kolagar@iis.fraunhofer.de

Alessandra Zarcone
Fraunhofer IIS, Erlangen, Germany
Technische Hochschule Augsburg, Germany
alessandra.zarcone@tha.de

Abstract

Automatically generated summaries can be evaluated along different dimensions, one being how faithfully the uncertainty from the source text is conveyed in the summary. We present a study on uncertainty alignment in automatic summarization, starting from a two-tier lexical and semantic categorization of linguistic expression of uncertainty, which we used to annotate source texts and automatically generate summaries. We collected a diverse dataset including news articles and personal blogs and generated summaries using GPT-4. Source texts and summaries were annotated based on our two-tier taxonomy using a markup language. The automatic annotation was refined and validated by subsequent iterations based on expert input. We propose a method to evaluate the fidelity of uncertainty transfer in text summarization. The method capitalizes on a small amount of expert annotations and on the capabilities of Large language models (LLMs) to evaluate how the uncertainty of the source text aligns with the uncertainty expressions in the summary.

1 Introduction and Motivation

Uncertainty is a multifaceted construct and can stem from various sources. It may stem from a lack of knowledge or information or constraints in data availability (epistemic uncertainty), or variability or noise in the data (aleatoric uncertainty) (Lahlou et al., 2023; Hüllermeier and Waegeman, 2021; Sankararaman and Mahadevan, 2011; Hofer et al., 2002). Or it can result from a model's limitation or approximation, preventing it from perfectly representing the underlying data patterns, whether due to inherent model constraints or approximations (Kuhn et al., 2023).

Different linguistic expressions or textual elements may convey uncertainty, suggesting doubt, possibility, ambiguity, or a lack of precision (Auger and Roy, 2008; Juanchich et al., 2017; Walley and De Cooman, 2001) and ultimately influencing comprehension and decision-making processes (Wang et al., 2018; Juanchich et al., 2017; Gkatzia et al., 2016). Understanding and effectively conveying uncertainty within textual content is a crucial aspect of natural language processing (NLP) and has been explored in downstream tasks such as text and document classification (Hu and Khan, 2021; Mukherjee and Awadallah, 2020; Chen et al., 2020; He et al., 2020; Zhang et al., 2019), question-answering (Ünlü and Arisoy, 2021; Li et al., 2021; Lyu et al., 2020), and natural language generation (NLG, Kuhn et al., 2023; Xiao and Wang, 2021; Rieser and Lemon, 2009) among others.

Despite the emergence of advanced methodologies in NLP, including pre-trained models like GPT-3/4 (Koubaa, 2023; OpenAI, 2023), BLOOM (Scao et al., 2022; Science, 2023), Llama models (Touvron et al., 2023), as well as specialized variants such as InstructGPT (Ouyang et al., 2022), ChatGPT (OpenAI, 2022), and Falcon-40B-instruct (Almazrouei et al., 2023; Penedo et al., 2023; Xu et al., 2023), the identification and assessment of uncertainty remain difficult tasks. These advancements have notably enhanced NLG performance; however, they have also introduced new dimensions for uncertainty exploration and investigation, including but not limited to issues such as hallucination (Zhang et al., 2023b,a; Chen et al., 2023; Ji et al., 2023), factuality and truthfulness (Raj et al., 2023; Quelle and Bovet, 2023; Augenstein et al., 2023), logical reasoning and self-consistency (Xiong et al., 2023; Chen and Mueller, 2023; Cheng et al., 2023) within LLM-generated output.

Text summarization aims to distill comprehensive information into shorter, more concise versions while retaining the essential information and

Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024), pages 41–61

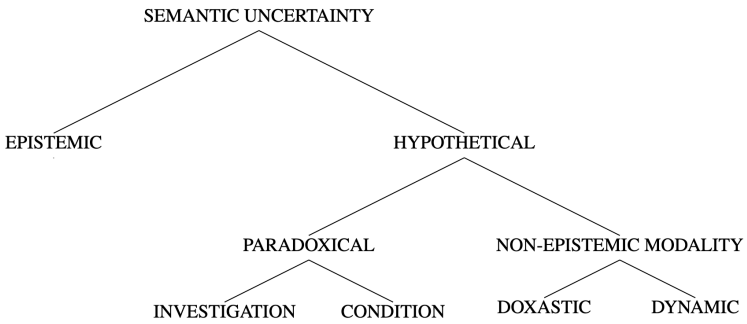

Figure 1: Categorization of semantic uncertainty introduced by (Vincze, 2014b)

preserving coherence (Allahyari et al., 2017; El-Kassas et al., 2021; Nenkova and McKeown, 2012). Uncertainty is important in the evaluation of summaries, as it directly impacts the fidelity and accuracy of condensed information. Recognizing and appropriately handling uncertainty expressions within a source text and effectively transferring them to summaries is crucial for ensuring the integrity and relevance of the distilled information (Zablotskaia et al., 2023; Xu et al., 2020).

Our work is guided by the linguistic taxonomy of uncertainty described in Section 3, and tries to answer the following research questions:

- RQ 1: How can LLMs be employed to identify and annotate expressions of uncertainty in text based on the taxonomy introduced in Section 3?
- RQ 2: How faithful are LLM-generated summaries regarding the dimension of uncertainty, and how do the uncertainty expressions in the summary align with the corresponding expressions in the source texts?

Firstly, we aim to establish a linguistically oriented taxonomy of uncertainty in textual content, building upon previous work (Section 2.1), to be used as a simplified ontology for text annotation. The taxonomy is described in Section 3. Section 4 describes the material gathered from various sources and describes the data annotation process. Finally, in Section 5 we evaluate the fidelity of uncertainty transfer in summarization processes.

2 Background and Related Work

2.1 Uncertainty from a Linguistic Perspective

The linguistic conceptualization and nuances of uncertainty find their origins in philosophy, particularly in decision theory work by Luce and Raiffa (1989). They delineate three key situations: certainty, risk, and uncertainty, each depending on the probabilities and possibilities associated with the potential consequences of an action. Bradley and Drechsler (2014) further detail three distinctions based on the nature of judgments, the object of judgments, and the severity of uncertainty reflected in the experience of the agent.

Rubin (2006) categorizes linguistic expressions of uncertainty based on proximity to certainty and proposes a four-dimensional model involving certainty level, perspective, focus, and time (cf. Fig. 4 in the Appendix). Expanding on Rubin's work, Szarvas et al. (2012) and later Vincze (2014b) present a classification of uncertainty across semantic, discourse, and pragmatic levels. At the semantic tier, propositions are deemed uncertain under truth-conditional semantics, branching into epistemic (uncertainty due to the lack of knowledge) and hypothetical uncertainties with hypothetical further branching into "investigation" (uncertainty in exploring certain aspects or information) and "condition" (uncertainty about certain conditions or criteria) under paradoxical and "doxastic" (uncertainty about beliefs or opinions) and "dynamic" (uncertainty associated with variability or change) under non-epistemic modality (Fig. 1).

EPISTEMIC: It **may** be raining.

DYNAMIC: I **have to** go.

DOXASTIC: He **believes** that the Earth is flat.

INVESTIGATION: We **examined** the role of NF-Kappa B in protein activation.

CONDITION: **If** it rains, we'**ll** stay in.

The discourse level concentrates on sources' fuzziness and subjectivity, categorizing expressions

such as "hedges", "weasels" and "peacocks" as shown in examples below (Vincze, 2014a,b, 2013; Ganter and Strube, 2009).

WEASEL: **Some** note that the number of deaths during confrontations with police is relatively proportional for a city the size of Cincinnati.

HEDGE: Magdalene Asylums were a **generally** accepted social institution until well into the second half of the 20th century.

PEACOCK: The main source of their inspiration was native Georgia, with its **rich** and **complex** history and culture, its **breathtaking** landscape and its **courageous** and hardworking people.

Pragmatic uncertainty arises when speakers obscure their evidence or source, violating conversational maxims of informativeness and evidence provision (Grice, 1975). Auger and Roy (2008) expands these categories to encompass both linguistic and extra-linguistic environments of information.

2.2 Linguistic Uncertainty Detection in NLP

Detection of linguistic cues of uncertainty in NLP was first systematically introduced in CoNLL-2010 Shared Task, centering on identifying uncertainty cues in English biological papers and Wikipedia articles (Farkas et al., 2010). Earlier work on uncertainty detection has mostly focused on rule-based approaches (Light et al., 2004; Chapman et al., 2007), followed by supervised approaches (Morante et al., 2009; Morante and Sporleder, 2012; Farkas et al., 2010; Vincze, 2014a).

With the emergence of LLMs and their remarkable generation capabilities, research on linguistic uncertainty expression has taken multiple paths. One approach has aimed at prompting models to gauge their confidence levels. Lin et al. (2022) introduced the concept of vanilla verbalized confidence by prompting LLMs to both generate answers and express uncertainty. Prompt strategies were suggested by van der Gaag et al. (2013) (CoT prompt strategy) and Tian et al. (2023) (Top-K prompt strategy). Additionally, exploring human-like behavior in the face of uncertainty, methods such as self-consistency extension by Wang et al. (2022) have been pursued.

Another strand of research has focused on uncertainty estimation in Question Answering tasks with LLMs, with Si et al. (2022) introducing logit calibration. Kuhn et al. (2023) introduced semantic entropy to handle uncertainty in Question Answering by incorporating linguistic invariances. Tian et al. (2023) evaluated computationally feasible methods to extract confidence scores from probabilities output by Reinforcement Learning-trained LLMs, and Chen and Mueller (2023) introduced BSDETECTOR for detecting speculative answers. Baan et al. (2023) emphasizes a more principled treatment of uncertainty, characterizing major sources of uncertainty in NLG, and proposes a taxonomy of uncertainty linked to the data and the model.

To our knowledge, although various methodologies have been introduced to investigate linguistic uncertainty cues within textual data, particularly in the realm of NLG, none of these endeavors have specifically concentrated on the task of summarization, in identifying linguistic elements of uncertainty and understanding their fidelity and transfer from the article to the generated summaries.

2.3 Uncertainty Annotated Datasets

Corpora across various domains and linguistic levels have undergone annotation for uncertainty expressions. Concerning different domains, these include biology (Nawaz et al., 2010; Kim et al., 2008; Settles et al., 2008; Shatkay et al., 2008; Vincze et al., 2008; Medlock and Briscoe, 2007), medicine (Uzuner et al., 2009), news (Rubin, 2010; Rubin et al., 2006), encyclopedic content (Vincze, 2014b; Farkas et al., 2010), reviews (Konstantinova et al., 2012; Díaz, 2013), and social media (Wei et al., 2018). At the linguistic level, Medlock and Briscoe (2007) annotated hedge phrases, while Vincze (2013) annotated for weasels, hedges, and peacock expressions. The CoNLL 2010 dataset also includes hedge phrases and weasels (Farkas et al., 2010). Furthermore, Vincze (2014b) focused on semantic category annotation across various corpora while Rubin (2010) annotated for epistemic modality in the context of information seeking, contributing to the exploration of linguistic uncertainty. While each dataset contributes valuable insights into understanding uncertainty, it is important to note that each dataset may capture certain annotations while potentially missing others, leading to a varied representation of linguistic uncertainty. Furthermore, the categorization of uncertainty expressions across these datasets may exhibit overlaps and variations, indicating that the definition and

taxonomy of such categories are not universally standardized

3 A Two-tier Linguistic Taxonomy of Uncertainty

This section delineates our linguistic taxonomy of uncertainty to annotate linguistic cues in textual data. We extend previous linguistic frameworks for categorizing uncertainty, aiming at developing a comprehensive yet easily adaptable framework encompassing various linguistic levels of uncertainty representation in text. The taxonomy classifies representations of uncertainty concerning the lexical material which conveys the uncertainty (lexical) and the type of uncertainty (semantic).

At the lexical level, our taxonomy distinguishes between uncertainty expressions at word level, phrase level, sentence level, section level, and discourse level (with discourse referring to the entire analyzed text). Specifically, we elaborate on word- and phrase-level uncertainty expressions based on the grammatical functions of words or phrases within their units. At the word level, we identify parts of speech such as adjectives, adverbs, auxiliaries, verbs, conjunctions, and nouns. At the phrase level, we look at adjective phrases, adverbial phrases, noun phrases, prepositional phrases, verb phrases, conjunctional phrases, infinitive phrases, and participle phrases (cf. Fig. 5 in the Appendix). For this research, we only focus on uncertainty expressions at the word and phrase level.

At the semantic level, we focus on the semantic uncertainty distinctions outlined by Szarvas et al. (2012) and Vincze (2013, 2014b,a), which includes categories like epistemic, dynamic, doxastic, investigation, and condition (cf. Fig. 6 in the Appendix). We deliberately exclude discourse-level uncertainty expressions such as hedges, weasels, and peacocks to avoid dependencies on external information or knowledge beyond the provided task information. This decision is made to streamline the annotation process and prevent potential complexities in evaluating the summaries at a later stage.

4 Data Acquisition and Annotation

4.1 Data Collection and Preprocessing

We extracted data from "Education Week" [1], an educational website featuring a variety of articles on educational topics, as this was part of a larger research project related to the topic of education. Additionally, we gathered content from "An Easy & Proven Way to Build Good Habits & Break Bad Ones" [2] website, a personal blog authored by James Clear, known for his expertise as a life coach. The reason for adding data from a personal blog was that the text normally contains linguistic expressions of uncertainty. This decision serves to add variety to the data collected from the educational website.

After conducting a minimal preprocessing step, we filtered the articles to a word count range of 600-700 words. This decision serves two objectives: to control the variation in text length, which may lead to significant statistical differences in uncertainty expressions across articles, and to facilitate more consistent and informed human evaluations, as previous research suggests that time-constrained human assessments with large texts may yield inconsistent results (Krishna et al., 2023). Ultimately, we acquired 150 article samples.

4.2 Generating Summaries

We provided a straightforward prompt to OpenAI's GPT4-8k, instructing it to generate summaries of the articles within a maximum limit of 200 words.

4.3 Uncertainty Annotation Leveraging LLMs

As outlined in Section 3, a starting point to understand LLM performance in transferring uncertainty expressions to summaries is to annotate articles and summaries. We propose a markup-based annotation syntax exemplified in Figure 2, inspired by Yamauchi et al. (2023). The annotation relies on XML syntax, employing a structured set of elements enclosing content within a start tag <Uncertainty> and an end tag </Uncertainty>, with two attributes: "POS" and "semantic".

The financial market appeared to be <Uncertainty POS="Adjective phrase" semantic="dynamic">highly unstable</Uncertainty>.

Figure 2: Sample annotation of uncertainty in text using XML syntax

Utilizing a markup language allows for precise and fine-grained annotation, enabling the categorization of uncertainty expressions based on the lexical and semantic categories. Furthermore, the predefined markup structure ensures consistency and standardization in annotation practices across

[1] https://www.edweek.org/

[2] https://jamesclear.com/

Even the best models of the world are <Uncertainty POS="adjective" semantic="epistemic">imperfect</Uncertainty>. This insight is important to remember if we want to learn how to make decisions and take action on a daily basis. For example, consider the work of Albert Einstein. During the ten-year period from 1905 to 1915, Einstein developed the general theory of relativity, which is one of the most important ideas in modern physics. Einstein's theory has held up remarkably well over time. For example, general relativity predicted the existence of gravitational waves, which scientists finally confirmed in 2015—a full 100 years after Einstein originally wrote it down. However, even Einstein's best ideas were <Uncertainty POS="adjective" semantic="epistemic">imperfect</Uncertainty>. While general relativity explains how the universe works in many situations, it breaks down in certain <Uncertainty POS="adjective" semantic="condition">extreme cases</Uncertainty> (like inside black holes).

Figure 3: Sample GPT-4-8K annotated text from the dataset.

	Errors per Category				Tot. Reviewed	GPT-4 Annotation
Correct Elements	Missing Elements	Semantic Attribute Error	POS Attribute Error	Span Error	Elements	Accuracy
189	39	49	29	15	321 (107 tags)	58.8 %

Table 1: Expert evaluation of GPT-4 annotation for 15 selected articles led to the review of 321 elements across three categories: semantic, POS, and annotation spans. This process introduced 13 new attributes (39 elements).

different datasets and annotators. Lastly, the compatibility of markup annotations with various text processing tools and their seamless integration into NLP workflows significantly enhances their usability across diverse applications.

We employed the GPT4-8k model to systematically generate uncertainty annotations for the articles and the summaries. We used a structured prompt template containing specific components: an "instruction" guiding the model through the annotation task, a "context" providing taxonomy descriptions and categories, an "input example" illustrating a text excerpt, and an "output example" showcasing the desired output format using the markup language (cf. Fig. 7 in the Appendix). We presented the model with articles and summaries from the dataset described in Section 4.1, each preceded by this prompt as a prefix. The resulting annotations were saved alongside the unlabeled articles (cf. Fig. 3).

4.4 Uncertainty Annotation Evaluation and Refinement

To ensure the accuracy of the annotations created using the GPT4-8k model, we adopted two approaches: an expert evaluation and review, and a self-refinement process guided by expert judgments.

Expert Evaluation Initially, we randomly selected 15 out of 150 samples from our annotated dataset and engaged two linguists as experts to review these annotations. Their task was to examine the annotations for consistency and correctness based on the span's accuracy, the correct POS, and semantic labeling. The experts categorized the extracted examples as either correct or incorrect, introducing a new "evaluation" element under the <Uncertainty> tag, and were asked to generate the correct annotations for the incorrect ones. Furthermore, we requested that they provide explanations for incorrect annotations, which we also collected. The task instructions for the expert evaluation are reported in Fig. 8 and a sample expert correction is shown in Fig. 9.

Our initial prompt resulted in 94 uncertainty annotation tags across the 15 selected samples. Given that the experts needed to evaluate the annotation span, as well as the POS and semantic attributes, we have a total of 282 elements (span and attributes) to review. Table 1 illustrates the results of the refined annotations performed by the linguists across the 15 selected samples on these 94 tags and 282 elements, which resulted in 107 tags (and 213 elements) after the expert review and the addition of missing tags and elements.

Expert Guided Self-Refinement Using Post-hoc Prompting Previous studies have shown that relying solely on LLMs for self-evaluation leads to suboptimal outcomes due to their limited self-assessment capabilities (Kolagar et al., 2023). To enhance the model's self-correction through reasoning, we leveraged both correct and incorrect annotations alongside the associated rationales provided by expert linguists.

Kim et al. (2023) and Shinn et al. (2023) propose a three-step self-correction prompting approach where the initial response serves as the standard prompt and is then reviewed by the model using a review prompt, where the model is asked to produce feedback on the previous response, followed by the model's new response to the original question with the feedback produced by the model itself. Huang et al. (2023) however, observed that the self-correction behavior of the model does not yield an improvement in the original reasoning of the model when the external feedback (oracle) is removed from the self-correction process. Hence, they suggest integrating human-provided labels to enhance the model's reasoning based on its own generated response.

We adopted a post-hoc prompting method similar to Huang et al. (2023), Kim et al. (2023), and Shinn et al. (2023), but adapted and refined their prompting strategy to suit our annotation refinement task, introducing the subsequent elements:

- The "context" used for the initial prompting
- The annotated text with examples of correct and incorrect labels
- The annotations provided by the experts for missing elements
- The accompanying justifications and corrections provided by the experts

After conducting two rounds of self-review and refinement with the model on each of the 15 expert-refined samples, we noticed remarkable improvements in the model's ability to fully correct its initially generated annotations. The performance enhancements observed align with the outcomes discussed in Huang et al. (2023), Kim et al. (2023), and Shinn et al. (2023), indicating the model's capacity for self-correction and improvement. However, fewer rounds of post-hoc self-correction are required when the model is provided with the reasoning in addition to the correct and incorrect labels. Table 2 displays the model's correction accuracy following the self-refinement process for the 15 samples.

Building on this progress, we extended the three-round review and refinement process to the annotation of the remaining articles in our dataset as well as to the summariesAs those were not annotated by our experts, we incorporated excerpts from the refined 15 expert annotations as examples into the prompt to guide the model's self-correction (Fig. 11).

Stages	GPT-4 Accuracy
After expert assessment	58.8%
After the 1st round	89.3%
After the 2nd round	100%

Table 2: GPT-4 annotation accuracy at different stages of the post-hoc refinement process compared to the base results after expert assessment.

We recognize a potential concern that the self-correction ability of the model could diminish once the expert refinement guidance is withdrawn during post-hoc assessments. To investigate this, from the pool of 135 remaining data after two rounds of refinement, we randomly selected 2 articles for expert assessment by linguists. Their evaluation revealed a decrease in the model's refinement accuracy to **80.4%** (76.8 % for semantic attribute). We consider this accuracy acceptable for the analysis discussed in section 5.1.

4.5 The Dataset

The final dataset comprises the articles with an average of 653.7 words, the summaries with an average of 153.5 words, the annotations, a sample subset of corrected annotations verified by linguistic experts, and the enhanced annotations after each round of post-hoc refinement.

5 Analyzing Uncertainty Transfer in Summarization

5.1 Evaluation of Uncertainty Representation in Summaries

In this study, our attention is drawn to evaluating the transference of uncertainty representation in the summaries. We concentrate solely on analyzing how well **semantic** annotation of uncertainty expressions is conveyed in the summaries based on the 5 semantic labels outlined in Section 3. We exclude the analysis of POS in this evaluation, as POS alterations might occur in summarization without necessarily affecting the fidelity of uncertainty expressions. We also exclude a comprehensive evaluation of other summary quality aspects, as previous research confirms that LLMs excel in the task of summarization, often surpassing human-generated summaries in terms of preferred outcomes (Goyal

Semantic Type	Instances in the Article	Matches in the Article	Instances in the Summary	Matches in the Summary	Precision	Recall
Epistemic	795	485	356	243	0.68	0.50
Dynamic	163	96	55	31	0.56	0.32
Doxastic	197	122	89	61	0.68	0.50
Investigation	221	133	97	79	0.81	0.59
Condition	49	36	35	12	0.34	0.33
Total	1425	865	632	426	0.67	0.49

Table 3: Precision and Recall calculated across all semantic categories for total found and matched instances in the articles and the summaries.

et al., 2023; Kolagar et al., 2023; Pu et al., 2023; Zhang et al., 2023c).

To execute this analysis, we need to align sentences or clauses containing uncertainty annotation in the summary to the corresponding sections in the article. For this, we use semantic similarity scores between sentences. Initially, we performed sentence segmentation on the summaries using SpaCy's Sentencizer for English (Honnibal et al., 2020) [3]. Since sentences may encompass multiple annotations, including coordinated conjunctions or clauses, we further identified the boundaries of each clause, using the main verb (ROOT dependency tag) analysis provided by SpaCy's dependency parser [4]. For the article, we divided it into sections containing around 20-30 words, ensuring sections concluded at a full stop to avoid mid-sentence breaks. The reason for that is that the summaries may refer to longer sections rather than individual sentences or clauses in the article.

We used Sentence-Bert (Reimers and Gurevych, 2019) [5] to conduct a semantic similarity analysis between segments in the summary containing a label and the sections in the article. This process aimed to identify the section in the article most closely related to the summaries. We then evaluated the highest-ranking section to ascertain the presence of a label within it.

We compute precision and recall specifically when there's a precise match, signifying an **exact alignment** between a semantic label in the summary and one or more identical labels in the article, for the section in the article where the summary stems from. The following cases were identified between the matched instances (cf. Fig. 12 for samples of the identified cases):

1. No annotation was found in the matched section of the article, resulting in a score of 0.
2. One annotation was found in the matched section of the article containing the same semantic label, resulting in a score of 1.
3. One annotation was found in the matched section of the article containing a different semantic label, resulting in a score of 0.
4. Multiple annotations were found in the matched section of the article containing the same label, resulting in a score of 1.
5. Multiple annotations were found in the matched section of the article containing different labels, resulting in a score of 0.

We have discovered a total of 1425 semantic labels within the article and 632 semantic labels within the summaries. Among these 632 labels, there were 425 exact matches found corresponding to 865 instances in the articles. Consequently, the precision and recall were computed as follows. Precision measures the accuracy of the **aligned labels** in the summary concerning the **total labels in the summary**, while recall measures the coverage of the **aligned labels** in the summary concerning the **total labels in the article sections** that have matches in the summary.

$$\text{Precision} = \frac{\text{Number of aligned labels in summary}}{\text{Total labels in summary}}$$

$$\text{Recall} = \frac{\text{Number of aligned labels in summary}}{\text{Total labels in the matched sections of article}}$$

Table 3 shows precision and recall results for the different uncertainty classes as well as general precision and recall.

5.2 Discussion

We need to highlight two crucial aspects. Firstly, we did not account for the ranking or significance

[3] https://spacy.io/api/sentencizer

[4] https://spacy.io/usage/linguistic-features#dependency-parse

[5] https://www.sbert.net/docs/usage-/semantic_textual_similarity.html

of uncertainty expressions in the article and summaries; our focus remained solely on alignment. We assigned equal importance to all expressions for precision and recall calculations, assuming that only relevant and vital information appears in the summaries. However, a more accurate assessment requires further exploration into the significance and hierarchy of these expressions.

Secondly, the automatic annotation; even after the self-refinement procedure using expert annotations and revisions, still yielded a lower accuracy on the randomly selected articles, potentially influencing the precision and recall outcomes. Variations in precision outcomes seem to also arise from the differing number of semantic types available in the article and summary. Conversely, the lower recall is acceptable, considering that the frequency of uncertainty expressions is much less in the summaries. This demonstrates the challenging nature of aligning uncertainty between source text and summaries.

Notwithstanding these constraints, our methodology demonstrates how precision and recall metrics can be used to assess the summary's faithfulness to the source text, providing an evaluation approach to assess the effectiveness of LLM-based summarization. Our analysis emphasizes the need for further exploration in evaluating summaries, particularly in domains requiring uncertainty alignment, particularly in safety-critical scenarios such as summarizing medical reports.

6 Conclusion and Future Work

In this study, we provided a framework to evaluate automatic summarization. We introduced a two-tier annotation taxonomy that categorizes linguistic uncertainty expressions within the text, emphasizing lexical and semantic expressions, and developed an XML-based syntax framework to standardize the annotation process for these expressions. We conducted experiments involving expert linguists to refine annotations and utilized their expert rationale to guide the LLM's self-evaluation, enhancing its ability to revise previous responses. We then evaluated the fidelity of uncertainty transfer in summaries using a straightforward precision and recall method, offering clear insights into how well the summaries align with the articles in terms of uncertainty expressions.

For future research, one avenue to explore can be additional dimensions of uncertainty in alignment with how human beings identify and solve uncertainty. We believe a multi-modal approach, integrating diverse linguistic cues beyond textual information, could significantly enhance the overall understanding of uncertainty. This approach might provide additional benefits to the study of uncertainty.

Another avenue of research could be exploring the practical application of enhanced uncertainty understanding in decision-making tools reliant on the summarization of lengthy documents across various sectors, including healthcare, finance, or risk assessment domains, offering insights into the level and nature of uncertainty within data or information sources.

Limitations

Primarily, our evaluation focused solely on uncertainty as a measure of summarization quality, neglecting other essential facets that might impact the assessment, thereby confining the scope of our study. Additionally, post-hoc evaluation process can get costly if more rounds of self-correction are required. Finally, in this study, we only focused on lexical and semantic expressions of uncertainty expressions at the word or phrase level and did not consider e.g., discourse-level expressions of uncertainty.

Ethics Statement

Web-Based Content for Research Purposes

Initially, we ensured that the content we gathered from web sources was obtained from websites explicitly permitting web scraping. The collected content was exclusively utilized for the sole purpose of this research, focusing on identifying textual uncertainty and creating summaries, as highlighted in other sections of this study. Given the diverse range of content collected from the internet—comprising personal blogs, news articles, opinion pieces, among others—it is possible that certain content might contain biased opinions or lack factual accuracy. Therefore, we urge the NLP community to utilize this dataset for its intended purpose, specifically for uncertainty annotation and evaluation, while being mindful of potential biases or inaccuracies inherent in the collected content.

Experiment Involving Human Participants

To conduct human evaluations, we recruited two linguists, who were recruited voluntarily and had

the option to withdraw at any time. Compensation rates followed the community norms for their involvement and effort. Participants were informed beforehand that any content conflicting with their values or indicating bias did not reflect the authors' opinions. We provided a feedback section for participants to flag such articles, ensuring removal from the experiment and final results for the research community. However, no feedback or comments regarding such content were received.

Acknowledgements

We extend our heartfelt gratitude to the linguists whose invaluable contributions were instrumental in the completion of this research.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

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.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Alain Auger and Jean Roy. 2008. Expression of uncertainty in linguistic data. In *2008 11th International Conference on Information Fusion*, pages 1–8. IEEE.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Richard Bradley and Mareile Drechsler. 2014. Types of uncertainty. *Erkenntnis*, 79:1225–1248.

Wendy Chapman, John Dowling, and David Chu. 2007. Context: An algorithm for identifying contextual features from clinical text. In *Biological, translational, and clinical language processing*, pages 81–88.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators.

Wenshi Chen, Bowen Zhang, and Mingyu Lu. 2020. Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1384.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2023. Relic: Investigating large language model responses using self-consistency. *arXiv preprint arXiv:2311.16842*.

Noa P Cruz Díaz. 2013. Detecting negated and uncertain information in biomedical and review texts. In *Proceedings of the Student Research Workshop associated with RANLP 2013*, pages 45–50.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task*, pages 1–12.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural language generation enhances human decision-making with uncertain information. *arXiv preprint arXiv:1606.03254*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372.

Eduard Hofer, Martina Kloos, Bernard Krzykacz-Hausmann, Jörg Peschke, and Martin Woltereck. 2002. An approximate epistemic uncertainty analysis

approach in the presence of epistemic and aleatory uncertainties. *Reliability Engineering & System Safety*, 77(3):229–238.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Yibo Hu and Latifur Khan. 2021. Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 628–636.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

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.

Marie Juanchich, Amélie Gourdon-Kanhukamwe, and Miroslav Sirota. 2017. "i am uncertain" vs "it is uncertain". how linguistic markers of the uncertainty source affect uncertainty communication. *Judgment and Decision Making*, 12(5):445–465.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9:1–25.

Zahra Kolagar, Sebastian Steindl, and Alessandra Zarcone. 2023. Eduquick: A dataset toward evaluating summarization of informal educational content for social media. In *Eval4NLP*, pages 32–48.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Lrec*, pages 3190–3195.

Anis Koubaa. 2023. Gpt-4 vs. gpt-3.5: A concise showdown.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2023. Deup: Direct epistemic uncertainty prediction.

Minghan Li, Ming Li, Kun Xiong, and Jimmy Lin. 2021. Multi-task dense retrieval via model uncertainty fusion for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 274–287.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 workshop: linking biological literature, ontologies and databases*, pages 17–24.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

R Duncan Luce and Howard Raiffa. 1989. *Games and decisions: Introduction and critical survey*. Courier Corporation.

Zhihao Lyu, Danier Duolikun, Bowei Dai, Yuan Yao, Pasquale Minervini, Tim Z Xiao, and Yarin Gal. 2020. You need only uncertain answers: Data efficient multilingual question answering. *TWorkshop on Uncertainty and Ro-Bustness in Deep Learning*.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 992–999.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

Roser Morante, Vincent Van Asch, and Antal van den Bosch. 2009. Joint memory-based learning of syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 25–30.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33:21199–21212.

Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 69–77.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

OpenAI. 2022. OpenAI: Introducing ChatGPT. [Online; posted 30-November-2022].

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Dorian Quelle and Alexandre Bovet. 2023. The perils & promises of fact-checking with large language models. *arXiv preprint arXiv:2310.13549*.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2023. Measuring reliability of large language models through semantic consistency.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Verena Rieser and Oliver Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Conference of the European Association for Computational Linguistics*, pages 105–120. Springer.

Victoria L Rubin. 2006. *Identifying certainty in texts*. Ph.D. thesis, Syracuse University, NY.

Victoria L Rubin. 2010. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540.

Victoria L Rubin, Elizabeth D Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. *Computing attitude and affect in text: Theory and applications*, pages 61–76.

Shankar Sankararaman and Sankaran Mahadevan. 2011. Model validation under epistemic uncertainty. *Reliability Engineering & System Safety*, 96(9):1232–1241.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Big Science. 2023. Introducing the world's largest open multilingual language model: Bloom.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Merve Ünlü and Ebru Arisoy. 2021. Uncertainty-aware representations for spoken question answering. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 943–949. IEEE.

Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association*, 16(1):109–115.

Linda C van der Gaag, Silja Renooij, Cilia LM Witteman, Berthe MP Aleman, and Babs G Taal. 2013. How to elicit many probabilities. *arXiv preprint arXiv:1301.6745*.

Veronika Vincze. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles.

Veronika Vincze. 2014a. Uncertainty detection in hungarian texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1844–1853.

Veronika Vincze. 2014b. Uncertainty detection in natural language texts. *PhD, University of Szeged*, 141.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Peter Walley and Gert De Cooman. 2001. A behavioral model for linguistic uncertainty. *Information Sciences*, 134(1-4):1–37.

Hai Wang, Zeshui Xu, and Xiao-Jun Zeng. 2018. Linguistic terms with weakened hedges: A model for qualitative decision making under uncertainty. *Information Sciences*, 433:37–54.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2018. An empirical study on uncertainty identification in social media context. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 79–88. World Scientific.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. *arXiv preprint arXiv:2010.07882*.

Ryutaro Yamauchi, Sho Sonoda, Akiyoshi Sannai, and Wataru Kumagai. 2023. Lpml: Llm-prompting markup language for mathematical reasoning.

Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. *arXiv preprint arXiv:2304.08653*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. 2023a. Sac^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023c. Benchmarking large language models for news summarization.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. *arXiv preprint arXiv:1907.07590*.

A Appendix

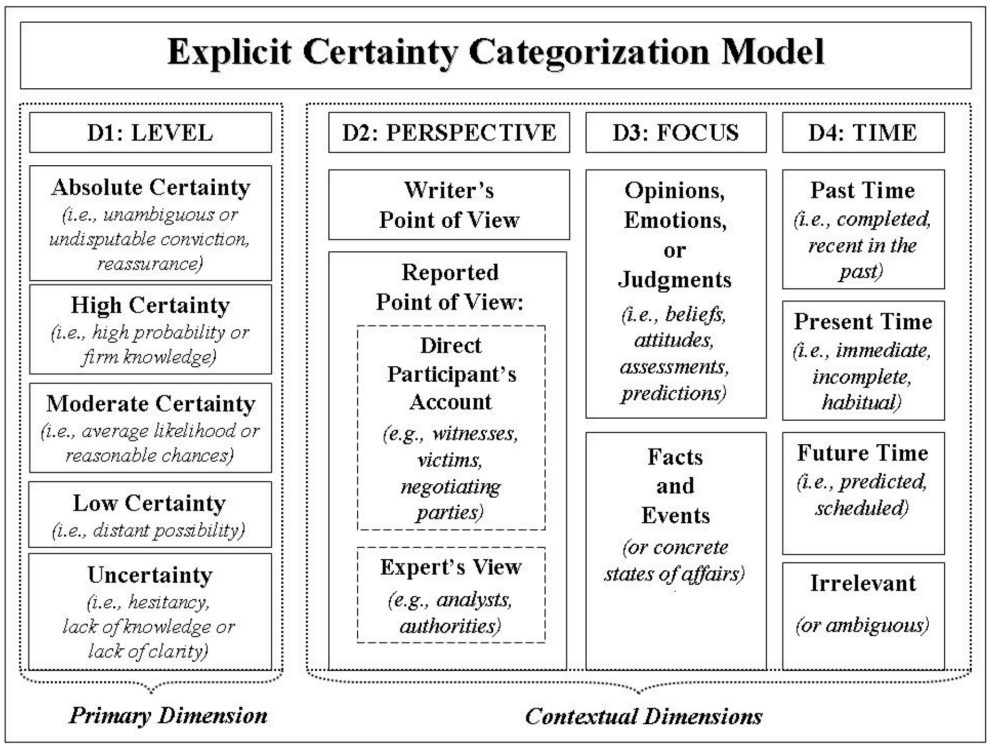

Figure 4: Explicit certainty categorization model introduced by (Rubin, 2006)

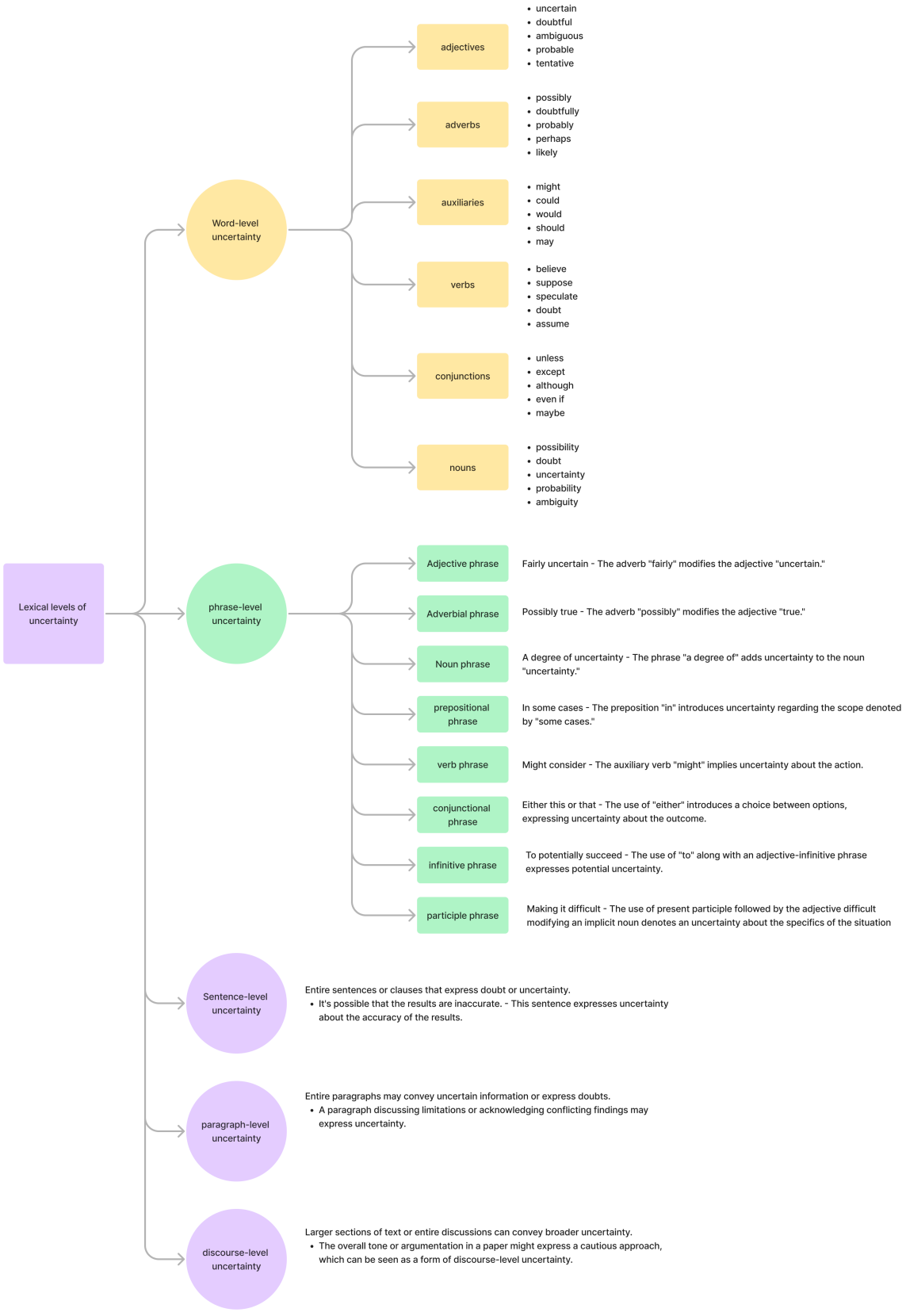

Figure 5: Lexical taxonomy of the linguistic expressions of uncertainty along with examples for each category

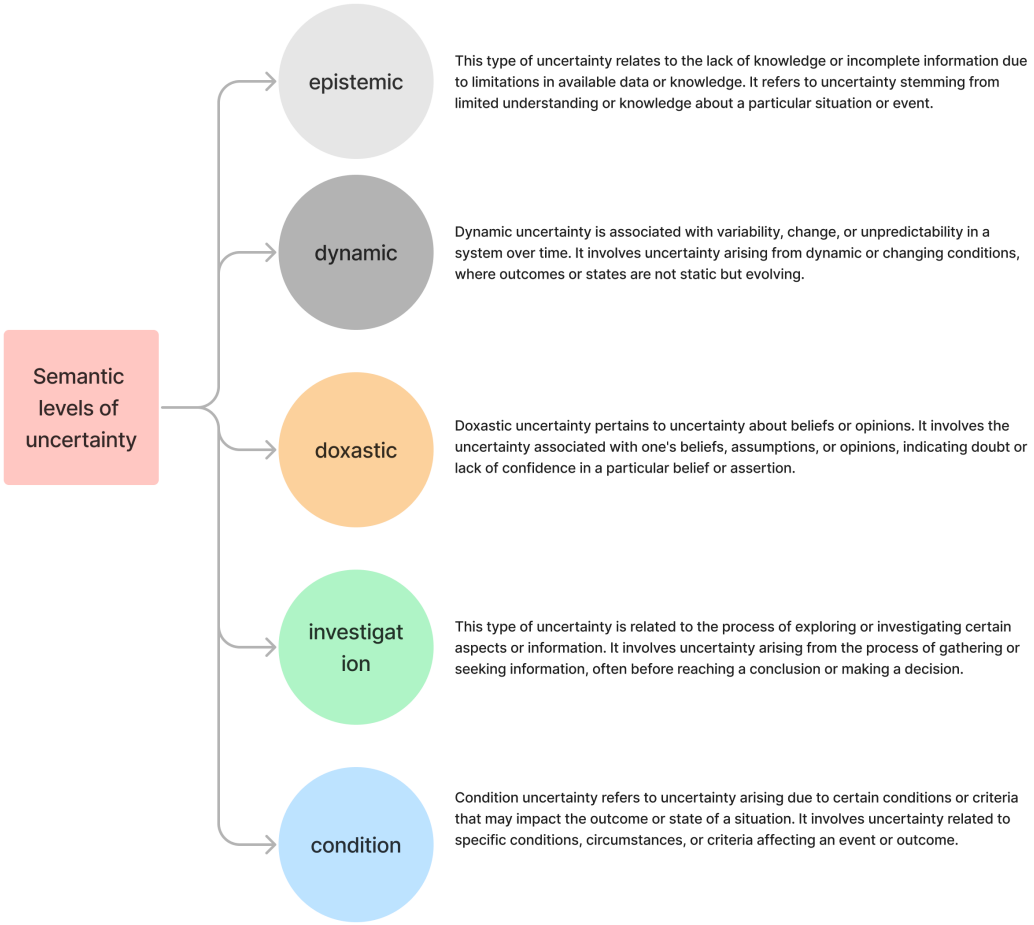

Figure 6: Lexical taxonomy of the linguistic expressions of uncertainty along with examples for each category

Instruction - Read the Context, the Input example, and the output example carefully. Apply the analysis to the Input Text, following the structure demonstrated in the Output example.

Context - Presented below is a lexical and semantic taxonomy of uncertainty in text, including examples for each category.

Lexical uncertainty:

1. Word-level uncertainty examples

    a. Adjective: uncertain, doubtful
    b. Adverb: possibly, likely
    c. Auxiliaries: might, could
    d. Verbs: assume, suppose
    e. Conjunctions: unless, except
    f. Nouns: possibility, ambiguity

1. Phrase-level uncertainty examples

    a. Adjective phrase: fairly uncertain
    b. Adverbial phrase: possibly true
    c. Noun phrase: a degree of uncertainty
    d. Prepositional phrase: in some cases
    e. Verb phrase: might consider
    f. Conjunctional phrase: either this or that
    g. Infinitive phrase: to potentially succeed
    h. Participle phrase: making it difficult

Semantic uncertainty:

1. Epistemic: It is related to the lack of knowledge or incomplete information due to limitations in available data or knowledge. It refers to uncertainty stemming from limited understanding or knowledge about a particular situation or event.
2. Dynamic: It is associated with variability, change, or unpredictability in a system over time. It involves uncertainty arising from dynamic or changing conditions, where outcomes or states are not static but evolving.
3. Doxastic: It is related to uncertainty about beliefs or opinions. It involves the uncertainty associated with one's beliefs, assumptions, or opinions, indicating doubt or lack of confidence in a particular belief or assertion.
4. Investigation: It is related to the process of exploring or investigating certain aspects or information. It involves uncertainty arising from the process of gathering or seeking information, often before reaching a conclusion or making a decision.
5. Condition: It refers to uncertainty arising due to certain conditions or criteria that may impact the outcome or state of a situation. It involves uncertainty related to specific conditions, circumstances, or criteria affecting an event or outcome.

Input example:

"Maintaining a strict diet plan can sometimes feel like an overwhelming task, especially when faced with various contradictory information about what is healthy. It's hard to determine which sources to trust, as some claim certain foods are beneficial, while others refute those claims entirely. Additionally, personal experiences and changing dietary trends may also influence one's understanding of what healthy diet is."

Output example:

"Maintaining a strict diet plan can sometimes feel like an <Uncertainty POS="adjective phrase" semantic="epistemic">overwhelming task</Uncertainty>, especially when faced with various <Uncertainty POS="adjective phrase" semantic="epistemic">contradictory information</Uncertainty> about what's deemed healthy. It's hard to determine which sources to trust, as some <Uncertainty POS="verb" semantic="epistemic">claim</Uncertainty> certain foods are beneficial, while others <Uncertainty POS="verb" semantic="epistemic">refute</Uncertainty> those claims entirely. Additionally, personal experiences and <Uncertainty POS="adjective" semantic="dynamic">changing</Uncertainty> dietary trends <Uncertainty POS="auxiliary" semantic="epistemic">may</Uncertainty> also influence one's understanding of what healthy diet is."

Input text: {text}

Figure 7: The prompt presented to GPT4-8k model to perform linguistic uncertainty annotation

Dear Linguists,

Thank you for participating in our experiment. In this study, we are presenting you with 15 texts, each containing annotations focused on linguistic uncertainty representations in text. These annotations are applied at both word and phrase levels (Lexical level) within each text, encompassing Part-of-Speech (POS) and semantic aspects as described below.

Lexical uncertainty:

1. Word-level uncertainty examples

    a. Adjective: uncertain, doubtful
    b. Adverb: possibly, likely
    c. Auxiliaries: might, could
    d. Verbs: assume, suppose
    e. Conjunctions: unless, except
    f. Nouns: possibility, ambiguity

1. Phrase-level uncertainty examples

    a. Adjective phrase: fairly uncertain
    b. Adverbial phrase: possibly true
    c. Noun phrase: a degree of uncertainty
    d. Prepositional phrase: in some cases
    e. Verb phrase: might consider
    f. Conjunctional phrase: either this or that
    g. Infinitive phrase: to potentially succeed
    h. Participle phrase: making it difficult

Semantic uncertainty:

1. Epistemic: It is related to the lack of knowledge or incomplete information due to limitations in available data or knowledge. It refers to uncertainty stemming from limited understanding or knowledge about a particular situation or event.
2. Dynamic: It is associated with variability, change, or unpredictability in a system over time. It involves uncertainty arising from dynamic or changing conditions, where outcomes or states are not static but evolving.
3. Doxastic: It is related to uncertainty about beliefs or opinions. It involves the uncertainty associated with one's beliefs, assumptions, or opinions, indicating doubt or lack of confidence in a particular belief or assertion.
4. Investigation: It is related to the process of exploring or investigating certain aspects or information. It involves uncertainty arising from the process of gathering or seeking information, often before reaching a conclusion or making a decision.
5. Condition: It refers to uncertainty arising due to certain conditions or criteria that may impact the outcome or state of a situation. It involves uncertainty related to specific conditions, circumstances, or criteria affecting an event or outcome.

Upon reviewing the provided examples and explanations, kindly assess the following:

1. Consistency and Accuracy Assessment: Evaluate the annotations for their consistency and accuracy using specific criteria. Please pay particular attention to: a. The accuracy of the selected text spans. b. The correct labeling of Part-of-Speech (POS) elements. c. Ensuring accurate semantic labels that depict uncertainty expressions.
2. Annotation Categorization: Categorize the extracted examples as either correct or incorrect. For any identified incorrect annotations, please create new annotations, providing the correct labels under the designated "evaluation" element within the <Uncertainty> tag.

    As you <Uncertainty POS="auxiliaries" semantic="dynamic" evaluation="incorrect">might</Uncertainty> expect, the story shortened over time as participants forgot certain details.

3. Explanation for Incorrect Annotations: Offer clear and detailed explanations for any identified incorrect annotations encountered. Document these explanations along with the corrected annotations for reference purposes.

Thank you for your valuable contributions to this experiment.

Figure 8: The instruction presented to the linguists for the correction of the GPT-4-8K annotated texts.

"Years before he arrived, the district's finance team made a mistake that led to some <Uncertainty POS="noun" semantic="dynamic" evaluation="incorrect">misspending</Uncertainty> of a small amount of state grant dollars.

Expert Reasoning: Misspending is not an example of uncertainty. A good example of dynamic uncertainty noun could be 'changing', 'vary', 'instability', 'unpredictability' and so on.

The story illustrates the long tail district leaders face as auditors <Uncertainty POS="verb" semantic="investigation" evaluation="incorrect">swoop in</Uncertainty> to examine how they spent billions in federal COVID relief aid that's poured into K-12 schools since the COVID-19 pandemic began.

Expert Reasoning: swoop in is not an example of investigation uncertainty verb. Instead, you should have annotated the verb examine in the above example.
<Uncertainty POS="verb" semantic="investigation" >examine</Uncertainty>
Other examples of investigation uncertainty verb are explore, research, investigate, examine, study, test, inquire, discover, uncover.

But spending federal dollars prudently <Uncertainty POS="adverb" semantic="epistemic" evaluation="correct">possibly</Uncertainty> requires complex rules, tight schedules, and dense bureaucracies.

Expert Reasoning: Possibly is an epistemic adverb, so the annotation is correct. Other examples of epistemic adverb can be 'likely', 'doubtfully', assumedly'.

Figure 9: An illustration of sample corrections implemented by the linguists. The color coding has been included solely to enhance visual clarity.

Context: Below is the linguistic taxonomy detailing uncertainty expressions in text, followed by analyses conducted by linguists. The 'evaluation' tag denotes whether annotations were “correct” or “incorrect”, accompanied by expert reasoning, and the text you have annotated before.

Lexical uncertainty:

Word-level uncertainty examples

Adjective: uncertain, doubtful
Adverb: possibly, likely
Auxiliaries: might, could
Verbs: assume, suppose
Conjunctions: unless, except
Nouns: possibility, ambiguity

Phrase-level uncertainty examples

Adjective phrase: fairly uncertain
Adverbial phrase: possibly true
Noun phrase: a degree of uncertainty
Prepositional phrase: in some cases
Verb phrase: might consider
Conjunctional phrase: either this or that
Infinitive phrase: to potentially succeed
Participle phrase: making it difficult

Semantic uncertainty:

Epistemic: It is related to the lack of knowledge or incomplete information due to limitations in available data or knowledge. It refers to uncertainty stemming from limited understanding or knowledge about a particular situation or event.
Dynamic: It is associated with variability, change, or unpredictability in a system over time. It involves uncertainty arising from dynamic or changing conditions, where outcomes or states are not static but evolving.
Doxastic: It is related to uncertainty about beliefs or opinions. It involves the uncertainty associated with one's beliefs, assumptions, or opinions, indicating doubt or lack of confidence in a particular belief or assertion.
Investigation: It is related to the process of exploring or investigating certain aspects or information. It involves uncertainty arising from the process of gathering or seeking information, often before reaching a conclusion or making a decision.
Condition: It refers to uncertainty arising due to certain conditions or criteria that may impact the outcome or state of a situation. It involves uncertainty related to specific conditions, circumstances, or criteria affecting an event or outcome.

Expert corrections and reasoning:
{ Expert corrections and reasoning will appear here}

Previously annotated text: { annotated_text}

Review your previous annotation, and identify issues with the annotations.
{ Identified issues will appear here}

Based on the problem you have found, improve your annotation of uncertainty expressions.
{ Refined annotations will appear here}

Figure 10: The post-hoc prompt presented to the GPT-4 model to reason about and correct previous annotations for the 15 selected samples.

Context: Below is the linguistic taxonomy detailing uncertainty expressions in text, followed by analyses conducted by linguists. The 'evaluation' tag denotes whether annotations were “correct” or “incorrect”, accompanied by expert reasoning, and the text you have annotated before.

Lexical uncertainty:

1. Word-level uncertainty examples

    a. Adjective: uncertain, doubtful
    b. Adverb: possibly, likely
    c. Auxiliaries: might, could
    d. Verbs: assume, suppose
    e. Conjunctions: unless, except
    f. Nouns: possibility, ambiguity

1. Phrase-level uncertainty examples

    a. Adjective phrase: fairly uncertain
    b. Adverbial phrase: possibly true
    c. Noun phrase: a degree of uncertainty
    d. Prepositional phrase: in some cases
    e. Verb phrase: might consider
    f. Conjunctional phrase: either this or that
    g. Infinitive phrase: to potentially succeed
    h. Participle phrase: making it difficult

Semantic uncertainty:

1. Epistemic: It is related to the lack of knowledge or incomplete information due to limitations in available data or knowledge. It refers to uncertainty stemming from limited understanding or knowledge about a particular situation or event.
2. Dynamic: It is associated with variability, change, or unpredictability in a system over time. It involves uncertainty arising from dynamic or changing conditions, where outcomes or states are not static but evolving.
3. Doxastic: It is related to uncertainty about beliefs or opinions. It involves the uncertainty associated with one's beliefs, assumptions, or opinions, indicating doubt or lack of confidence in a particular belief or assertion.
4. Investigation: It is related to the process of exploring or investigating certain aspects or information. It involves uncertainty arising from the process of gathering or seeking information, often before reaching a conclusion or making a decision.
5. Condition: It refers to uncertainty arising due to certain conditions or criteria that may impact the outcome or state of a situation. It involves uncertainty related to specific conditions, circumstances, or criteria affecting an event or outcome.

Expert corrections and reasoning:

Text: A recent article in the New York Times shared research on longevity that revealed that the people who live the longest not only live healthy lifestyles, but also tend to engage and connect with the people around them.
Expert Reasoning: ‘tend’ indicates a tendency or likelihood in this context rather than a scientific fact.
Annotation for tend to: <Uncertainty POS="verb" semantic="epistemic">tend</Uncertainty>
Other examples of epistemic verb could be, seem or appear.

Text: You’re left with just the basics and <Uncertainty POS="conjunctional phrase" semantic="condition" evaluation="correct">whether or not</Uncertainty> you have mastered them.
Expert Reasoning: The annotation is correct. Other examples of conditional phrase could be ‘either this or that’.

Text: I <Uncertainty POS="verb" semantic="doxastic" evaluation="correct">believe</Uncertainty> that this is also one of the reasons why history repeats itself.
Expert Reasoning: The annotation is correct. Believe is an example pf doxastic uncertainty based on one’s opinion. Other examples might include: consider, think, hypothesize, expect and assume.

Text: Because of the capabilities of artificial intelligence, schools will need to evaluate student learning, according to the panelists
Expert Reasoning: evaluate is an example of uncertainty type known as investigation, which is missing from the text. <Uncertainty POS="verb" semantic="investigation" >evaluate</Uncertainty>
Other examples of investigation uncertainty are, determine, investigate, examine, test, study.

Text: Cardona told Education Week that the move is underway, “we might change the department of education to be inclusive.”
Expert Reasoning: might change is an example of dynamic uncertainty and it is a verb phrase. Other examples of dynamic uncertainty can be fluctuate, evolve, change, or unpredictability as a noun.
<Uncertainty POS="verb phrase" semantic="dynamic" >might change</Uncertainty>

Previously annotated text: {annotated_text}

Review your previous annotation, and identify issues with the annotations.
{Identified issues will appear here}

Based on the problem you have found, improve your annotation of uncertainty expressions.
{Refined annotations will appear here}

Figure 11: The post-hoc prompt presented to the GPT-4 model to correct the rest of the dataset.

Case 1: **No annotation was found in the matched section of the article, resulting in a score of 0.**

Article: In fact, fewer than 1 in 6 educators—13 percent—surveyed by the EdWeek Research Center earlier this year say that A through F or numeric grades are a not “very effective way” to give feedback to students or evaluate their progress.

Summary: Some educators are <Uncertainty POS="adjective phrase" semantic="epistemic">somewhat uncertain</Uncertainty> that the scoring system captures student progress consistently.

Case2: **One annotation was found in the matched section of the article containing the same semantic label, resulting in a score of 1.**

Article: Better demographic data about young children with disabilities who need and receive federally funded early intervention services, such as physical therapy, <Uncertainty POS="verb" semantic="epistemic">could</Uncertainty> help policymakers address barriers to access.

Summary: Better data about young children with disabilities <Uncertainty POS="verb" semantic="epistemic">could</Uncertainty> help address barriers.

Case3: **One annotation was found in the matched section of the article containing a different semantic label, resulting in a score of 0.**

Article: In cases like these, when we are attempting to do something that is complex and multi-faceted, I <Uncertainty POS="verb" semantic="doxastic">believe</Uncertainty> that being wrong is actually a sign that you’re doing something right.

Summary: The text suggests that being wrong <Uncertainty POS="verb phrase" semantic="epistemic">might be</Uncertainty> part of the process of making complex decisions.

Case4: **Multiple annotations were found in the matched section of the article containing the same label, resulting in a score of 1.**

Article: In the early phases of any activity like going to the gym or starting a new diet, it's <Uncertainty POS="adjective" semantic="epistemic"> probable</Uncertainty> that some errors <Uncertainty POS="auxiliary" semantic="epistemic"> might</Uncertainty> occur that results in getting negative feedback.

Summary: The initial stages of any endeavour are <Uncertainty POS="adverb" semantic="epistemic">likely</Uncertainty> to be filled with mistakes.

Case5: **Multiple annotations were found in the matched section of the article containing different labels, resulting in a score of 0.**

Article: I <Uncertainty POS="adverb" semantic="doxastic">believe</Uncertainty> that consistency is <Uncertainty POS="adverb" semantic="epistemic">probably</Uncertainty> very important for making progress, doing better work, getting in shape, and achieving some level of success in different areas of life.

Summary: The author suggests that <Uncertainty POS="conjunction" semantic="condition">if</Uncertainty> you are consistent, you see progress.

Figure 12: Example of matched sections of the articles and the summary for the cases explained in Section 5.1.