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

Zahra Kolagar, Alessandra Zarcone

February 17, 2024

Fraunhofer Institute for
Integrated Circuits IIS

Motivation

Analyzing how faithfully the linguistic uncertainty from the source text is conveyed to the summaries.

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.

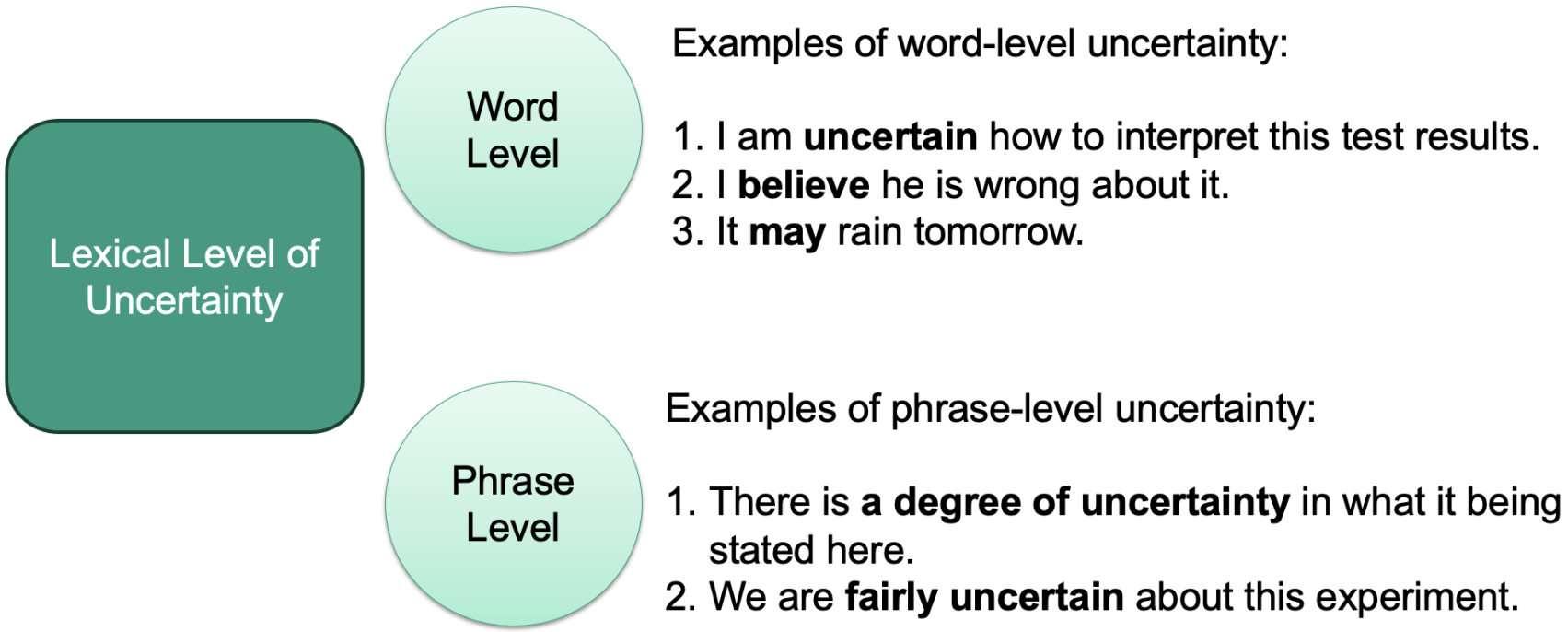
Research Questions

- How faithful are LLM-generated summaries, and how do the uncertainty expressions in the summary align with the corresponding expressions in the source texts?

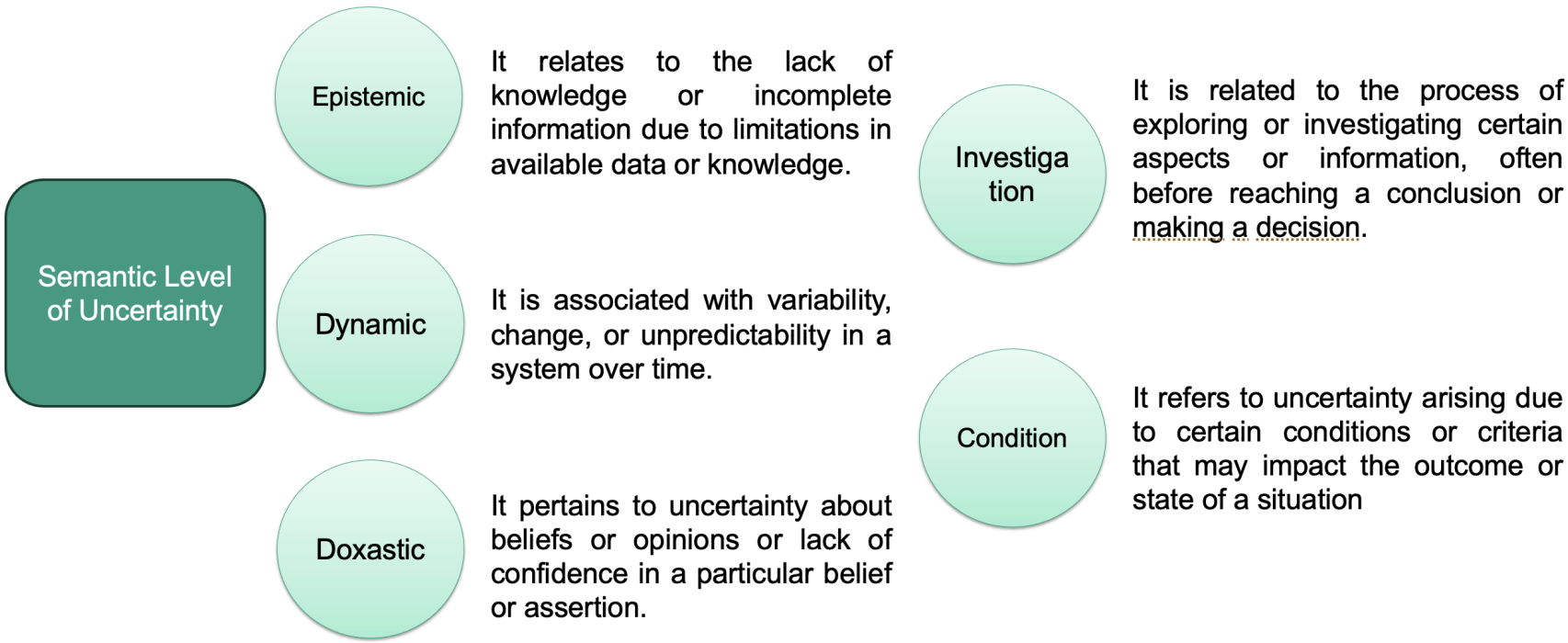
- How can LLMs be employed to identify and annotate expressions of uncertainty in text?

Linguistic Uncertainty

Example of Lexical Uncertainty Expressions



Semantic Uncertainty Expressions



Semantic Uncertainty Expressions

Examples of Semantic Uncertainty

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.

Data Acquisition and Annotation

Data Collection & Summarization

- "Education Week" (https://www.edweek.org/), an educational website featuring various articles on educational topics.
- "An Easy Proven Way to Build Good Habits Break Bad Ones" (https://jamesclear.com/) website, a personal blog.
- **150** articles of **600-700 words** to control the variation in text length and to facilitate more consistent and informed human evaluations.
- Instructed GPT-4 to generate summaries within a maximum limit of **200 words**

Uncertainty Annotation Leveraging LLMs

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

Markup-based annotation:

- allows for **precise** and **fine-grained** annotation
- ensures **consistency** and **standardization** in annotation practices across different datasets and annotators
- ensures **compatibility** of markup annotations with various text processing tools

Uncertainty Annotation Evaluation & Refinement

Expert Evaluation

15 out of 150 samples were reviewed by two linguists.

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

Expert Reasoning: 'might expect' should have been annotated as a verb phrase including an auxiliary + verb. Also, 'might expect' is an example of doxastic uncertainty. The correct annotation is therefore, <Uncertainty POS="verb phrase" semantic="doxastic" >might expect </Uncertainty>

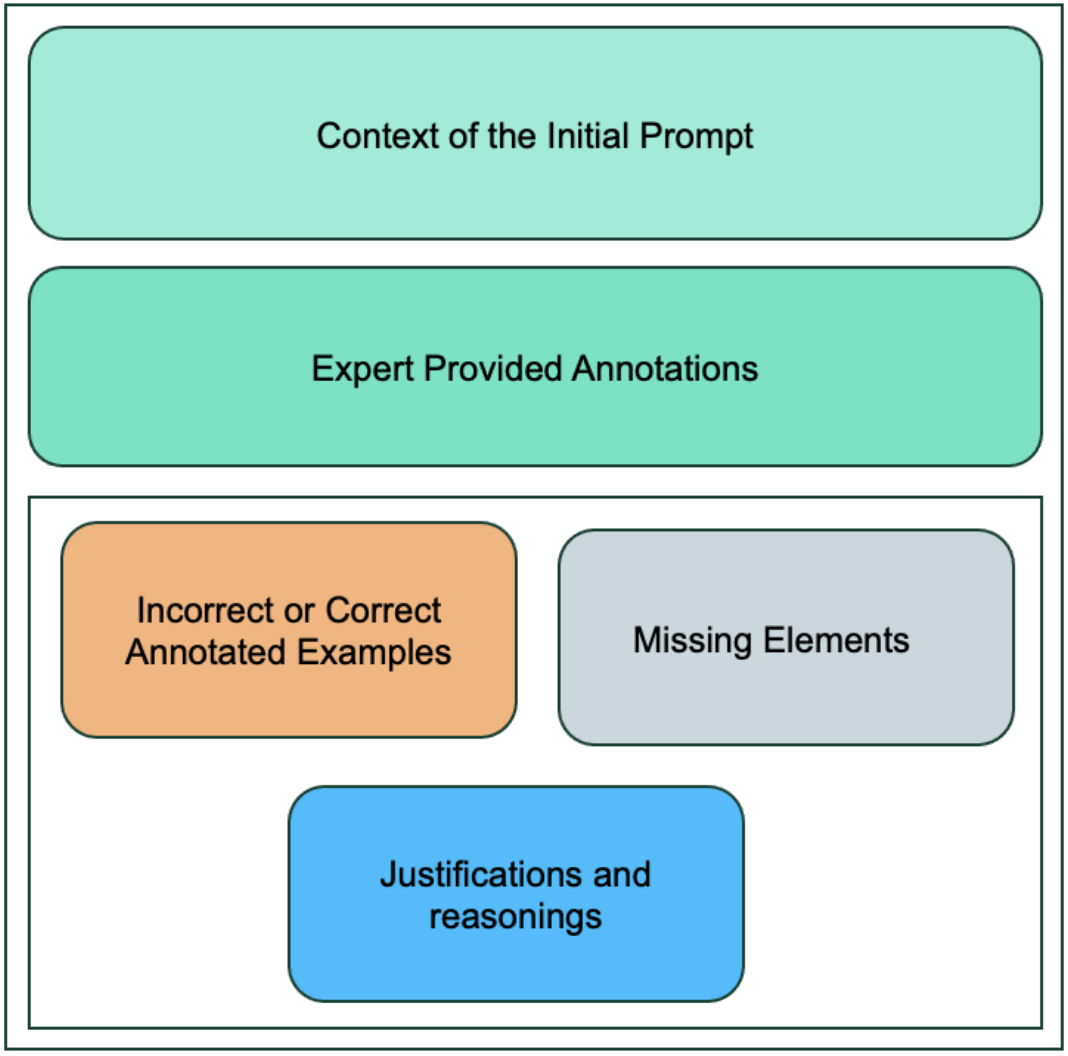
Expert Evaluation

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).

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

Expert Guided Self-Refinement Using Post-hoc Prompting

Post-hoc Prompt Elements:



Expert Guided Self-Refinement Using Post-hoc Prompting

GPT-4 annotation accuracy at different stages

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

- Extended the refinement process to the **135** remaining samples
- Incorporated **excerpts** from the refined 15 expert annotations as examples into the prompt to guide the model
- Randomly selected **2 articles** for expert assessment
- Observed a decrease in the model's refinement accuracy to **80.4%** (76.8% for semantic attribute)

Analyzing Uncertainty Transfer in Summarization

Evaluation of Uncertainty Representation in Summaries

- Only analyzed the **semantic** annotation of uncertainty based on the 5 semantic labels namely, condition, investigation, epistemic, dynamic, and doxastic
- Excluded the analysis of **POS** in this evaluation as POS alterations might occur in summarization without necessarily affecting the fidelity of uncertainty expressions
- Excluded a **comprehensive evaluation** of other summary quality aspects

Evaluation of Uncertainty Representation in Summaries

We need to align sentences or clauses containing uncertainty annotation in the summary to the corresponding sections in the article

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.

Exact Alignment

We computed **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.

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

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

Result

Semantic Type	Precision	Recall
Epistemic	0.68	0.50
Dynamic	0.56	0.32
Doxastic	0.68	0.50
Investigation	0.81	0.59
Condition	0.34	0.33
Total	**0.67**	**0.49**

Discussion

- We did not account for the **ranking or significance** of uncertainty expressions
- The automatic annotation yielded a **lower accuracy** on the 135 sample articles, potentially influencing the precision and recall outcomes
- Variations in precision outcomes seem to also arise from the **differing number of semantic types**
- The lower recall is acceptable, considering that the frequency of uncertainty expressions are much less in the summaries

Conclusion

- We introduced a **two-tier annotation taxonomy** that categorizes linguistic uncertainty expressions within the text
- Developed an **XML-based syntax** framework to standardize the annotation process for these expressions
- We conducted experiments involving **expert linguists** to refine annotations
- Utilized their expert rationale to guide the LLM's self-evaluation, using **post-hoc prompting technique**
- Evaluated the **fidelity of uncertainty transfer** in summaries using a straightforward precision and recall method