Explaining Sources of Uncertainty in Automated Fact-Checking

Anonymous ACL submission

Abstract

Understanding sources of a model's uncertainty regarding its predictions is crucial for effective human-AI collaboration. Prior work proposes to use numerical uncertainty or hedges ("I'm not sure, but. . . "), which do not explain uncertainty arising from conflicting evidence, leaving users unable to resolve disagreements or rely on the output. We introduce CLUE (**C**onflict-&Agreement-aware **L**anguage-model **U**ncertainty **E**xplanations), the first framework to generate natural language explanations of model uncertainty by: (i) identifying relationships between spans of text that expose claim-evidence or inter-evidence conflicts/agreements driving the model's predictive uncertainty in an unsupervised way; and (ii) generating explanations via prompting and attention steering to verbalize these critical interactions. Across three language models and two fact-checking datasets, we demonstrate that CLUE generates explanations that are more faithful to model uncertainty and more consistent with fact-checking decisions than prompting for explanation of uncertainty without span-interaction guidance. Human evaluators find our explanations more helpful, more informative, less redundant, and better logically aligned with the input than this prompting baseline. CLUE requires no fine-tuning or architectural changes, making it plug-and-play for any white-box language model. By explicitly linking uncertainty to evidence conflicts, it offers practical support for fact-checking and readily generalizes to other tasks that require reasoning over complex information.

1 Introduction

Large Language Models (LLMs) are increasingly prevalent in high-stakes tasks that involve reasoning about information reliability, such as fact-checking (Wang et al., 2024; Fontana et al., 2025). To foster effective use of such models in fact-checking tasks, these models must explain the rationale for their predictions (Atanasova et al., 2020; Kotonya and Toni, 2020).

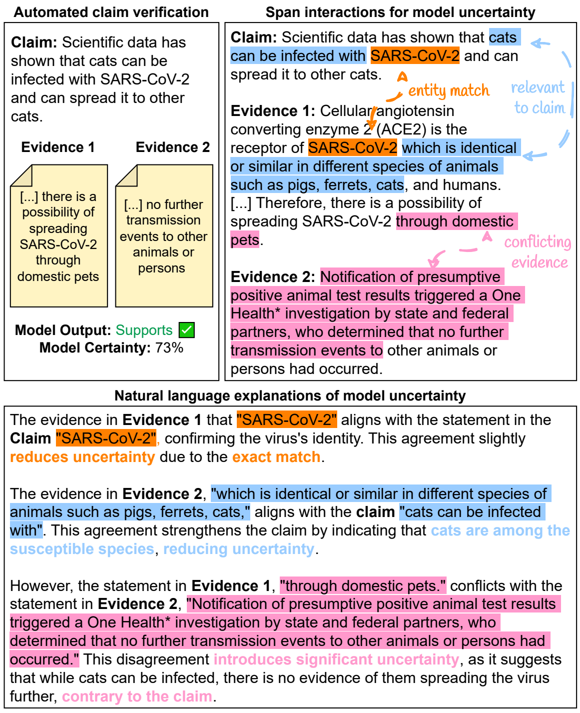

Figure 1: Example of claim and evidence documents, alongside span interactions for uncertainty and generated natural language explanations.

However, current methods in automated fact-checking have been criticised for their failure to address practical explainability needs of fact-checkers (Warren et al., 2025) and for their disconnect from the tasks typically performed by fact-checkers (Schlichtkrull et al., 2023). For example, although fact-checking involves complex reasoning about the reliability of evidence, which may be conflicting, existing automatic fact-checking techniques focus only on justifying the verdict (Atanasova et al., 2020; Stammbach and Ash, 2020; Zeng and Gao, 2024). Such methods do not explain the uncertainty associated with their predictions, which is crucial for their users to determine whether some of the uncertainty is resolvable, and if so, which aspects of this uncertainty within the evidence to address (e.g., by retrieving additional information) (Warren et al., 2025).

Uncertainty in model predictions is often communicated through numerical scores (e.g., "I am 73% confident"), however, metrics can be hard to contextualize and lack actionable insights for end-users (Zimmer, 1983; Wallsten et al., 1993; van der Waa et al., 2020; Liu et al., 2020). Recent efforts have instead used natural language expressions (e.g., "I'm not sure") to convey uncertainty (Steyvers et al., 2025; Yona et al., 2024; Kim et al., 2024), but these approaches have limitations: users may overestimate model confidence (Steyvers et al., 2025) and such expressions often fail to faithfully reflect model uncertainty (Yona et al., 2024). Existing explainable fact-checking systems exhibit two critical limitations: they focus solely on justifying veracity predictions through generic reasoning summaries of the input sequence (see Figure 2), while failing to (1) communicate model uncertainty or (2) explicitly surface evidentiary conflicts and agreements that relate to it. This constitutes a fundamental methodological gap, as effective fact-checking requires precisely identifying the sources of uncertainty, for example from conflicting evidence, to guide targeted verification.

We propose CLUE, a pipeline that generates natural language explanations (NLEs) of model uncertainty by explicitly capturing conflicts and agreements in the input (e.g., a claim and its supporting or refuting evidence). The pipeline first identifies the salient span-level interactions that matter to the prediction of the model through an unsupervised approach, providing an input-feature explanation that highlights key relationships between separate input segments (e.g., claim and evidence) (Ray Choudhury et al., 2023). These interactions have been shown to be both faithful to the model and plausible to humans (Sun et al., 2025). CLUE then converts these signals into uncertainty-aware explanations by explicitly discussing the interactions and the conflict/agreement relations they express. CLUE does not require gold-label explanations, avoids fine-tuning, and operates entirely at inference time.

Across three language models (§4.2) and two fact-checking datasets (§4.1), we evaluate two variants of CLUE. Automatic metrics show that both variants generate explanations that are more faithful to each model's uncertainty and agree more closely with the gold fact-checking labels than a prompting baseline that lacks conflict-/agreement-span guidance(§5.5). Human judgements likewise rate the CLUE explanations as more helpful, more informative, less redundant, and better logically aligned with the input. We also observe a trade-off between two variants of our CLUE framework, one attains higher faithfulness, the other higher plausibility, highlighting a promising avenue for future work to achieve both simultaneously(§5.5).

2 Related Work

2.1 Uncertainty Quantification in LLMs

Recent work on LLM uncertainty quantification primarily relies on logit-based methods such as answer distribution entropy (Kadavath et al., 2022), summing predictive entropies across generations (Malinin and Gales, 2021), and applying predictive entropy to multi-answer question-answering (Yang et al., 2025), while estimating uncertainty in long-form tasks involves measuring semantic similarity between responses (Duan et al., 2024; Kuhn et al., 2023; Nikitin et al., 2024). Quantifying uncertainty in black-box models often relies on verbalizing confidence directly (Lin et al., 2022; Mielke et al., 2022b), though these measures are overconfident and unreliable (Yona et al., 2024; Tanneru et al., 2024). Alternative approaches measure output diversity across paraphrased prompts (Zhang et al., 2024a; Chen and Mueller, 2024), but this method can introduce significant computational overhead and conflate model uncertainty with prompt-induced noise, obscuring interpretability. Accordingly, in this work, we focus on the uncertainty of open-source models, which are readily accessible and widely used. We adopt *predictive entropy*, a straightforward white-box metric computed from the model's answer logits, as our uncertainty measure for fact-checking tasks. This choice balances interpretability and computational efficiency while avoiding potential noise introduced by multiple prompts.

2.2 Linguistic Expressions of Uncertainty

Numerical uncertainty estimates do not address the sources of uncertainty, and are therefore difficult for end-users, such as fact-checkers, to interpret and act upon (Warren et al., 2025). Linguistic expressions of uncertainty may be more intuitive for people to understand than numerical ones, (Zimmer, 1983; Wallsten et al., 1993; Windschitl and

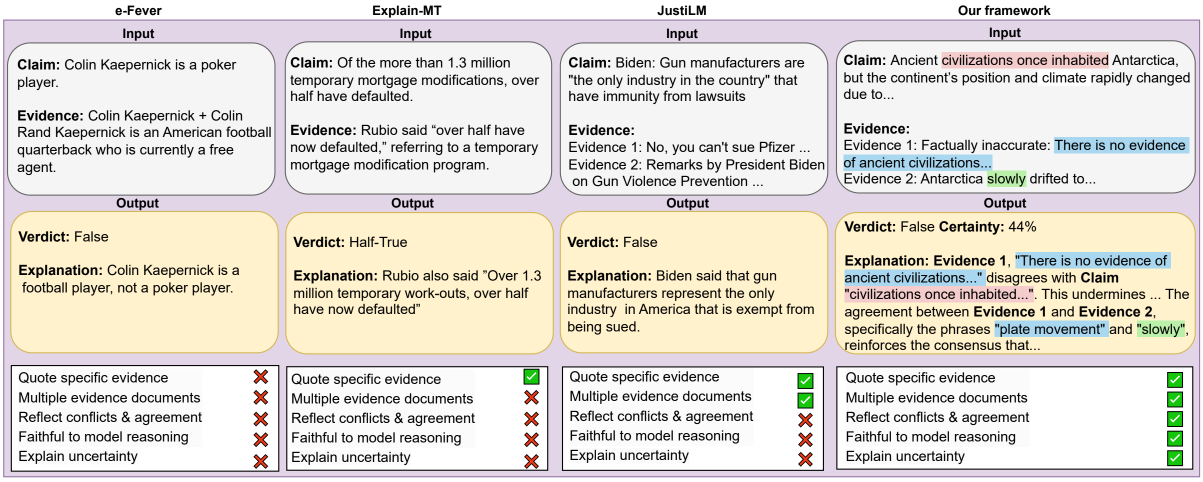

Figure 2: Explanations produced by earlier systems, e-FEVER (Stammbach and Ash, 2020), Explain-MT (Atanasova et al., 2020), and JustiLM (Zeng and Gao, 2024), compared with those from our CLUE framework. CLUE is the only approach that explicitly traces model uncertainty to the conflicts and agreements between the claim and multiple evidence passages.

Wells, 1996), and recent work has proposed models that communicate uncertainty through hedging phrases such as "I am sure" or "I doubt" (Mielke et al., 2022b,a; Lin et al., 2022; Zhou et al., 2023; Tian et al., 2023; Xiong et al., 2023; Ji et al., 2025; Zheng et al., 2023; Farquhar et al., 2024). However, these expressions are not necessarily faithful reflections of the model's uncertainty (Yona et al., 2024) and tend to overestimate the model's confidence (Tanneru et al., 2024), risking misleading users (Steyvers et al., 2025). Moreover, they do not explain *why* the model is uncertain. In this paper, we propose a method that explains sources of model uncertainty by referring to specific conflicting or concordant parts of the input that contribute to the model's confidence in the output. This approach ensures a more faithful reflection of model uncertainty and provides users with a more intuitive and actionable understanding of model confidence.

2.3 Generating Natural Language Explanations for Fact-Checking

Natural language explanations provide justifications for model predictions designed to be understood by laypeople (Wei Jie et al., 2024). NLEs have typically been evaluated by measuring the similarity between generated NLEs and human-written reference explanations using surface-level metrics such as ROUGE-1 (Lin, 2004) and BLEU (Papineni et al., 2002). In fact-checking, supervised methods have been proposed that involve extracting key sentences from existing fact-checking articles and using them as explanations (Atanasova et al., 2020). Later work proposed a post-editing mechanism to enhance the explanation coherence and fluency (Jolly et al., 2022), while others have fine-tuned models on data collected from fact-checking websites to generate explanations (Feher et al., 2025; Raffel et al., 2020; Beltagy et al., 2020). Recent work has shifted towards few-shot methods requiring no fine-tuning, for example, using few-shot prompting with GPT-3 (Brown et al., 2020) to produce evidence summaries as explanations (Stammbach and Ash, 2020) and incorporating a planning step before explanation generation (Zhao et al., 2024) to outperform standard prompting approaches. Zeng and Gao (2024) focuses on generating fact-checking justifications based on retrieval-augmented language models. However, existing methods are often not faithful to model reasoning (Atanasova et al., 2023; Siegel et al., 2024, 2025), have limited utility in fact-checking (Schmitt et al., 2024), and fail to address model uncertainty, which has been identified as a key criterion for fact-checking (Warren et al., 2025).

To this end, we introduce the first framework designed for the task of explaining sources of uncertainty in multi-evidence fact-checking. Our method analyzes span-level agreements and conflicts correlated with uncertainty scores. Unlike conventional approaches that align with human NLEs (reflecting human perspectives rather than model reasoning), our method generates explanations that are both faithful to model uncertainty and helpful to people in a fact-checking context.

3 Method

3.1 Preliminaries and Overall Framework

Our objective is to *explain why* a LLM is uncertain about a multi-evidence fact-checking instance by grounding that uncertainty in specific agreements or conflicts within the input.

Problem setup. Each input instance is a triple $X = (C, E_1, E_2)$ consisting of a claim C and two evidence pieces E_1, E_2. Note that, in this work, we set the number of evidence pieces to two for simplicity. For clarity, we denote their concatenation as $X = [x_1, \ldots, x_{|C|+|E_1|+|E_2|}]$. The task label comes from the set $\mathcal{Y} = \{\textsc{Supports}, \textsc{Refutes}, \textsc{Neutral}\}$.

Pipeline overview. Our framework proceeds in three stages:

1. **Uncertainty scoring.** We compute *predictive entropy* from the model's answer logits to obtain a scalar uncertainty score $u(X)$ (Section 3.2). This logit-based measure is model-agnostic.
2. **Conflicts/Agreement extraction.** We capture the agreements and conflicts most relevant to the model's reasoning by identifying the text-span interactions between C, E_1, and E_2 that embody these relations (Section 3.3).
3. **Explanation generation.** The model receives the extracted spans as soft constraints and produces a natural-language rationale $Y_R = [y'_1, \ldots, y'_r]$ along with its predicted label $\hat{y}$ to the identified interactions (Section 3.4).

Outputs. For each instance X, the framework returns the predicted task label $\hat{y} \in \mathcal{Y}$; the numeric uncertainty score $u(X)$; and the textual explanation $Y_R = [y'_1, \ldots, y'_r]$ that grounds the source of uncertainty in the specific agreements or conflicts between C, E_1, E_2.

3.2 Predictive Uncertainty Score Generation

To get the uncertainty of the model towards generating an answer label on a specific input sequence, we follow the previous work and get the predictive uncertainty with the entropy theory, which does not require multiple runs and is widely used in open-source models.

Specifically, we define the numeric uncertainty score u as the entropy of the softmax distribution over the model's output logits for a set of candidate answers $\mathcal{Y} = \{\textsc{Supports}, \textsc{Refutes}, \textsc{Neutral}\}$. For each candidate label $y_i \in \mathcal{Y}$:

$$P(y_i \mid X) = \frac{\exp(\text{logit}(y_i))}{\sum_{j=1}^{|\mathcal{Y}|} \exp(\text{logit}(y_j))} \quad (1)$$

where $\text{logit}(y_i)$ is the model's output logit towards candidate answer y_i given input X. $P(y_i \mid X)$ is the confidence score of model for selecting y_i as the final answer across all candidate answers within $\mathcal{Y}$. And finally, the model's uncertainty towards the input sequence X is:

$$u(X) = -\sum_{y_i \in \mathcal{Y}} P(y_i \mid X) \log P(y_i \mid X) \quad (2)$$

3.3 Conflict and Agreement Span Interaction Identification for Answer Uncertainty

To surface the conflicts and agreements that drive a model's uncertainty, we extract and then label salient span interactions among the claim C and two evidence passages, E_1 and E_2.

Span interaction extraction. For each ordered input part pair $(F, T) \in \{(C, E_1), (C, E_2), (E_1, E_2)\}$, we follow previous work (Ray Choudhury et al., 2023; Sun et al., 2025) to extract the important span interactions and their importance score to model's answer by (i) identifying the most important attention head to the model's answer prediction from its final layer, (ii) obtaining its attention matrix $\mathbf{A} \in \mathbb{R}^{(|F|+|T|)\times(|F|+|T|)}$, and (iii) symmetrizing the cross-part scores:

$$a'_{p,q} = \tfrac{1}{2}\big(\mathbf{A}_{p,q} + \mathbf{A}_{q,p}\big), \quad x_p \in F,\ x_q \in T.$$

Treating $a'_{p,q}$ as edge weights yields a bipartite token graph, which we partition into contiguous spans with the Louvain algorithm (Blondel et al., 2008). Given a $\text{span}_w \subset F$ and a $\text{span}_v \subset T$, their interaction importance is

$$a_{wv} = \frac{1}{|\text{span}_w|\,|\text{span}_v|} \sum_{x_p \in \text{span}_w} \sum_{x_q \in \text{span}_v} a'_{p,q}. \quad (3)$$

The scored interactions for (S, T) form $S_{(S,T)} = \{((\text{span}_w, \text{span}_v),\ a_{wv})\}$.

Relation labeling. To tag each span pair as an *agreement*, *disagreement*, or *unrelated*, we prompt GPT-4o (Team, 2024)[1] to assign a label $r_{wv} \in \{\text{agree}, \text{disagree}, \text{unrelated}\}$,balancing scalability and accuracy (See templates in App. H.6).

[1] https://openai.com/index/hello-gpt-4o/

After labeling all three pairs, the complete interaction set for instance X is

$$S_R = S_R(C, E_1) \cup S_R(C, E_2) \cup S_R(E_1, E_2), \quad (4)$$

where, for example, $S_R(C, E_1) = \{((\text{span}_w, \text{span}_v), a_{wv}, r_{wv})\}$. Each element links two spans with an importance score and a relation label, thereby supplying the conflict- or agreement-span interactions used in later stages.

3.4 Uncertainty Natural Language Explanation Generation

To turn the extracted conflict- and agreement spans into rationales towards model uncertainty, we rely on two complementary mechanisms. (i) **Instruction-driven prompting** embeds the spans directly in the input so the model is told which segments to reference. (ii) **Intrinsic attention steering** guides the model's own attention toward those same segments while it is generating the rationale. Both mechanisms use *self-rationalization*: the model first states its verdict $\hat{y}$ and then explains Y_R, a sequencing shown to improve faithfulness over pipeline approaches (Wiegreffe et al., 2021; Marasovic et al., 2022; Siegel et al., 2025).

Instruction-based NLE. For each instance X, we rank all labelled interactions by importance and keep the top $K = 3$, denoted $S_R^{(K)}$, to avoid too long explanations. These three span pairs are slotted into a three-shot prompt (See App.F.1), which instructs the model to explain how the highlighted agreements or conflicts influence its confidence. Finally, the standard transformer decoding process emits both the predicted label $\hat{y}$ and the accompanying explanation Y_R.

Attention steering. Instead of explicit instructions, we can guide generation by modifying attention on the fly with PASTA (Zhang et al., 2024b). Starting from the same $S_R^{(K)}$, we collect all token indices that fall inside any selected span,

$$\mathcal{I} = \{p : (\text{span}_w, \text{span}_v) \in S_R^{(K)},\ p \in \text{span}_w \cup \text{span}_v\}. \quad (5)$$

For each attention head (ℓ, h) deemed relevant to model uncertainty, let $\mathbf{A}$ be its attention matrix. We down-weight non-target tokens by β:

$$\tilde{A}_{ij} = \frac{A_{ij}}{Z_i} \begin{cases} 1 & \text{if } j \in \mathcal{I}, \\ \beta & \text{otherwise}, \end{cases} \quad (6)$$

$$Z_i = \sum_{j \in \mathcal{I}} A_{ij} + \beta \sum_{j \notin \mathcal{I}} A_{ij}. \quad (7)$$

All other heads remain unchanged. Following Zhang et al. (2024b), we steer $|H| = 100$ heads and set $\beta = 0.01$ to balance steering efficacy and prevent degeneration; see App. B for the head-selection procedure. With the steered attention in place, the transformer generates $\hat{y}$ followed by the rationale Y_R, now naturally centered on the conflict- or agreement spans that drive its uncertainty.

4 Experimental Setup

4.1 Datasets

We select two fact-checking datasets, one specific to the health domain, HealthVer (Sarrouti et al., 2021), and one closer to a real-world fact-checking scenario, DRUID (Hagström et al., 2024). These datasets were chosen because they provide multiple evidence pieces per claim, making them well-suited to our goal of explaining model uncertainty arising from the inter-evidence conflicts and agreements. For experiments, we select six hundred instances that consist of a claim and multiple pieces of evidence, and a golden label $y \in \{\text{SUPPORTS}, \text{REFUTES}, \text{NEUTRAL}\}$ from each dataset.[2]

4.2 Models

We compare three generation strategies for NLEs towards model uncertainty:

- **`Prompt_Baseline`**: A three-shot prompt baseline extending the prior few-shot NLE work (Stammbach and Ash, 2020; Zeng and Gao, 2024; Zhao et al., 2024) by explicitly asking the model to highlight conflicting or supporting spans that shape its uncertainty (See prompt template in App.F.1).
- **`CLUE-Span`**: The instruction-based variant of our CLUE where the extracted span interactions are filled into a three-shot prompt to guide the explanation generation (§3.4; prompt template in App.F.2).
- **`CLUE-Span+Steering`**: The attention steering variant of our CLUE where the same prompt as `CLUE-Span` is used. Additionally, attention steering is applied to instinctively guide the model's explanation generation toward the identified spans (§3.4; prompt template in App.F.2).

Experiments are run on three recent, open-weight, instruction-tuned LLMs of comparable

[2] While DRUID has six fine-grained fact-checking labels, we merge the labels into the above three categories to balance the label categories.

size: Qwen2.5-14B-Instruct[3] (Qwen Team, 2024), Gemma-2 9B-IT[4] (Gemma Team, 2024), and OLMo-2-1124-13B-Instruct[5] (Team OLMo et al., 2024). Each backbone is used consistently across our pipeline for span-interaction extraction, answer prediction, and NLE generation on four NVIDIA A100-SXMS-40GB GPUs. We chose these models to balance capability (reasoning and instruction-following quality) with practical constraints on inference latency and GPU memory.

5 Automatic Evaluation

5.1 Faithfulness

To assess whether the NLEs produced by our CLUE are faithful to the model's uncertainty, we adapt the Correlational Counterfactual Test (CCT) (Siegel et al., 2024) and propose an Entropy-CCT metric.

Following Siegel et al. (2024), we start by inserting a random adjective or noun into the original instance X to obtain a perturbed input X' (See App. D for details). Let $u(X)$ denote the model's uncertainty score defined by Eq. 2, unlike CCT(See details of original CCT in App.E), we measure the impact of the perturbation on the model's uncertainty with Absolute Entropy Change (AEC):

$$\Delta u(X) = \left|u(X) - u(X')\right| \tag{8}$$

For each perturbation, we record whether the inserted word appears in the generated NLE, using its presence as a proxy for importance. This yields a binary mention flag $m \in \{0, 1\}$, following Siegel et al. (2024); Atanasova et al. (2023).

Let D_m denote the set of perturbed examples where the NLE *mentions* the inserted word and $D_{\neg m}$ is the complementary set where it does not, we correlate the continuous variable Δu with the binary mention flag m via the point-biserial correlation r_{pb} (Tate, 1954). The Entropy-CCT statistic is:

$$\text{CCT}_{\text{entropy}} = r_{\text{pb}} = \frac{\mathbb{E}_m[\Delta u] - \mathbb{E}_{\neg m}[\Delta u]}{\text{Std}(\Delta u)} \cdot \sqrt{\frac{|D_m| \cdot |D_{\neg m}|}{(|D_m| + |D_{\neg m}|)^2}} \tag{9}$$

where $\mathbb{E}_m[\Delta u]$ and $\mathbb{E}_{\neg m}[\Delta u]$ are the mean absolute entropy changes for these two groups, respectively. $\text{Std}(\Delta u)$ is the standard deviation of absolute entropy changes across the full dataset.

Ultimately, this metric quantifies the alignment between changes in model uncertainty and explanatory references to input perturbations, thereby measuring how faithfully the NLEs reflect the model's uncertainty.

5.2 Span-Coverage

An uncertainty explanation should surface *all* information conveyed by the selected span interactions. We therefore compute **Span-Coverage**: the fraction of reference interactions that are explicitly mentioned in the generated NLE. Let S_{NLE} be the set of span interactions extracted from the explanation, and let $S_R(k)$ be the reference set supplied in the prompt (see §3.4). Then

$$\text{Span-Coverage} = \frac{|S_{\text{NLE}} \cap S_R(k)|}{|S_R(k)|}. \tag{10}$$

A higher value indicates the NLE covers a higher proportion of the information supplied by the extracted span interactions.

5.3 Span-Extraneous

Ideally, the explanation should mention *only* the provided interactions. We measure the proportion of mentioned interactions that *do not* belong to the reference set, denoted **Span-Extraneous**:

$$\text{Span-Extraneous} = \frac{|S_{\text{NLE}} \setminus S_R(k)|}{|S_{\text{NLE}}|}. \tag{11}$$

A lower value indicates closer alignment with the intended span interactions.

5.4 Label-Explanation Entailment

We evaluate how well the uncertainty explanation agrees with the model's predicted label by treating the task as a natural-language inference (NLI) problem. First, we convert the predicted label into a hypothesis using the template *"The claim is supported by / refuted by / neutral to the evidence."* The generated explanation serves as the premise. The resulting premise–hypothesis pair is fed to a widely used off-the-shelf language-inference model, DeBERTa-v3[6] (He et al., 2023). The Label-Explanation Entailment (LEE) score is the proportion of examples for which the NLI model predicts ENTAILMENT.

5.5 Results

For brevity, we refer to Qwen2.5-14B-Instruct, OLMo-2-1124-13B-Instruct, and Gemma-2-9B-it simply as Qwen, OLMo, and Gemma, respectively.

[3] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[4] https://huggingface.co/google/gemma-2-9b-it

[5] https://huggingface.co/allenai/OLMo-2-1124-13B-Instruct

[6] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

	HealthVer				DRUID			
Approach	Faith. (↑)	Span-Cov. (↑)	Span-Ext. (↓)	LEE (↑)	Faith. (↑)	Span-Cov. (↑)	Span-Ext. (↓)	LEE (↑)
Qwen2.5-14B-Instruct								
`Prompt`$_{\texttt{Baseline}}$	-0.028	–	–	0.74	-0.08	–	–	0.60
`CLUE-Span`	0.006	0.33	0.68	0.75	0.089	0.20	0.38	**0.78**
`CLUE-Span+Steering`	**0.033**	**0.44**	**0.53**	**0.80**	**0.102**	**0.28**	**0.20**	0.77
OLMo-2-1124-13B-Instruct								
`Prompt`$_{\texttt{Baseline}}$	-0.10	–	–	0.55	-0.13	–	–	0.53
`CLUE-Span`	0.005	0.10	0.83	0.61	0.014	0.08	0.79	0.65
`CLUE-Span+Steering`	**0.020**	**0.23**	**0.77**	0.68	**0.099**	**0.15**	**0.70**	**0.69**
Gemma-2-9B-It								
`Prompt`$_{\texttt{Baseline}}$	-0.105	–	–	0.66	-0.12	–	–	0.57
`CLUE-Span`	0.007	0.34	0.59	0.82	0.043	0.23	0.43	0.76
`CLUE-Span+Steering`	**0.021**	**0.39**	**0.50**	**0.85**	**0.098**	**0.30**	**0.47**	**0.81**

Table 1: Uncertainty NLE evaluation results across the HealthVer and DRUID datasets (§4.1). For each model (§4.2) we compare `Prompt`$_{\texttt{Baseline}}$, `CLUE-Span`, and `CLUE-Span+Steering` on four metrics: Faith. (§5.1), Span-Cov. (§5.2), Span-Ext. (§5.3), and LEE (§5.4). Bold values mark the best result per metric for each dataset–model pair; "–" indicates inapplicable metrics for `Prompt`$_{\texttt{Baseline}}$, as it is not supplied with extracted span interactions.

Faithfulness. We use Entropy-CCT, a point–biserial correlation bounded by $-1 \leq r_{\text{pb}} \leq 1$ (Eq. 9), to measure the faithfulness of the NLEs to the model's uncertainty (§5.1). When $r_{\text{pb}} = 0$, the explanation mentions high- and low-impact perturbation words equally often; every $+0.01$ adds roughly *one percentage point (pp)* to the chance that the explanation names a token that is *truly influential for the model's predictive uncertainty* (App. G).

Table 1 shows that **`Prompt`$_{\texttt{Baseline}}$ is *non-faithful* in all six settings** with r_{pb} are all negative values ranging from -0.03 to -0.13. Thus its NLEs mention truly influential tokens 3–13 pp *less* often than uninfluential ones—the opposite of faithful behaviour. **Both variants of our CLUE reverse this trend.** Presenting span interactions in the prompt (`CLUE-Span`) raises every correlation to non-negative values and peaks at $r_{\text{pb}} = 0.089$ on the DRUID–Qwen setting. This means the explanation now mentions about 17 pp more often than `Prompt`$_{\texttt{Baseline}}$($r_{\text{pb}} = -0.080$). Adding attention steering (`CLUE-Span+Steering`) lifts the r_{bp} scores to 0.033 on HEALTHVER and 0.102 on DRUID with Qwen model, i.e., net gains of +6 pp and +18 pp over `Prompt`$_{\texttt{Baseline}}$. Moreover, four of the six positive correlations produced by `CLUE-Span+Steering` are significant at $p < 0.01$ (Table 3), confirming that the improvements are both substantial and statistically reliable. **Particularly large jumps of OLMo on Druid dataset (up to $\Delta r_{\text{pb}} = +0.23 \approx +23$ pp)** suggest that span-interaction guidance from our CLUE framework is most beneficial for models that initially struggle to align explanations with predictive uncertainty.

Other Properties We also evaluate three properties of the generated NLEs: (i) **Span-Coverage** of extracted conflict-/agreement- span interactions(§5.2) and (ii) **Span-Extraneous**: mention of non-extracted spans(§5.3), (iii) **Label-Explanation Entailment** with the generated fact-checking label(§5.4). As Table 1 shows, **`CLUE-Span+Steering` outperforms `CLUE-Span` in both span-coverage and span-extraneous**, consistent with the attention steering method's effectiveness in helping the model better focus on provided highlights during generation (Zhang et al., 2024b). Absolute numbers, however, remain modest (peak span-coverage: .44, span-extraneous: .20 with Qwen). A span-coverage of 1 means the NLE cites every extracted interaction, while a span-extraneous of 0 means it adds none beyond them. This gap highlights considerable headroom for better integrating critical span interactions into the explanations. Among the three backbones, **Qwen attains the highest span-coverage and the lowest span-extraneous scores**, a trend that likely reflects its stronger instruction-following ability (see benchmark scores in Appendix A), and thus larger or more capable models might narrow the gap further. **Both variants of our framework achieve stronger label-explanation entailment scores than the baseline**, yielding explanations that stay logically aligned with the predicted labels while remaining faithful to the model's uncertainty patterns (as demonstrated in our faithfulness analysis).

6 Human Evaluation

6.1 Method

We recruited N=12 participants from Prolific.com (https://www.prolific.com/) to rank explanations generated by **Prompt$_{\text{Baseline}}$**, **CLUE-Span**, **CLUE-Span+Steering** for 40 instances (20 from DRUID, 20 from HealthVer) (See details in App.H.1). Adapting Atanasova et al. (2020), participants ranked explanations in descending order (1st, 2nd, 3rd) according to five criteria, complementary to our automatic evaluation metrics:

- **Helpfulness.** The explanation offers information that aids readers to judge the claim and fact-check.
- **Coverage.** The explanation captures *all* salient information in the input that matters for the fact check. This differs from automatic Span-Coverage (§5.2), which counts overlap with pre-extracted spans.
- **Non-redundancy.** The explanation does not offer irrelevant or repetitive information to the input. This differs from automatic Span-Extraneous (§5.3) which counts mentions outside the extracted spans.
- **Consistency.** The explanation contains logically contradictory statements to the input. This differs from automatic Label-Explanation Entailment (§5.4), which tests label–explanation alignment.
- **Overall Quality.** Overall ranking of explanations by their overall quality, considering all criteria above.

6.2 Results

Table 4 in App. H.2 shows the study participant evaluation results. Annotator agreement was moderate-low, which we attribute to the relative complexity of the task and individual differences in how the information was perceived (see App. H.7).

The explanations generated by CLUE were preferred by our evaluators to those generated using **Prompt$_{\text{Baseline}}$**: the **explanations generated by CLUE-Span+Steering were rated as most helpful, highest coverage, and containing the least amount of redundant information**, while **those from CLUE-Span were judged to have the highest consistency and overall quality**. Although **CLUE-Span+Steering** achieves the highest faithfulness (see §5.5), our participants judged its overall quality slightly lower than that of **CLUE-Span**. A possible reason for this is that although **CLUE-Span+Steering** adheres closely to the top-K=3 extracted span interactions (as reflected in its higher Span-Coverage and lower Span-Extraneous scores), it may produce explanations that are slightly less internally consistent or fluent. In contrast, **CLUE-Span** is less faithful to those extracted spans, but may capture additional points that study participants deemed important, likely because the spans identified as important for model do not fully overlap with those identified by humans (Ray Choudhury et al., 2023), highlighting the well-documented trade-off between faithfulness and plausibility (Agarwal et al., 2024). Future work on improving the plausibility of the span interactions while retaining their faithfulness may therefore improve the human evaluation scores for **CLUE-Span+Steering**.

Finally, we observed slight variation between datasets: **CLUE-Span+Steering** tended to be rated higher than **CLUE-Span** for DRUID, and vice versa for HealthVer. This may arise from differences in length and complexity of the input: DRUID evidence documents, retrieved from heterogeneous online sources, may have benefited from the attention steering more than HealthVer evidence documents, consisting of focused, shorter extracts from scientific abstracts.

7 Conclusion

We present the first framework, CLUE, for generating NLEs of model uncertainty by referring to the conflicts and agreements between claims and multiple pieces of evidence in a fact-checking task. Our method, evaluated across three language models and two datasets, demonstrates significant improvements in both faithfulness to model uncertainty and label consistency compared to standard prompting. Evaluations by human participants further demonstrate that the explanations generated by CLUE are more helpful, more informative, less redundant, and better logically aligned with the input. This work establishes a foundation for explainable fact-checking systems, providing end users (e.g., fact-checkers) with grounded, faithful explanations that reflect the model's uncertainty.

Limitations

Our paper proposes a novel framework for generating NLEs towards the model's uncertainty by explicitly pointing to the conflicts or agreements within the claim and multi-evidence interactions. While our framework demonstrates improved explanation quality through rigorous evaluation across three language models and two datasets, we acknowledge several limitations that present opportunities for future research.

Regarding the model selection, our experiments are constrained to medium-sized models (Qwen2.5-14B-Instruct, Gemma2-9B-it, and OLMo2-13B-Instruct) due to computational limitations. Although these models show significant improvements over baseline performance, our results suggest that larger models (e.g., 70B parameter scale) with enhanced instruction-following and reasoning capabilities might further improve explanation quality — particularly for coverage and redundancy metrics. Our framework's modular design readily accommodates such scaling.

In this study we focus on HealthVer and DRUID datasets, where claims are paired with discrete pieces of evidence, ideal for studying evidence-conflict scenarios. Future work could investigate more complex evidence structures (e.g., long-form documents), diverse fact-checking sources, and scenarios with more than two pieces of evidence per claim to better reflect real-world fact-checking challenges.

While our evaluation with laypeople confirms that our framework produces explanations of higher quality than prompting, expert evaluations (e.g., with professional fact-checkers) are needed to assess practical utility in high-stakes settings.

Regarding the scope of the uncertainty sources, our work specifically explains model uncertainty arising from evidence conflicts. While this captures a critical subset of cases, real-world uncertainty may also stem from other sources, including insufficient evidence, knowledge gaps in the model, and context-memory conflicts. We view this work as a foundational step toward broader research on model uncertainty explanation.

Ethical Considerations

Our work is limited to examining claims, evidence, and explanations in English, and so our results may not be generalisable to other languages. As the task involved complex reasoning about technical subjects, we screened our participants to be native English speakers to ensure that they could fully understand the material and increase the chances of high-quality responses (see H.1 for details). However, this criteria may also introduce or reinforce existing biases and limit the generalisability of our findings. Participants were informed about the study and its aims before agreeing to provide informed consent. No personal data was collected from participants and they received fair payment for their work (approximately 9 GBP/hour).

This work concerns automated fact-checking, which aims to reduce the harm and spread of misinformation, but nevertheless has the potential for harm or misuse through model inaccuracy, hallucination, or deployment for censorship. Our current work aims to provide explanation that allow users to examine the outputs of these systems more critically, and so we do not see any immediate risks associated with it.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200.

Karl Cobbe, Vivek Kosaraju, Mohammad Bavarian, et al. 2021. GSM8K: A Dataset for Grade School Math Word Problems. In *Proceedings of the Neural Information Processing Systems Datasets and Benchmarks Track*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting Hallucinations in Large Language Models using Semantic Entropy. *Nature*, 630(8017):625–630.

Darius Feher, Abdullah Khered, Hao Zhang, Riza Batista-Navarro, and Viktor Schlegel. 2025. Learning to Generate and Evaluate Fact-Checking Explanations with Transformers. *Engineering Applications of Artificial Intelligence*, 139:109492.

Nicolo' Fontana, Francesco Corso, Enrico Zuccolotto, and Francesco Pierri. 2025. Evaluating open-source large language models for automated fact-checking. *Preprint*, arXiv:2503.05565.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology.

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. *Preprint*, arXiv:2412.17031.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations. *arXiv preprint arXiv:2503.14477*.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Maurice G Kendall and B. Babington Smith. 1939. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. *arXiv preprint*. ArXiv:2011.03870 [cs].

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *arXiv preprint arXiv:2302.09664*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*. `https://openreview.net/forum?id=8s8K2UZGTZ`.

Dawn Liu, Marie Juanchich, Miroslav Sirota, and Sheina Orbell. 2020. The Intuitive Use of Contextual Information in Decisions Made with Verbal and Numerical Quantifiers. *Quarterly Journal of Experimental Psychology*, 73(4):481–494.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022a. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022b. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel Language Entropy: Fine-Grained Uncertainty Quantification for LLMs from

Semantic Similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 21(140):1–67.

Sagnik Ray Choudhury, Pepa Atanasova, and Isabelle Augenstein. 2023. Explaining interactions between text spans. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12709–12730, Singapore. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2157–2174, New York, NY, USA. Association for Computing Machinery.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. The Probabilities Also Matter: A More Faithful Metric for Faithfulness of Free-Text Explanations in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–546.

Noah Y Siegel, Nicolas Heess, Maria Perez-Ortiz, and Oana-Maria Camburu. 2025. Faithfulness of LLM Self-Explanations for Commonsense Tasks: Larger Is Better, and Instruction-Tuning Allows Trade-Offs but Not Pareto Dominance. *arXiv preprint arXiv:2503.13445*.

Dominik Stammbach and Elliott Ash. 2020. e-FEVER: Explanations and Summaries for Automated Fact Checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, pages 1–11.

Jingyi Sun, Pepa Atanasova, and Isabelle Augenstein. 2025. Evaluating input feature explanations through a unified diagnostic evaluation framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10559–10577.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying Uncertainty in Natural Language Explanations of Large Language Models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR.

Robert F Tate. 1954. Correlation between a Discrete and a Continuous Variable. Point-Biserial Correlation. *The Annals of mathematical statistics*, 25(3):603–607.

OpenAI Team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. 2020. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, 144:102493.

Thomas S Wallsten, David V Budescu, Rami Zwick, and Steven M Kemp. 1993. Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numerical Terms. *Bulletin of the Psychonomic Society*, 31(2):135–138.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul D Windschitl and Gary L Wells. 1996. Measuring Psychological Uncertainty: Verbal versus Numeric Methods. *Journal of Experimental Psychology: Applied*, 2(4):343.

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*.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024b. Tell your model where to attend: Post-hoc attention steering for LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated Fact-Checking with Planning and Customized Action Reasoning Using Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. 2023. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

Alf C Zimmer. 1983. Verbal vs. Numerical Processing of Subjective Probabilities. In *Advances in psychology*, volume 16, pages 159–182. Elsevier.

A Backbone model performance on public benchmarks

Table 2 summarises the publicly reported five-shot results on two standard reasoning benchmarks. All figures are taken verbatim from the official model cards or accompanying technical reports. Figures are copied from the official model cards.

These numbers corroborate our claim that Qwen2.5-14B-Instruct is the strongest of the three for instruction-following and reasoning.

B Method: Selecting attention heads to steer

Following Zhang et al. (2024b), we steer only a selected subset of attention heads rather than all of

Model	Params	MMLU	GSM8K
Qwen2.5-14B-Instruct (Qwen Team, 2024)	14.7 B	79.7	90.2
Gemma-2-9B-IT (Gemma Team, 2024)	9.0 B	71.3	68.6
OLMo-2-1124-13B-Instruct (Team OLMo et al., 2024)	13 B	67.5	54.2

Table 2: Benchmark scores on MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) are used to characterize instruction-following and reasoning strength.

them, because targeted steering yields larger gains in output quality. Our selection criterion, however, differs from theirs: instead of ranking heads by their impact on task accuracy, we rank them by how strongly they affect the model's *predictive uncertainty* during fact-checking.

Concretely, for each fact-checking dataset chosen in this work(see details in §4.1), D, we draw a validation subset D_d with $|D_d| = 300$ examples. For every input $X \in D_d$, we compute the model's baseline uncertainty score $u(X)$ when it predicts the fact-checking label as stated in §3.2. Then, for each attention head identified by layer ℓ and index h, we zero out that head, re-run the model, and measure the absolute change in uncertainty

$$\Delta u(X, \ell, h) \;=\; \big| u(X) \;-\; u_{/o(l,h)}(X) \big|.$$

Averaging $\Delta u(X, l, h)$ over all $X \in D_d$ yields a single importance score for head (ℓ, h). We rank the heads by this score and keep the top t heads for each dataset and each model. Note that we set $t = 100$ in line with the recommendation of Zhang et al. (2024b) and to balance steering effectiveness against the risk of degeneration.

C Prompt Example for Assigning Relation Labels to Captured Span Interactions

To identify agreements and conflicts between the claim and the two evidence passages, we use the prompt in Figure 3 to label each extracted span interaction (see §3.3).

```
You are a helpful assistant. Your task:

1. Read the claim and its two evidence passages (E1,
     E2).
2. For each supplied span interaction, decide
     whether the two spans
   AGREE, DISAGREE, or are UNRELATED, taking the
        full context into account.
3. Output the span pairs exactly as given, followed
     by
   "relation: agree|disagree|unrelated".

Return format:
  1. "SPAN A" - "SPAN B"  relation: <agree|disagree|
      unrelated>
  2. ...
  3. ...

### SHOT 1  (annotated example)
Claim: [...]
Evidence 1: [...]
Evidence 2: [...]

Span interactions (to be labelled):
  1. "[...]" - "[...]"
  2. "[...]" - "[...]"
  3. "[...]" - "[...]"

Expected output:
  1. "[...]" - "[...]"  relation: ...
  2. "[...]" - "[...]"  relation: ...
  3. "[...]" - "[...]"  relation: ...

### SHOT 2  % omitted for brevity
### SHOT 3  % omitted for brevity

### NEW INSTANCE  (pre-filled for each new example)
Claim: {CLAIM}
Evidence 1: {E1}
Evidence 2: {E2}
Span interactions:
  1. "{SPAN1-A}" - "{SPAN1-B}"
  2. "{SPAN2-A}" - "{SPAN2-B}"
  3. "{SPAN3-A}" - "{SPAN3-B}"
```

Figure 3: Prompt template for span interaction relation labelling.

D Perturbation details for faithfulness measurement

To evaluate how faithfully each NLE reflects model uncertainty, we generate multiple counterfactuals per instance, following Atanasova et al. (2020) and Siegel et al. (2024) (see §5.1). For every input, comprising one claim and two evidence passages, we first tag part-of-speech with spaCy, then choose seven random insertion sites. At each site we insert either (i) a random adjective before a noun or (ii) a random adverb before a verb. The candidate modifiers are drawn uniformly from the full WordNet lists of adjectives and adverbs. Because we sample three random candidates for each of the four positions, this procedure yields $4 \times 3 = 12$ perturbations per instance, providing a sufficient set for the subsequent Entropy-CCT evaluation, in which we check whether the NLE mentions the inserted word and correlate that mention with the uncertainty change induced by each perturbation.

E Differences Between Entropy-CCT and CCT

In CCT test, Total Variation Distance (TVD) is computed between two probability distributions P and Q as $\text{TVD}(P, Q) = \frac{1}{2}\sum_i |P_i - Q_i|$, measuring the absolute change in class-wise probabilities. We instead operate on the entropies of those distributions, yielding a single-valued measure of uncertainty shift.

F Prompt template for **Prompt$_{\text{Baseline}}$**, **CLUE-Span** and **CLUE-Span+Steering** on Healthver and Druid dataset

We designed two prompt templates for our experiments. The baseline prompt (Figure 4) gives the model no span interactions; instead, it must first identify the relevant agreements or conflicts and then discuss them in its explanation. In contrast, the prompt used by our CLUE framework (Figure 5) supplies the three pre-extracted span interactions (§3.3). The model is explicitly instructed to base its explanation on these spans, ensuring that the rationale remains grounded in the provided evidence.

F.1 Prompt template for **Prompt$_{\text{Baseline}}$**

To generate NLEs about model uncertainty without span-interaction guidance, we craft a three-shot prompt that instructs the model to identify the interactions most likely to affect its uncertainty and to explain how these relations they represent affect it. (See Figure 4).

F.2 Prompt template for **CLUE-Span** and **CLUE-Span+Steering**

To generate NLEs about model uncertainty with the span-interaction guidance, we craft a three-shot prompt that instructs the model to discuss how these interactions, along with the relations they represent, affect its uncertainty. (See Figure 5).

G Extended Statistical Analysis of Faithfulness Scores

This section elaborates on the statistical evaluation of faithfulness regarding (i) recalling the definition and intuitive interpretation of the point–biserial coefficient r_{pb}(E.q. 9), (ii) outlining the t-test used to assess significance, (iii) reporting the faithfulness results (§5.1) along with statistical results. Note that, each dataset is evaluated on $n = 600 \times 12 =$ 7,200 perturbations with 600 instances with 12 perturbations each (see App. D). and (iv) demonstrating through concise numerical summaries that both **CLUE-Span** and **CLUE-Span+Steering** are significantly more faithful than the **Prompt$_{\text{Baseline}}$**.

```
You are a helpful assistant. Your tasks:
1. Determine the relationship between the claim and
     the two evidence passages.
2. Explain your prediction's uncertainty by
     identifying the three most
   influential span interactions from Claim-Evidence
        1, Claim-Evidence 2,
   and Evidence 1-Evidence 2, and describing how
       each interaction's relation
   (agree, disagree, or unrelated) affects your
       overall confidence.
Return format: [Prediction] [Explanation]

### SHOT 1
Input
  Claim: [...]
  Evidence 1: [...]
  Evidence 2: [...]
Output
  [Prediction: ...] [Explanation: ...]

### SHOT 2   % omitted for brevity
### SHOT 3   % omitted for brevity

### NEW INSTANCE
Claim: {CLAIM}
Evidence 1: {E1}
Evidence 2: {E2}
Your answer:
```

Figure 4: Three-shot prompt for **Prompt$_{\text{Baseline}}$** (Shots 2–3 omitted) on the HealthVer and DRuiD datasets.

G.1 Interpreting r_{pb} and Δr_{pb}

The Entropy-CCT score is the point-biserial correlation (Tate, 1954) between the absolute entropy change $|\Delta u|$ and the binary mention flag m. Because it is mathematically identical to a Pearson r computed between one continuous and one binary variable, it obeys $-1 \leq r_{\text{pb}} \leq 1$. When $r_{\text{pb}} = 0$, it means the high- and low-impact perturbations are mentioned equally often. If the two strata are roughly balanced, every $+0.01$ in r_{pb} increases the probability that a truly uncertainty-influential token is mentioned by about one percentage point (pp). A *gain* Δr_{pb} therefore translates to an *absolute* improvement of $\approx |\Delta r_{\text{pb}}| \times 100$,pp in mention rate. For instance, moving from -0.08 to $+0.06$ is a swing of 0.14, corresponding to, 14,pp.

G.2 Significance testing

Because the point-biserial is a Pearson correlation, the familiar t–test applies:

```
You are a helpful assistant. Your tasks:
1. Determine the relationship between the claim and
     the two evidence passages.
2. Explain your prediction's uncertainty by
     referring to the three span
   interactions provided below (Claim-Evidence 1,
        Claim-Evidence 2,
   Evidence 1-Evidence 2) and describing how each
        interaction's relation
   (agree, disagree, or unrelated) affects your
        overall confidence.
Return format: [Prediction] [Explanation]

### SHOT 1
Input:
  Claim: [...]
  Evidence 1: [...]
  Evidence 2: [...]
  Span interactions:
    1. ''[...]'' - ''[...]''  (C-E1)  relation:
         [...]
    2. ''[...]'' - ''[...]''  (C-E2)  relation:
         [...]
    3. ''[...]'' - ''[...]''  (E1-E2) relation:
         [...]
Output:
  [Prediction: ...] [Explanation: ...]

### SHOT 2   % omitted for brevity
### SHOT 3   % omitted for brevity

### NEW INSTANCE
Claim: {CLAIM}
Evidence 1: {E1}
Evidence 2: {E2}
Span interactions (pre-filled):
    1. ''{SPAN1-A}'' - ''{SPAN1-B}''  (C-E1)
         relation: {REL1}
    2. ''{SPAN2-A}'' - ''{SPAN2-B}''  (C-E2)
         relation: {REL2}
    3. ''{SPAN3-A}'' - ''{SPAN3-B}''  (E1-E2)
         relation: {REL3}
Your answer:
```

Figure 5: Three-shot prompt for **CLUE-Span** and **CLUE-Span+Steering** (Shots 2–3 omitted) on the HEALTHVER and DRUID datasets.

$$t = r_{\text{pb}} \sqrt{\frac{n-2}{1-r_{\text{pb}}^2}}, \tag{12}$$

$$t \sim t_{(n-2)} \qquad \text{under } H_0 : r_{\text{pb}} = 0. \tag{13}$$

With $n = 7,200$ we have df $= 7,198$; the critical two-sided values are $|t| > 1.96$ for $p < 0.05$ and $|t| > 2.58$ for $p < 0.01$.

G.3 Faithfulness with significance results

Table 3 shows the point-biserial coefficients r_{pb}, which is our faithfulness measurement for model uncertainty(See, E.q.9), the associated t statistics, and two-sided p values for every model–method pair. Values that meet the stricter $p < 0.01$ criterion are highlighted in bold.

Across both datasets and all three backbones, the **Prompt$_{\text{Baseline}}$** exhibits negative correlations, implying an *non-faithful* tendency to highlight low-impact tokens within the generation NLEs, with mean $= -0.094$. The prompt-only variant of our CLUE framework **CLUE-Span** neutralises this bias and turns the average into $+0.027$; three of its six coefficients are clear $p < 0.01$, indicating a modest but significant improvement regarding faithfulness.

The full **CLUE-Span+Steering** variant pushes the mean to $+0.062$ and achieves $p < 0.01$ in four of six settings. Interpreting these numbers via §G.1, the switch from -0.094 to $+0.062$ yields a *absolute* increase of $(0.062 - (-0.094)) \times 100! \approx$!16, pp in the probability that a truly influential token of uncertainty is named in the NLE, which is easily noticeable in qualitative inspection.

The consistently positive, statistically significant gains therefore substantiate the claim made in the main text: CLUE produces markedly more faithful NLEs towards model uncertainty than the **Prompt$_{\text{Baseline}}$**, and the steer variant is particularly beneficial for models that initially struggle with uncertainty attribution.

H Human Evaluation Details

H.1 Participants and Materials

Participants We recruited N=12 participants from Prolific.com (https://www.prolific.com/), screened to be native English speakers from Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. The study was approved by our institution's Research Ethics Committee (reference number to be added after anonymous review period).

Materials Explanations for 40 instances (20 from DRUID, 20 from HealthVer, selected at random) were evaluated in total. Each participant annotated explanations for 10 instances (5 labelled 'True', 5 labelled 'False'), in addition to two attention check instances which were used to screen responses for quality. For each instance, participants were provided with a claim, two evidence documents, model verdict, model numerical certainty, and three alternative explanations (see Figure 6 in H.6). Explanations were generated using Qwen2.5-14b-instruct (Qwen Team, 2024) based on its automatic evaluation performance.

Procedure Participants read information about the study (see H.3) and provided informed consent (see H.4) before reading detailed task instructions and completing a practice example of the task (see H.5). The task took approximately 20 minutues, and participants were paid £3 for their work.

Model	Method	r_{pb}	t	p
		HealthVer		
Qwen2.5-14B-Instruct	Prompt$_{\text{Baseline}}$	-0.028	-2.38	1.7×10^{-2}
	CLUE-Span	$+0.006$	$+0.51$	6.1×10^{-1}
	CLUE-Span+Steering	$+0.033$	$+2.80$	$\mathbf{5.1 \times 10^{-3}}$
OLMo-2-1124-13B-Instruct	Prompt$_{\text{Baseline}}$	-0.100	-8.53	$\mathbf{< 10^{-15}}$
	CLUE-Span	$+0.005$	$+0.42$	6.7×10^{-1}
	CLUE-Span+Steering	$+0.020$	$+1.70$	9.0×10^{-2}
Gemma-2-9B-IT	Prompt$_{\text{Baseline}}$	-0.105	-8.96	$\mathbf{< 10^{-15}}$
	CLUE-Span	$+0.007$	$+0.59$	5.5×10^{-1}
	CLUE-Span+Steering	$+0.021$	$+1.78$	7.5×10^{-2}
		DRUID		
Qwen2.5-14B-Instruct	Prompt$_{\text{Baseline}}$	-0.080	-6.81	$\mathbf{9.8 \times 10^{-12}}$
	CLUE-Span	$+0.089$	$+7.58$	$\mathbf{3.4 \times 10^{-14}}$
	CLUE-Span+Steering	$+0.102$	$+8.70$	$\mathbf{< 10^{-15}}$
OLMo-2-1124-13B-Instruct	Prompt$_{\text{Baseline}}$	-0.130	-11.12	$\mathbf{< 10^{-15}}$
	CLUE-Span	$+0.014$	$+1.19$	2.3×10^{-1}
	CLUE-Span+Steering	$+0.099$	$+8.44$	$\mathbf{< 10^{-15}}$
Gemma-2-9B-IT	Prompt$_{\text{Baseline}}$	-0.120	-10.26	$\mathbf{< 10^{-15}}$
	CLUE-Span	$+0.043$	$+3.65$	$\mathbf{2.6 \times 10^{-4}}$
	CLUE-Span+Steering	$+0.098$	$+8.35$	$\mathbf{< 10^{-15}}$

Table 3: Detailed faithfulness evaluation results for baseline method **Prompt$_{\text{Baseline}}$**, and two variants of our CLUE framework **CLUE-Span** and **CLUE-Span+Steering** on Healthver and Druid dataset based on Qwen2.5-14B-Instruct(Qwen Team (2024)), OLMo-2-1124-13B-Instruct(Team OLMo et al. (2024))and Gemma-2-9B-IT(Gemma Team (2024)). Point-biserial correlation r_{pb} is our Entropy-CCT measurement(§5.1), along with t statistic and two-sided p-value for each model–method pair (n = 7,200, df = 7,198). Entries with $p < 0.01$ are bold.

H.2 Human evaluation results

Due to space limitations, we present the human evaluation results in Table 4.

H.3 Human evaluation information screen

Thank you for volunteering to participate in this study! Before you decide whether you wish to take part, please read this information screen carefully.

1. What is the project about?

Our goal is to make sure that AI fact-checking systems can explain the decisions they produce in ways that are understandable and useful to people. This survey is part of a project to help us understand what kinds of explanations are helpful and why.

2. What does participation entail?

You are invited to help us explore what kinds of explanations work better in fact-checking. In this task you will see claims, an AI system's prediction about whether this claim is true or false and corresponding evidence used to make the prediction. You will also see an explanation for why the AI system is certain or uncertain about its prediction to help you decide how to interpret the true/false prediction. We ask you to evaluate the explanations along 5 different dimensions (the detailed explanation of the task is on the next page). All participants who complete the survey will receive a payment of £3. There is no cost to you for participating. You may refuse to participate or discontinue your involvement at any time without penalty.

3. Source of funding

This project has received funding from <redacted for anonymous review>

4. Consenting to participate in the project and withdrawing from the research

You can consent to participating in this study by ticking the box on the next page of the study. Participation in the study is completely voluntary. Your decision not to consent will have no adverse consequences. Should you wish to withdraw during the experiment you can simply quit the webpage. All incomplete responses will be deleted. After you have completed the study and submitted your responses, it will no longer be possible to withdraw from the study, as your data will not be identifiable and able to linked to you.

5. Possible benefits and risks to participants

By participating in this study you will be contributing to research related to understanding what kinds of explanations are useful to people who use or who are impacted by automated fact checking systems. This is a long-term research project, so the benefits

	Prompt$_{\text{Base}}$	CLUE-S	CLUE-SS
Helpfulness			
Overall	2.025	1.892	**1.867**
DRUID	1.9	1.917	**1.767**
HealthVer	2.15	**1.867**	1.967
Consistency			
Overall	1.875	**1.783**	1.817
DRUID	1.717	1.75	**1.617**
HealthVer	2.033	**1.817**	2.017
Non-redundancy			
Overall	2.05	1.908	**1.833**
DRUID	1.983	1.983	**1.683**
HealthVer	2.117	**1.833**	1.983
Coverage			
Overall	1.967	1.775	**1.758**
DRUID	1.767	1.75	**1.617**
HealthVer	2.167	**1.8**	1.9
Overall Quality			
Overall	1.967	**1.908**	1.925
DRUID	1.9	1.9	**1.817**
HealthVer	2.033	**1.917**	2.033

Table 4: Mean Average Rank (MAR) for the five human-evaluation criteria applied to explanations from **Qwen2.5-14B-Instruct** on the HEALTHVER and DRUID datasets (chosen for its high faithfulness; see §5.5). `Prompt_Baseline`, `CLUE-Span` (`CLUE-S`), and `CLUE-Span+Steering` (`CLUE-SS`) are compared. Lower MAR means a better (higher) average rank; the best score in each row is boldfaced.

of the research may not be seen for several years. It is not expected that taking part will cause any risk, inconvenience or discomfort to you or others.

6. What personal data does the project process?

The project does not process any personal data.

7. Participants' rights under the <data regulation redacted for anonymous review>

As a participant in a research project, you have a number of rights under <data regulation redacted for anonymous review>. Your rights are specified in the <institution redacted for anonymous review> privacy policy. <link redacted for anonymous review>

8. Person responsible for storing and processing of data

<redacted for anonymous review>

Please click 'Next' to read more about consenting to participate in the study.

H.4 Human Evaluation Consent Form

We hereby request your consent for processing your data. We do so in compliance with <data regulation redacted for anonymous review>. See the information sheet on the previous screen for more details about the project and the processing of your data.

- I confirm that I have read the information sheet and that this forms the basis on which I consent to the processing of my data by the project.
- I hereby give my consent that <institution> may register and process my data as part of the <redacted for anonymous review> project.
- I understand that any data I provide will be anonymous and not identifiable to me.
- I understand that my anonymous response data will be retained by the study team.
- I understand that after I submit my responses at the end of the study, they cannot be destroyed, withdrawn, or recalled, because they cannot be linked with me.
- I understand that there are no direct benefits to me from participating in this study
- I understand that anonymous data shared through publications or presentations will be accessible to researchers and members of the public anywhere in the world, not just the <location redacted for anonymous review>.
- I give my consent that the anonymous data I provided may be stored in a database for new research projects after the end of this project.
- I give permission for my anonymous data to be stored for possible future research related to the current study without further consent being required.
- I understand I will not be paid for any future use of my data or products derived from it.

By checking this box, I confirm that I agree to the above and consent to take part in this study.
☐ I consent

H.5 Evaluation Task Instructions

What do I have to do?
In this study you will see claims, an AI system's prediction about whether this claim is true or

false, how certain the system is about its label, and the corresponding evidence used to make the prediction. You will also see three different explanations for why the AI system is certain or uncertain about its prediction. These explanations are intended help you decide how to interpret the true/false prediction.

Your task is to **evaluate the quality of the explanations** provided, **not** the credibility of the claims and evidence.

What information will I be shown?

You will be shown examples of claims, evidence document, verdicts and explanations.

- A claim is some statement about the world. It may be true, false, or somewhere in between.
- Additional information is typically necessary to verify the truthfulness of a claim - this is referred to as evidence or evidence document. An evidence document consists of one or several sentences extracted from an external source for the particular claim. In this study, you will see two evidence documents that have been retrieved for a claim. These evidence documents may or may not agree with each other.
- Based on the available evidence, a verdict is reached regarding whether a claim is true or false.
- Uncertainty often arises when evaluating the claim and evidence to reach a verdict. Each verdict is accompanied by a numerical uncertainty score which represents the AI system's confidence that its predicted verdict is correct.
- You will see 3 alternative explanations for where uncertainty arises with regard to the verdict. Note that these explanations focus on the AI system's uncertainty, not the verdict itself.
- You are asked to evaluate the explanations according to 5 different properties. The properties are as follows:

  Helpfulness. The explanation contains information that is helpful for evaluating the claim and the fact check.

  Coverage. The explanation contains important, salient information and does not miss any important points that contribute to the fact check.

  Non-redundancy. The explanation does not contain any information that is redundant/repeated/not relevant to the claim and the fact check.

  Consistency. The explanation does not contain any pieces of information that are contradictory to the claim and the fact check.

  Overall Quality. Rank the explanations by their overall quality.

- Please rank the explanations in descending order. For example, you should rank the explanation that you think is most helpful as '1', and the explanation that you think is least helpful as '3'. If two explanations appear almost identical, you can assign them the same ranking, but as a general rule, you should try rank them in hierarchical order.

- The three explanations, Explanation A, Explanation B, and Explanation C, will appear in a different order throughout the study, so you may need to pay some attention to which is which.

Important: Please only consider the provided information (claim, evidence documents, and explanations) when evaluating explanations. Sometimes you will be familiar with the claim, but we ask you to approach each claim as new, whether or not you have seen it before. It doesn't matter whether you personally agree or disagree with the claim or evidence – we are asking you to evaluate what the AI produces: if you were to see this claim for the first time, would you find the explanation provided by the AI useful? On the next page, you will see an example of the task.

H.6 Example of human evaluation set-up

Here is an example of what you will see during the study. First, you will see a **Claim**, and two pieces of **Evidence**, along with an AI system's predicted **Verdict** and the system's **Certainty** that its prediction is correct.

The **parts of the claim and evidence that are most important to the AI system's certainty are highlighted.** Parts of the Claim are Red, parts of Evidence 1 are Blue, and parts of Evidence 2 are Green.

Underneath, you will see **three alternative explanations for the AI system's certainty**, Explanation A, Explanation B, and Explanation C. The parts of each explanation that refer to the claim and evidence are colour coded in the same way (Claim = Red, Evidence 1 = Blue, Evidence 3 = Green).

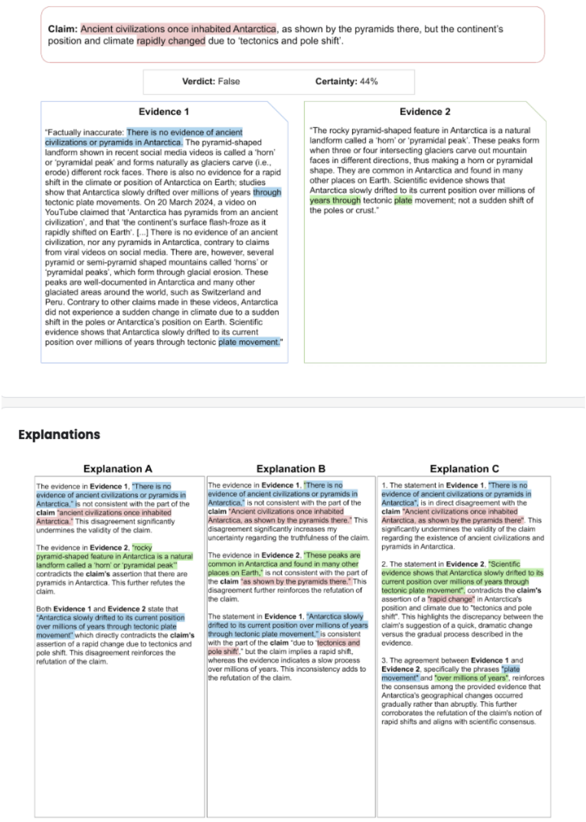

Figure 6: Example of human evaluation set-up

	DRUID		HealthVer	
	Set A	Set B	Set A	Set B
Helpfulness	.016	.079	.003	.013
Consistency	.44	.058	.017	.016
Non-redundancy	.005	.084	.005	.019
Coverage	.494	.113	.018	.027
Overall Quality	.005	.158	.01	.002

Table 5: Inter-rater agreement (Kendall's W) for human evaluation

Your task is to read the claim, evidence, and explanations, and rank each explanation based on five properties.

Now, you can try this example below!

H.7 Inter-rater agreement

In line with similar NLE evaluations carried out by previous studies (e.g., (Atanasova et al., 2020)), interrater agreement (Kendall's W (Kendall and Smith, 1939)) was moderate to low (see Table 5), which we attribute to the relative complexity of the task and individual differences in how the information was perceived.