Appendix A Explanations Generation Algorithm.

Appendix B Study Materials: Prompts, Illustrative Examples, and LLM Evaluation Case.

Table B1: Prompts Used in This Study.

Task	**Prompt**
Enhanced Explanation Generation	*You are assisting users with no ML experience to understand an ML model's predictions regarding fake news detection.* *I will provide you with LIME feature contribution explanations in the format (feature_ name, contribution).* *Convert the explanation into a simple narrative. Please provide a clear, concise, and understandable narrative that clearly explains the model's decision. Ensure the explanation is easy to understand and uses the fewest tokens possible.*
Comprehensive Explanation Generation	*You are assisting users with no ML experience to understand why a news is classified as Fake or Real.* *I will provide you with the news and its classification (Fake/Real).* *Please provide a step-by-step explanation of why the tweet is classified that way. Keep it simple, concise, and easy to understand.*
Automated Evaluation	*You will be provided with an explanation for a detected fake news article. Your task is to rate the explanation based on multiple metrics.* *Evaluation Criteria:* *Fluency (1-5): Assesses whether the explanation follows proper grammar and structural rules. A score of 1 indicates poor fluency, while 5 indicates excellent fluency.* *Informativeness (1-5): Measures how well the explanation provides new information, such as background and additional context. A score of 1 indicates the explanation is not informative, while 5 indicates it is very informative.* *Persuasiveness (1-5): Evaluates whether the explanation is convincing. A score of 1 means the explanation is not persuasive, while 5 means it is highly persuasive.* *Soundness (1-5): Describes whether the explanation is valid and logically sound. A score of 1 indicates it lacks soundness, while 5 indicates it is very sound.* *Evaluation Steps:* *Carefully read the explanation provided for the fake news detection.* *Analyze the explanation based on each of the four metrics: Fluency, Informativeness, Persuasiveness, and Soundness.* *For each metric, assign a score from 1 to 5, based on the explanation's quality in that category.*

Algorithm 1 Explanations Generation

```
1: Input: BERT model, input text $x$, $n$: number of perturbed samples, Interpretable model $g$, LLM, Prompts $T$ and $C$, Explanation Level: level $\in \{\text{Technical}, \text{Enhanced}, \text{Comprehensive}\}$.
2: Output: Explanation at the specified level.
3: Process:
4: Compute $y = BERT(x)$.                                  ▷ BERT Prediction
5: if level = Technical then
6:     Generate a set of perturbed samples $Y = \{x'_i\}_{i=1}^{n}$ from $x$, where each $x'_i$ is a slightly modified version of $x$.
7:     For each $x'_i \in Y$, compute the prediction $BERT(x'_i)$.
8:     Assign a weight $\pi_x(x'_i)$ to each perturbed sample $x'_i \in Y$:

           $\pi_x(x'_i) = \exp\left(\frac{-D(x, x'_i)^2}{\sigma^2}\right)$        (A1)

       where $D(x, x'_i)$ is Euclidean distance between $x$ and $x'$, and $\sigma$ controls the width of the neighborhood.
9:     Fit a local interpretable model $g$ (a linear model) to the perturbed samples $Y$ with weights $\pi_x(x'_i)$ and predictions $BERT(x'_i)$.
10:    Minimize the weighted loss function:

           $\arg\min_{g} \sum_{x'_i \in Y} \pi_x(x'_i) \left(BERT(x'_i) - g(x'_i)\right)^2 + \Omega(g)$        (A2)

       where $\Omega(g)$ is a regularization term to ensure the simplicity of $g$.
11:    Extract feature contributions $\{\beta_j\}_{j=1}^{m}$ from the coefficients of $g$.
12:    Return: Technical Explanation = $\{\beta_j\}_{j=1}^{m}$.
13: end if
14: if level = Enhanced then
15:    Perform steps 5–11 to compute $\{\beta_j\}_{j=1}^{m}$.
16:    Format the contributions $\{\beta_j\}_{j=1}^{m}$ for each feature $x_j$ as:

           $T(x, \beta) = \{\text{Feature } x_j : \text{ Contribution } \beta_j\}_{j=1}^{m}$

17:    Return: Enhanced Explanation = LLM($T(x, \beta)$).
18: end if
19: if level = Comprehensive then
20:    Construct the prompt $C(x, y)$ = Explain why $BERT(x) = y$.
21:    Return: Comprehensive Explanation = LLM(C(x, y)).
22: end if
```

Table B2: Example of Explanation Levels Proposed for a Specific News.

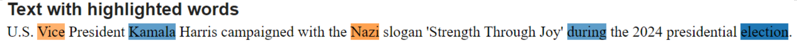
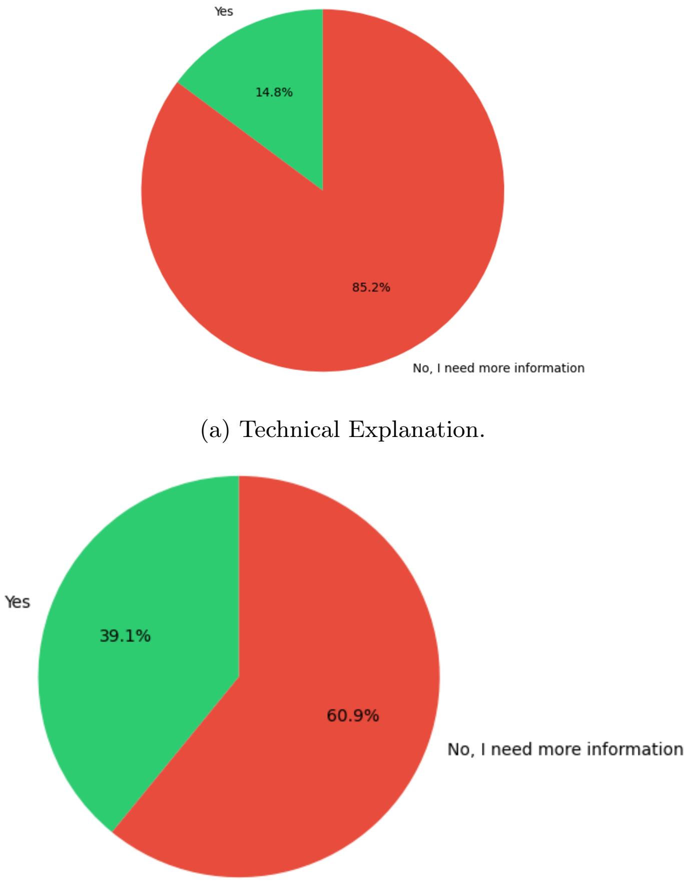
News	U.S. Vice President Kamala Harris campaigned with the Nazi slogan "Strength Through Joy" during the 2024 presidential election. [**label: Fake**]
Technical Explanation	 **Text with highlighted words** U.S. Vice President Kamala Harris campaigned with the Nazi slogan 'Strength Through Joy' during the 2024 presidential election.
Enhanced Explanation	*The model identified several key words to evaluate the news. The words 'election', 'Kamala', and 'during' slightly decreased the chance of the news being fake. However, the presence of 'Nazi' slightly increased the likelihood of the news being fake.*
Comprehensive Explanation	*1. Historical Context: The phrase "Strength Through Joy" was a slogan associated with Nazi propaganda during World War II. It is highly unlikely that a public figure, especially a U.S. Vice President, would knowingly use such a phrase due to its negative historical connotations.* *2. Lack of Credible Sources: If this claim were true, it would be widely reported by reliable news outlets. Since this claim isn't backed by trustworthy evidence, it raises doubts about its authenticity.* *3. Logical Inconsistency: A political figure like Kamala Harris, who is experienced and aware of public scrutiny, would avoid making a decision that would be politically and socially damaging.* *4. Misleading Information Patterns: Fake news often sensationalizes claims to provoke outrage or strong reactions. This statement has hallmarks of misinformation because it combines a high-profile individual with a controversial historical reference.* *5. Verification Tools: When fact-checked using reliable databases or tools, there is no evidence to support this claim.*

Table B3: Example of LLM Evaluation of Quality Criteria Across Different Explanation Depths.

Criteria		**Technical Explanation**	**Enhanced Explanation**	**Comprehensive Explanation**
Fluency	Score	**1.0**	**4.0**	**4.0**
	Explanation	*The explanation is not written in coherent, understandable English. It appears to be a list of words followed by numerical values, which does not form a grammatically correct or meaningful sentence.*	*The explanation is grammatically correct and follows proper structural rules.*	*The explanation is mostly fluent, with good grammar and structure. However, it could benefit from smoother transitions between points and a more varied sentence structure.*
Informativeness	Score	**1.0**	**3.0**	**4.0**
	Explanation	*The explanation does not provide any useful information or context regarding the fake news detection. It lacks background information or details that could help understand the issue.*	*The explanation provides some information about how the words influence the likelihood of the news being fake, but it lacks deeper context about the model used or why these particular words have such an effect.*	*The explanation provides a fairly detailed breakdown of why the news is considered false, including steps like source evaluation, fact-checking, and using logic.*
Persuasiveness	Score	**1.0**	**3.0**	**4.0**
	Explanation	*The explanation is not persuasive at all. Without a coherent argument or context, it is impossible to be convinced by what is presented.*	*The explanation is somewhat convincing but does not provide enough evidence or deeper analysis to make it highly persuasive.*	*The explanation is generally convincing, citing multiple reasons for the news being false. However, it could be more persuasive by providing more concrete evidence or specific historical sources.*
Soundness	Score	**1.0**	**4.0**	**5.0**
	Explanation	*The explanation is not logically sound. It does not present any argument or reasoning that could be evaluated for validity.*	*The explanation is logically sound in describing how the model uses the words to determine the likelihood of the news being fake, but it could be more detailed.*	*The explanation is logically sound and methodically addresses different aspects of the news to determine its falsity. It correctly identifies potential issues in source, factual accuracy, bias, and logical consistency.*

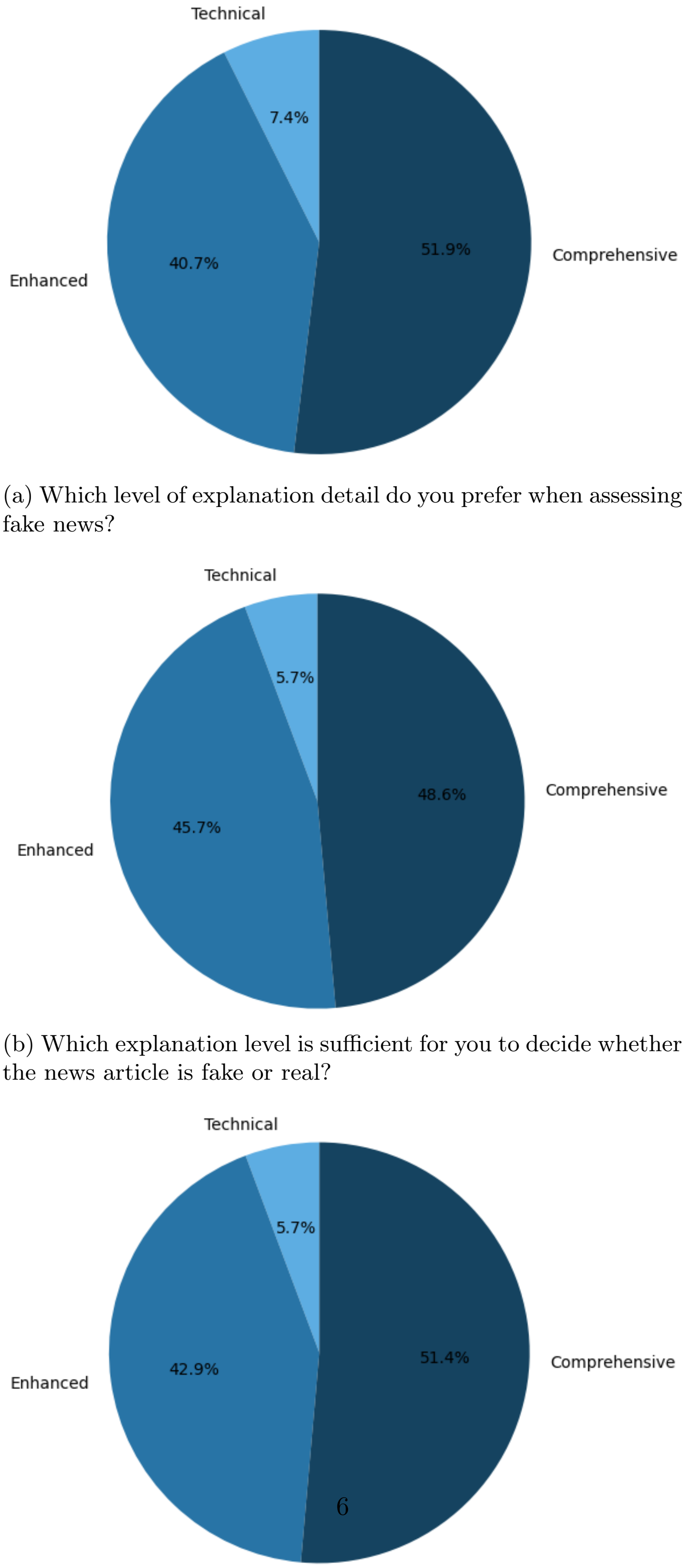
Appendix C Survey Results on Detailed Interactive Explanations



(a) Technical Explanation.

(b) Enhanced Explanation.

Fig. C1: Is this explanation enough?



(a) Which level of explanation detail do you prefer when assessing fake news?

(b) Which explanation level is sufficient for you to decide whether the news article is fake or real?

(c) Which explanation level provided you with enough information to make a decision about the news article's credibility?

Fig. C2: Explanation Level Preferences.

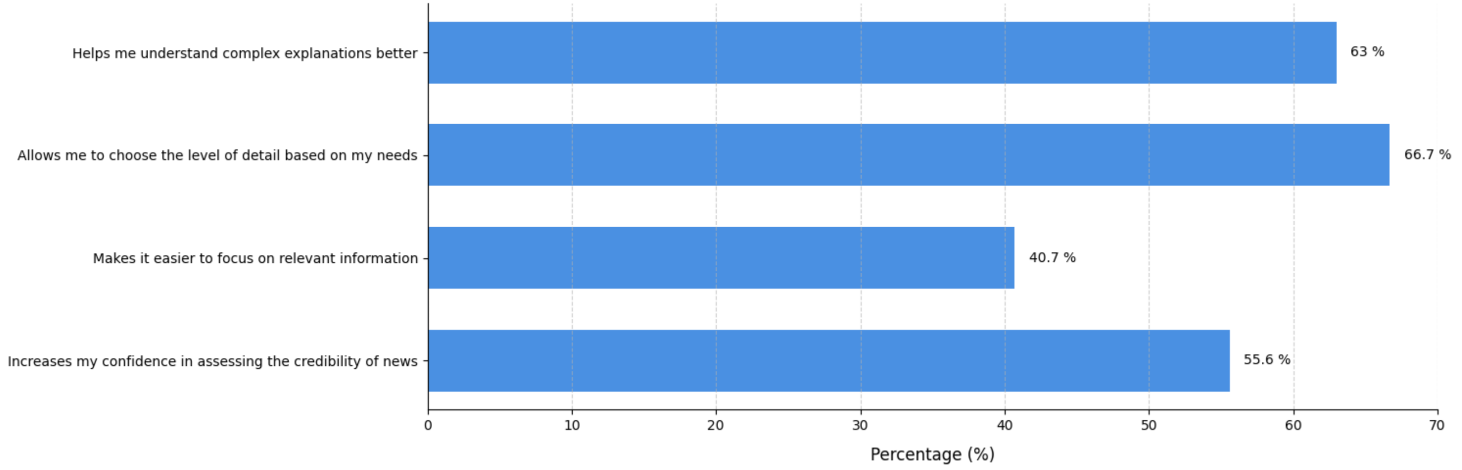


Fig. C3: What do you like the most about receiving different levels of fake news explanations with varying amounts of detail?

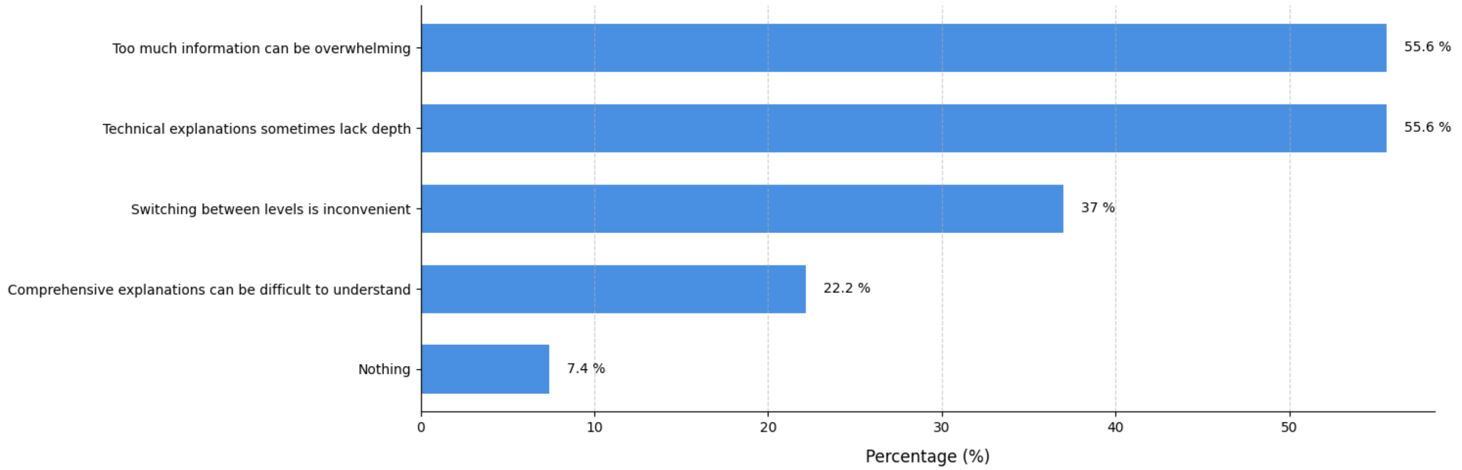


Fig. C4: What do you like the least about receiving different levels of fake news explanations with varying amounts of detail?

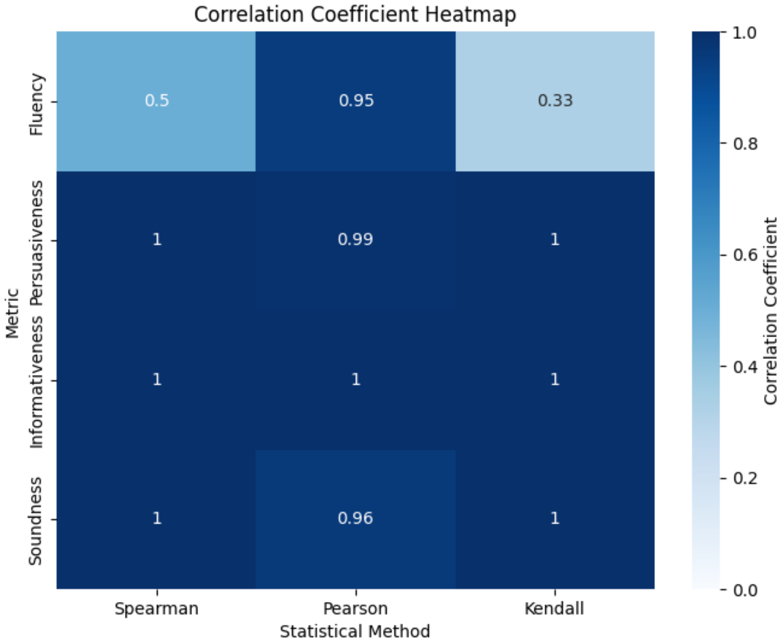


Fig. C5: Correlation heatmap between LLM and Human Evaluations.