Universiteit
Leiden

Master Computer Science

ACORN: Annotating Checkworthy Opinions with Reasoning

Name: Jiahe Xu
Student ID: s4162080
Date: 25/02/2026

Specialisation: Advanced Computing and System

1st supervisor: Michiel van der Meer
2nd supervisor: Suzan Verberne

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In this work, we present ACORN, a multimodal dataset designed to better understand how people decide whether a claim is worth fact-checking and how they explain that decision. Rather than treating check-worthiness as a binary outcome alone, ACORN explicitly captures the reasoning processes underlying human judgments of image–text claims.

Based on the HintsOfTruth corpus, we invited annotators to assess their check-worthiness, we developed a detailed questionnaire that guides annotators to express their reasoning in a structured manner, allowing us to observe their motivations behind their judgments and evidence cues that shape human fact-checking decisions factors that are often implicit or missing in existing datasets.

Using this dataset, we examine how human reasoning compares with that of large language models (LLMs). Instead of focusing only on judgment accuracy, we analyze how different reasoning strategies lead to agreement or disagreement between humans and models.

The main contributions of this work are: (1) a publicly available multimodal dataset with 260 samples annotated by 26 humans with their reasoning for check-worthiness judgments; (2) an analysis of the reasoning strategies differences across LLMs and human annotators; (3) a reproducible research framework for developing interpretable fact-checking systems.

Our experiments highlight a key difference: LLMs act like scanners, sticking to the rules about what might be misleading. Humans, on the other hand, act more as judges of value. They use a shortcut to filter out content they see as just for entertainment, too subjective, or unlikely to cause real-world harm, even if it might be untrue. In addition, we observe that images and annotators' emotional reactions often act as triggers for fact-checking.

Contents

1 Introduction **1**

2 Related Work **3**
2.1 Research Gaps in Multimodal Fact-Checking 3
2.2 From feature extraction to understanding human judgment 3
2.3 Challenges in Multimodal Judgment 3
2.4 Limitations of Existing Data Resources 4

3 Research Questions **4**

4 Methodology **5**
4.1 Data 5
4.1.1 Subset Selection 5
4.1.2 Data Structure 5
4.2 Questionnaire Design 6
4.2.1 Basic Claim properties 6
4.2.2 Reasoning 7
4.2.3 Multimodal Interaction 7
4.2.4 Emotional Influence. 7
4.3 Platform Deployment and Quality Assurance Strategy 8
4.3.1 Platform Selection 8
4.4 Data Collection Procedure 8
4.5 LLM-based Annotation Procedure 9

5 Experiments **10**
5.1 Experiment 1: Validating the Annotation Framework and Data Quality 10
5.1.1 Demographic Characteristics 10
5.1.2 Task Completion Status 11
5.1.3 Reliability of Core Judgments 11
5.1.4 Analysis of Accuracy Differences 12
5.1.5 Whether Complete Reasoning Chains Were Successfully Captured 14
5.1.6 Influence of Images on Check-worthiness Judgments 17
5.1.7 The Role of Emotional Factors in Check-worthiness Judgments 22
5.2 Experiment 2: Human and LLMs Comparisons 24
5.2.1 Runtime and Computational Cost Analysis 25
5.2.2 Human vs Model Reasoning Differences 26

6 Discussion and Limitations **38**
6.1 Discussion 38
6.1.1 Humans Weigh Values, Models Scan for Risks 38
6.1.2 Redefine The Criteria For Check-worthiness 38

6.1.3 The Influence Of Instruction On Human Judgment 39
6.2 Limitations 40
6.2.1 Limited participants 40
6.2.2 Annotation quality 40
6.2.3 The experimental environment is different from the real world environment 40
6.3 Future Work 41
6.3.1 From Risk Scanning to Value-Aware Reasoning 41
6.3.2 Cross-Cultural and Demographic Validation 41
6.3.3 Improve Annotation Quality 42

7 Conclusion **42**

A Keyword Extraction Rules in Experiment 1 **45**
A.1 Text preprocessing 45
A.2 Stopword and task-noise filtering 46
A.3 Manual normalization (lemma mapping) 46
A.4 Lightweight suffix normalization (for out-of-map tokens) 46
A.5 Label standardization used for grouping 47

B Model Versions in Experiment2 **47**

C Prompt Template For LLMs Annotation **47**

D Prompt and Model Parameter For Key words Exaction. **48**
D.1 Parameter Configuration. 48
D.2 Prompt 48

1 Introduction

With the rapid growth of social media platforms, the large-scale spread of misinformation and disinformation has become a serious threat to public trust and social stability [1]. The rise of generative artificial intelligence has further intensified this problem, it enables more diverse and subtle forms of manipulation, such as deepfaked images, manipulated media, and fabricated videos [2]. On social platforms, such content is often more attractive than verified news and therefore more likely to spread widely [3]. In high-risk domains, including public health crises, armed conflicts, and financial markets, misinformation can cause social harm that may even exceed the true information [4].

Automated fact-checking (AFC) has long been studied as a technical approach to addressing the large-scale spread of misinformation [3]. While recent advances in large language models (LLMs) have improved performance on several downstream components of fact-checking, such as claim verification and evidence-based reasoning, the problem of check-worthiness assessment has been recognized as challenging since early work in the field [5].

Check-worthiness assessment acts as a critical filter for fact-checking. It identifies which claims pose a potential social risk or public impact and need to be checked. On the other hand, personal opinions, jokes, or low-risk content that can be safely deprioritized.

Mistakes at this stage can lead to serious consequences. When the system marks low-value content, they waste the limited fact-checking resources. If the system fails to identify high-risk claims, harmful misinformation will continue to spread uncontrollably.

Compared with accuracy judgment, check-worthiness is more subjective and context-dependent. Different social groups, cultural backgrounds and value systems may have different views on what claims are checkworthy. For example, text-image posts on vaccine side effects can be regarded highly worthy of examination in a public health context, but they can be interpreted as personal narratives or satire in other contexts. Similarly, in political or social debates, the same claim may have a significant impact on one group but be largely irrelevant to another [6]. These background and social differences make check-worthiness assessment a complex cognitive task.

Most existing studies on check-worthiness focus on text-only claims. However, real-world misinformation is often multimodal and combines text with images or videos [7]. Multimodal content is more persuasive than text alone and is harder to evaluate. Images can support or weaken textual claims. Even without clear factual mistakes, images can guide audiences toward misleading interpretations.

For this reason, recent work has shifted from text-based fact-checking to Multimodal Fact-Checking (MFC). Researchers often consider LLMs with visual capa-

bilities, such as GPT and Gemini, as promising tools for MFC. However, previous studies show that these models still perform unreliably on key multimodal challenges, including image manipulation detection, out-of-context identification, and truth assessment, especially when text and images do not align [8]. More importantly, these models usually provide predictions without clear explanations. Their reasoning remains opaque, which creates a serious "black-box"problem [9].

This lack of transparency directly affects the reliability of LLMs outputs in the check-worthiness assessment. Reliability here extends beyond predictive accuracy to include the stability and consistency of decision criteria, as well as alignment with human fact-checking reasoning. In practice, model based check-worthiness judgments can determine whether content is flagged, down ranked, or escalated for human review. When models cannot provide interpretable and verifiable justifications, their outputs are difficult to trust and deploy, particularly in large-scale, high-risk information environments.

To address these challenges, this paper introduces ACORN (Annotating Check-worthy Opinions with Reasoning), a multimodal annotation dataset centered on check-worthiness judgments and their underlying reasoning processes. Unlike existing datasets such as FEVER [10], PoliClaim [6], and MFC-Bench [7], which primarily focus on true or false labels, ACORN places human reasoning at the core of its design. Through a structured questionnaire, it explicitly captures annotators' decision motivations, evidential considerations and value related orientations.

Rather than aiming to improve model performance, ACORN provides an interpretable and analyzable foundation for examining how humans form check-worthiness judgments in multimodal contexts. Building on this dataset, this study systematically compares human annotators and state-of-the-art LLMs in terms of judgment consistency, reasoning patterns, and semantic coverage. The findings clarify the strengths and limitations of current models in multimodal check-worthiness assessment and offer empirical support for the development of more transparent and trustworthy automated fact-checking systems.

The main contributions of this paper are threefold.

- **ACORN Dataset.** We introduce ACORN, the first multimodal dataset designed for check-worthiness assessment with explicit human reasoning chains, addressing the lack of structured reasoning annotations in existing fact-checking dataset.

- **Structured Annotation Framework.** We propose an annotation framework that captures check-worthiness decisions and the reasoning, enabling systematic analysis of interpretability in multimodal fact-checking.

- **Human Model Comparative Analysis.** Using ACORN, we conduct a systematic comparison between human annotators and state-of-the-art

LLMs in terms of judgment consistency, reasoning patterns, revealing key limitations of current models in multimodal check-worthiness detection.

2 Related Work

2.1 Research Gaps in Multimodal Fact-Checking

Current research on multimodal fact-checking mainly focuses on deciding whether a claim is true or false. Most studies combine text and images to improve this final judgment. The field has converged on several standard technical tasks. These tasks include image-text consistency checking, image manipulation detection, and context mismatch detection. Datasets such as FEVER, GIF-Check, and MFC-Bench support model training and evaluation for these tasks [3, 11, 7]. In particular, MFC-Bench provides a unified evaluation framework for cross-task comparison [7].

Despite this progress, most existing work treats fact-checking as an outcome-focused problem. These systems aim to decide whether a claim is correct. They pay much less attention to a more practical question, which claims are worth checking first. In real-world settings, fact-checking resources are limited. When systems ignore this prioritization step, they struggle to operate effectively.

2.2 From feature extraction to understanding human judgment

Early work on check-worthiness relied on surface-level features. These features included syntactic patterns and semantic cues used to separate factual from non-factual claims [5]. Later studies introduced social signals. These signals included public attention and topic popularity. This line of work highlighted the role of social impact in real-world fact-checking [12].

More recent studies question whether differences in human judgments should be treated as noise. Research such as AFaCTA shows that these differences often reflect stable variations in values, perspectives, and risk perception [6]. This shift changes how researchers understand reliability. Reliability no longer appears as a purely technical label. Instead, it reflects a decision process shaped by human judgment. However, most of these insights remain grounded in text-only settings.

2.3 Challenges in Multimodal Judgment

This limitation becomes more severe in multimodal contexts. Online misinformation increasingly relies on images and videos rather than text alone [11].

Visual content can shape interpretation quickly and strongly. Misleading images paired with suggestive text can influence judgment even when the text itself seems harmless [3].

Current multimodal benchmarks mainly target detection tasks such as image manipulation. They rarely examine how text and images jointly affect human judgment. When misleading cues arise from the interaction between modalities, single-modality approaches reach their limits. This gap highlights the need to study how multimodal content shapes check-worthiness decisions.

2.4 Limitations of Existing Data Resources

Most of the existing fact-checking datasets focus on the final label. They seldom record how the judgment was formed. Therefore, the reasoning process is often hidden.

Early datasets such as FEVER mainly provided accuracy labels and supporting evidence [10]. However, they offer almost no insights into why a claim is judged to be true or false.

Some newer datasets attempt to capture human interpretations. For example, AFaCTA has collected the fundamental principles of annotation and emphasized differences in values and perception of risk [6]. However, these explanations usually rely on predefined options or short responses. This format is difficult to reflect open-ended and context-dependent human reasoning.

These limitation shows that there is a lack of datasets in this field that link multimodal content with open human reasoning. ACORN aims to address this gap. It records how people interpret their decisions of check-worthiness, not just final labels.

3 Research Questions

Building on the gaps in previous work, this study focuses on developing a multimodal check-worthiness annotation framework that incorporates human reasoning characteristics. To this end, we pose the following core questions:

1. **Designing annotation tasks:** How can annotation tasks be structured to capture both check-worthiness judgments of multimodal claims and the explicit reasoning processes behind them, thereby providing a foundation for explainable modeling?

2. **Diversity across annotators:** How consistent or diverse are check-worthiness judgments and reasoning across annotators, and what mechanisms leads to such differences in multimodal contexts?

3. **Comparisons with LLMs:** How do LLMs perform in generating checkworthiness reasoning compared to human annotators, and what are their strengths and limitations in terms of interpretability?

4. **Challenges in complex scenarios:** What challenges would occur when the content being checked includes synthetic or manipulated content?

By addressing these questions, our study aims to move beyond label only datasets and establish a richer resource that integrates reasoning chains, emotion factors and multimodal alignment. This will not only enhance interpretability but also bring fact-checking systems closer to real-world human judgment.

4 Methodology

4.1 Data

This study constructs the ACORN dataset based on the **HintsOfTruth** benchmark, a multimodal dataset for check-worthiness detection [13]. HintsOfTruth contains paired claims and images, stemming from real-world news examples that have been fact-checked by reliable fact-checker organizations.

The original benchmark focuses on detecting misleading text-image combinations. Our study extends this setting by collecting explicit human judgments and reasoning for check-worthiness assessment.

4.1.1 Subset Selection

We select a subset from the HintsOfTruth for this annotation task. This subset mainly based on two key aspects.

- **Claim type:** The subset includes claims that were initially considered checkworthy and not checkworthy.

- **Domains coverage:** This subset covers the main domains of misinformation, including health, politics, science and entertainment.

This ensures that the final ACORN dataset reflects diverse and realistic misinformation scenarios.

4.1.2 Data Structure

Each data instance contains five core components.

- **Claim:** A short statement written in English.

- **Image:** An image associated with the claim.
- **Check-worthiness Label:** A label (*check-worthy* or *not check-worthy*) annotated by annotators.
- **Reasoning:** Explanations that describe why annotators think the claim is checkworthy or not.
- **Others:** Information about other influencing factors, such as images, emotions.

4.2 Questionnaire Design

To understand the process behind human check-worthiness judgments, we designed a questionnaire that explores many aspects of the annotator's decision process.

4.2.1 Basic Claim properties

At the beginning of the questionniare, we ask annotators to assess some basic properties of each claim. We want to make sure that annotators clearly understand what the claims mean, so that later judgments are reliable.

- **Comprehensibility.** First, we ask whether the claim is clearly understandable or not. If the claim is unclear or difficult to understand, it can be filtered out at an earlier stage and doesn't need to proceed to subsequent processes.
- **Clarity and Factuality.** Annotators indicate whether the claim is factual and verifiable or expresses an opinion or judgment. This follows common practice in check-worthiness research [6], which helps to reduce subjective disagreement across annotators.
- **Novelty and Familiarity.** Annotators report whether the claim is novel or familiar for them. This item is motivated by previous findings [14], which suggest that repeated exposure can reduce perceived risk. The question captures whether familiarity influences the check-worthiness results.
- **Domains.** The questionnaire asks annotators to assign the claim to a topical domain, such as health, politics, or finance. These labels can support later analysis of domain check-worthiness patterns.

4.2.2 Reasoning

The questionnaire requires the annotators to explain how they reach a decision. We focus on two core aspects: potential harm and suspicious evidence. We guide annotators to indicate whether their judgment is driven by concerns about possible social consequences or by doubts about the credibility of the evidence, so that we can avoid to get ambiguous or unexplainable reasons.

4.2.3 Multimodal Interaction

We explicitly capture how images influence check-worthiness judgments in multimodal content with the following questions.

- **Image-Text Alignment.** Annotators assess whether the image supports or misleads the textual claim.
- **Effect of image removal.** The questionnaire asks annotators whether their judgment would change if the image were removed. This question estimates the influence of visual information on check-worthiness assessment.
- **Suspicion of Manipulation.** We ask annotators whether the image appears manipulated or AI-generated. Annotators also report specific visual cues, such as lighting issues or visual artifacts. This design supports later analysis of visual manipulation signals [15].

4.2.4 Emotional Influence.

We also want to find out if emotion can influence annotators' jugdement.

- **Emotion perception.** The questionnaire measures whether the claim evokes emotions such as fear or anger. This design is inspired by emotion taxonomies used in GoEmotions [16]. The goal is to examine whether emotions influence check-worthiness judgment.
- **Bias Self-awareness.** At the end of the questionnaire, annotators will be asked to reflect on whether their personal experiences or feelings might have influenced their judgments. This is helpful for the subsequent analysis of potential sources of bias.

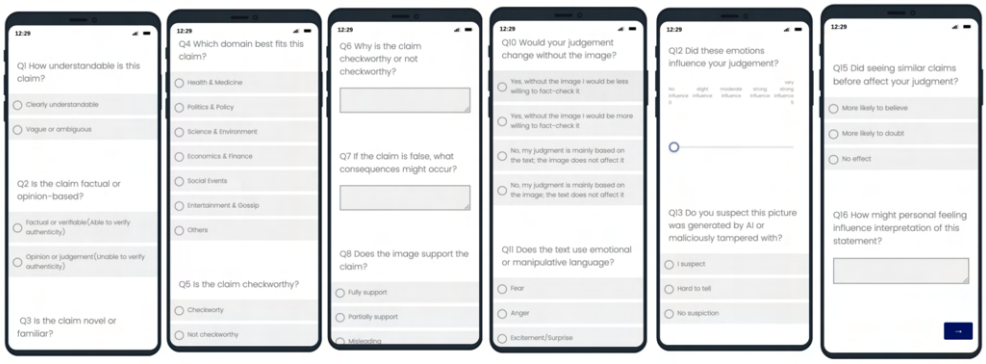

Figure 1: Questionnaire.

4.3 Platform Deployment and Quality Assurance Strategy

4.3.1 Platform Selection

We used Prolific to recruit annotators, and Qualtrics as the questionnaire platform. Qualtrics allowed us to implement a multi-page questionnaire with mandatory responses and conditional branching, and it can also integrate smoothly with Prolific, so it is convenient for participant management.

4.4 Data Collection Procedure

The data collection process consists of four stages.

Annotator Recruitment. We recruited annotators have basic English reading and writing skills to make sure they can understand the claims and give their clear reasons for check-worthiness.

Quality Assurance. To make sure the data we collect is effective and analyzable, we embedded quality control mechanisms within the questionnaire.

- **Instructions.**Before the main annotation task, all annotators have to read an instruction about the task. We demonstrated the full process from properties assessment to reasoning explanation with an example, particularly emphasis on expectations for open-ended responses. So that we can avoid interpretation variance.
- **Attention Checks.** We use explicit instruction-following items (e.g., "Please select banana") to identify inattentive responses.

- **Behavioral checking.** Responses with short completion times or obviously perfunctory reply were flagged for exclusion.

After the data was collected, there were no responses that failed the quality check. All the annotations are finally integrated into a structured JSON format to form the ACORN dataset.

4.5 LLM-based Annotation Procedure

In addition to human annotations, this study incorporates LLMs to examine whether the ACORN framework is also applicable to LLMs, and conduct a systematic comparison of human and LLMs reasoning behaviors.

Model Selection. To compare the performance differences between human annotators and in the multimodal check-worthiness judgment task, we selected five mainstream LLMs: **Gemini-2.5-Pro, GPT-5.1, Claude-Opus-4.5, Llama-4-Maverick** and **Grok-4-fast**. The above models all support generating natural language reasoning from image-text input and explicit judgment results. By using multiple models for the experiment, we can mimic the action of recruiting annotators in Experiment 1, reducing bias from a single model.

All model inferences were conducted between [**Dec 25,2025**] and [**Dec 26,2025**], accessed through the OpenRouter API. Specific model versions are listed in the Appendix B.

Prompt Design. In the experiment, we used a minimally constrained prompt. The prompt only contained constraints on the role setting and output format: the model was designated as an expert survey annotator and was required to output answers in English according to a fixed numbering format (Q1-Q16).

In addition to language and format requirements, the request did not provide any specific definition of check-worthiness, risk dimensions, keyword lists, or reasoning examples. This design aims to avoid explicitly conveying judgment standards to the model through instruction text, thereby reducing prompt induced bias, so that the model's judgment and reasoning reflect more of the model's own reasoning logic, rather than a restatement of the task instructions.

The complete prompt template used for all models is provided in Appendix **??**.

Data Description. Experiment 2 is based on three types of data sources: expert provided gold standard check-worthiness labels, human annotators' judgment results and their reasoning texts, and the reasoning texts from five LLMs on the same set of image-text claims. All analyses are performed on a strictly aligned sample set. The final dataset used in Experiment 2 contains 250 multimodal claim

judgments, each including a check-worthiness label (checkworthy / not checkworthy) and a corresponding free text explanation.

Model Inference Process. The model inference process was organized and managed through the dspy framework, and model calls were completed based on the aggregated API interface provided by OpenRouter. OpenRouter serves as a unified model access layer, used to call various LLMs under consistent request formats, improving efficiency while reducing the impact of interface differences on the experimental process. This setup allows different models to run under the same reasoning environment and calling logic, which can improve the consistency and reproducibility of experimental conditions.

5 Experiments

5.1 Experiment 1: Validating the Annotation Framework and Data Quality

Experiment 1 primarily aimed to verify that our questionnaire could effectively capture the annotators' reasoning process. We mainly rely on two types of evidence for validation.

- First, from the perspective of annotator consistency, evaluate whether the core check-worthiness judgments exhibit stability and agreement across different annotators;

- Secondly, by analyzing the open-ended reasoning text, evaluate whether the questionnaire successfully guides the annotators to provide a structured and complete reasoning process.

5.1.1 Demographic Characteristics

The demographic information of the 26 annotators we recruited is shown in the Table 1 below, sourced from the Prolific platform.

This experiment recruited 26 participants (excluding the ones without submission) for the human annotation task. The sample was mainly female (64%), with an average age of 37.6 years(from 23 to 65). The participants showed some geographic concentration, mainly from South Africa (60%), followed by the UK and other countries. Most of the participants (84%) used English as their native or primary language. The employment status was mainly full-time.

Statistic	Result
Sample size	$N = 26$
Gender ratio (Female:Male)	16:9
Age (mean, range)	$M = 37.6$, 23–65
Geographic distribution (main)	60% South Africa; 24% UK; 16% other
Ethnic composition (main)	60% Black; 36% White; 4% Asian
Primary language is English	84%
Employment status (full-time)	68%

Table 1: Demographic characteristics of human annotators.

Task	Participants	Valid Samples	Validity Rate	Avg. Time (min)	Min Time (min)	Max Time (min)
Task 1	6	6	100%	43.2	26.6	69.6
Task 2	5	5	100%	40.5	33.7	42.6
Task 3	5	5	100%	52.3	14.9	71.9
Task 4	5	5	100%	44.7	31.1	76.7
Task 5	5	5	100%	45.5	25.1	62.9
Overall	**26**	**25**	**100%**	**45.2**	**14.9**	**76.7**

Table 2: Task completion statistics across all annotation tasks.

5.1.2 Task Completion Status

The overall average completion time was 45.2 minutes (ranging from 14.9 to 76.7 minutes). Considering factors such as validity and completion time, we believe that the data quality of this experiment is relatively good.

5.1.3 Reliability of Core Judgments

Metrics and calculation methods. This study uses three metrics to evaluate agreement among annotators: Accuracy, majority agreement and Fleiss' κ.

- **Accuracy**: Accuracy measures the proportion of annotator labels that match the gold labels, where the gold labels are taken from the original HintsOfTruth dataset. This metric provides a direct view of agreement with gold labels.

- **Majority agreement**: The proportion of annotators who select the majority label. This metric reflects consistency among annotators at the sample level.

$$M_i = \frac{\max(n_{\text{check}}, n_{\text{not check}})}{K}$$

$$\bar{W} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

- **Fleiss' κ**: It measures agreement beyond chance and is suitable for multi-annotator settings with imbalanced label distributions.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Results of five test sets. Based on the above evaluation metrics, the results of the five test sets are summarized in Table 3.

Test Set	Accuracy	Maj.	Fleiss' κ	Ncw
1	0.47	0.77	0.05	0.70
2	0.62	0.78	0.18	0.64
3	0.40	0.88	-0.04	0.88
4	0.70	0.82	0.41	0.62
5	0.60	0.80	0.16	0.72
Overall	**0.56**	**0.81**	**0.15**	**0.71**

Table 3: Summary of accuracy, inter-annotator agreement, and proportion of 'not checkworthy' labels across five test sets.

Across test sets, majority agreement rates remain consistently high. This pattern indicates stable annotation behavior, where annotators tend to reach similar judgments for the same claims. In contrast, κ values vary substantially across test sets.

Low κ values appear mainly in test sets with highly imbalanced label distributions. For example, in test set 1, 44 out of 60 instances are labeled as not checkworthy. In these cases, annotators frequently select the same label for many claims. This behavior results in high agreement but low discriminative variation. As a consequence, κ is mechanically reduced.

Test sets with higher κ values show a different pattern. In these sets, both annotators agree with each other and distinguish claims more clearly. These results suggest that, in this study, majority agreement reflects annotator consensus, while κ reflects the discriminative structure of the test set rather than annotation quality.

5.1.4 Analysis of Accuracy Differences

To explain why some test sets had lower accuracy, we compared high accuracy samples with low accuracy samples, analyzing annotators' common reasoning chains across different samples.

Claim	Domain	Gold Label	Accuracy	Representative Reason
A video shows people in New Zealand celebrating the end of COVID-19 restrictions.	Medicine	Checkworthy	0.80	It might lead to more covid infections.
A video shows dozens of Muslims being tortured in Burma.	Politics	Checkworthy	1.00	May incite hatred and retaliation.
A busy street scene with many people walking and festive decorations.	Others	Not checkworthy	1.00	Just a statement, no impact.
A man is walking on a lonely street next to a bus stop.	Others	Not checkworthy	1.00	Just a statement, no impact.
Facebook requires users to send a photocopy of their ID to verify accounts.	Policy	Checkworthy	1.00	It's far in the past, no impact for now.

Table 4: Examples of samples with high annotation accuracy.

High accuracy samples. From these high accuracy samples, it can be observed that when the content is clearly harmless, or obviously involves social impact, public health, annotators tend to make stable judgments, with accuracy reaching 80%–100%. In other words, when claims involve clear potential to cause real-world impact or public misleading issues, annotators are more likely to reach judgments consistent with the gold label.

Claim	Domain	Gold Label	Accuracy	Representative Reason
The youngest parents in the world, aged 12 and 13.	Entertainment	Checkworthy	0.20	Perceived as absurd or implausible and therefore not worth verification.
A video clip shows Bruce Lee playing ping-pong with nunchaku.	Entertainment	Checkworthy	0.20	Interpreted as humorous or created purely for entertainment.
A photograph shows a young Nancy Pelosi as "Miss Lube Rack" of 1959.	Entertainment	Checkworthy	0.17	Considered irrelevant to public interest.
A video shows Osama bin Laden singing Lady Gaga's "Poker Face".	Entertainment	Checkworthy	0.00	Viewed as obviously fabricated and intended for amusement.
A photograph shows a short-tailed weasel.	Others	Checkworthy	0.20	Judged to have no meaningful real-world consequences.

Table 5: Examples of samples with low accuracy.

Low accuracy samples. These error-prone samples from Table 5 show highly consistent characteristics: for entertainment, absurd, or common sense content, annotators often directly terminate the check process and label them as harmless / no impact / like a joke / unimportant. This phenomenon indicates that when making check-worthiness judgments, humans do not always prioritize truthfulness, but are more inclined to make decisions based on personal subjective judgments: content perceived as high risk is scrutinized more strictly, while low risk content is more easily ignored.

5.1.5 Whether Complete Reasoning Chains Were Successfully Captured

In addition to consistency metrics, this experiment also focuses on a core question: whether the questionnaire prompts annotators to provide complete decision

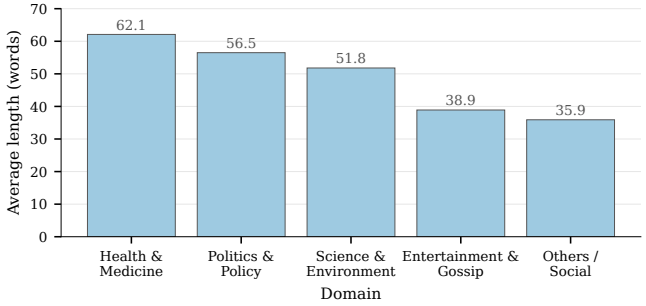

Figure 2: Average reasoning length (in words) across different domains in Experiment 1.

reasoning, rather than brief or superficial explanations. To this end, this study analyzes reasoning texts in three dimensions: reasoning length, diversity of reasoning expressions, and the influence of multiple factors on final judgments.

Reasoning text length We separately counted the length of the reasoning text according to the checkworthy and non-checkworthy labels, as well as by the content domain, to distinguish the reasoning differences in different contexts.

Label	**N**	**Average Length**
Checkworthy	88	61.3
Not checkworthy	172	41.7

Table 6: Average length of free-text reasoning responses by check-worthiness label.

The results from Table 7 can be intuitively observed: when annotators judge a claim as checkworthy, their explanations are usually longer, and they are more willing to provide background information and clearly describe potential risks and consequences; conversely, when a claim is judged as not checkworthy, explanations are often limited to one or two sentences.

From a domain perspective (Figure 1), explanations are usually more detailed in high-risk domains (such as health, public, policy, science, or environmental issues); while in entertainment or daily social content, explanations are significantly shorter. This pattern is consistent with previous observations: annotators impose stricter reasoning requirements on high-risk topics, while dismissing low-risk content more quickly.

Diversity of reasoning expressions. To examine the diversity of free text reasoning used by annotators during the decision making process, this study adopted

the following procedure to extract core keywords from reasoning texts:

- First, we collect all free-text explanations provided by annotators across the five test sets. These texts come from the open-ended reasoning questions, where annotators explain why they consider a claim to be checkworthy or not.

- We then clean and standardize these explanations to make the expressions of different annotators comparable. We convert all text to lowercase and remove punctuation. We also remove common function words and task-related filler terms (e.g., *claim, image, post*), which frequently appear in explanations but do not convey substantive reasoning.

- To reduce variation caused by different word choices, we manually normalize semantically similar expressions into shared forms. For example, words such as *misleading* and *misled* are treated as the same concept, as are variants like *harmful* and *harm*. This step allows us to focus on the underlying ideas expressed by annotators rather than surface wording differences.

- Finally, we group the cleaned explanations by check-worthiness label and examine which reasoning terms appear most frequently in each group. This process highlights common patterns in how annotators justify checkworthy and non-checkworthy judgments.

A complete list of the applied normalization and filtering rules is provided in Appendix A.

Table 7: Top core reasoning terms associated with *checkworthy* judgments

Rank	Word	Frequency	Typical Context Example
1	mislead	12	"could mislead the public"
2	public	11	"public interest", "mislead the public"
3	check	10	"worth checking", "fact check"
4	harm	8	"cause harm", "potentially harmful"
5	interest	7	"in the public interest"
6	health	7	"public health risk"
7	social	7	"social impact", "social media"
8	risk	6	"high risk", "safety risk"
9	panic	6	"cause unnecessary panic"
10	dangerous	5	"potentially dangerous information"

Based on the keyword analysis in Table 7, we observe a tendency in how annotators explain check-worthiness judgments. Annotators repeatedly refer to a small set of shared concerns, such as the potential to mislead, possible public exposure, involvement of high-risk domains, and anticipated real-world consequences.

For claims labeled as checkworthy, these considerations often appear together in annotators' explanations. However, the order and emphasis of these factors vary across individuals and cases. This suggests that check-worthiness judgments are shaped by a combination of risk-related considerations.

Table 8: Top core reasoning terms associated with *not checkworthy* judgments

Rank	Word	Frequency	Typical Context Example
1	irrelevant	66	"the statement is irrelevant"
2	care	66	"no one would care if it were false"
3	personal	54	"just a personal opinion"
4	opinion	54	"just a personal opinion"
5	feel	54	"opinion or feeling"
6	joke	36	"it is clearly a joke"
7	prank	36	"joke or prank"
8	predict	9	"predicts future events"
9	future	9	"predicts future events"
10	verify	9	"unable to verify authenticity"

In contrast, claims judged as not checkworthy are typically justified through a set of exclusion-oriented considerations. Annotators frequently describe such claims as irrelevant, personal, humorous, or unverifiable, emphasizing that these claims are unlikely to attract attention or lead to meaningful real-world consequences.Annotators deprioritize content that is perceived as lacking public relevance, serious impact, or a plausible path to societal harm. This judgment focuses on efficient use of fact-checking resources, independent of whether the claim itself is true or false.

5.1.6 Influence of Images on Check-worthiness Judgments

In multimodal fact-checking scenarios, images are often seen as background support for text. However, visual cues may play an independent role in credibility judgments, especially when image-text inconsistency exists or there is suspicion of manipulation. This section explores a basic question: whether images affect annotators' judgments and in what ways this influence is manifested.

To this end, this study characterizes the role of images from four perspectives:

1. Whether the image is consistent with the text content(Q8);

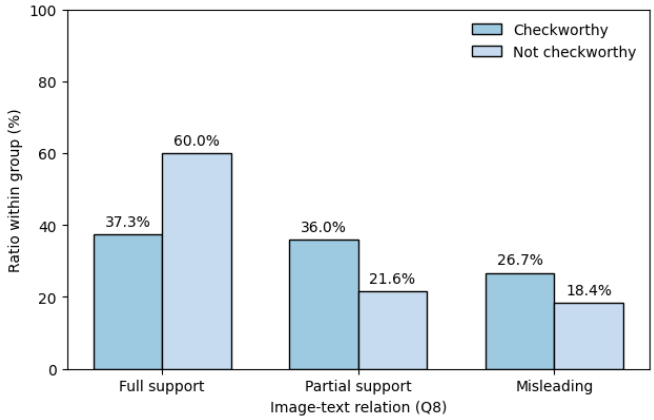

Figure 3: Check-worthiness ratio under different image-text consistency conditions (Q8).

2. Whether removing the image would change the judgment(Q10);
3. Whether annotators suspect the image is AI-generated(Q13);
4. Whether annotators explicitly cite the image as evidence in free-text reasoning(Q14).

Image-Text Consistency and Check-worthiness Triggering We first examine the semantic relationship between the image and the text. Question 8 asks annotators to assess the extent to which the image supports the textual claim, full support, partial support or misleading. Figure 2 presents the distribution of check-worthiness judgments under different image-text consistency conditions.

The results show a simple pattern. When the image clearly supports the text, annotators prefer to see the claim as something that do not need check. Most of these cases are judged as not checkworthy. When the image only partly supports the text, or appears misleading, annotators are much more likely to mark the claim as checkworthy.

This suggests that, a lack of clear alignment between the image and the text makes annotators vigilant, this uncertainty is often enough for them to think that the claim deserves further check.

Image dependency: whether removing the image changes judgment. However, image-text inconsistency does not necessarily mean that the image is actually used in decision-making. To distinguish between"seeing the image" and "relying on the image," the questionnaire directly asks in Q10: if only the text were presented and the image removed, would the judgment change?

Group	changes without image	unchanged without image
Checkworthy	48 (64.0%)	27 (36.0%)
Not checkworthy	85 (45.9%)	100 (54.1%)

Table 9: Distribution of judgment changes after image removal (Q10) within checkworthy and not checkworthy decisions (Q5).

The results reveal a clear difference between checkworthy and not checkworthy judgments. Among claims labeled as checkworthy, a majority of judgments (64.0%) change when the image is removed. This indicates that, for checkworthy claims, annotators often rely on visual information to form their judgments, and the textual content alone is insufficient to support the decision.

In contrast, not checkworthy judgments show the opposite tendency. More than half of these judgments (54.1%) remain unchanged after image removal, suggesting that annotators primarily base their decisions on the textual content. In such cases, the image does not provide decisive information, and the claim is perceived as clear, low-risk, or irrelevant even without additional visual context.

Taken together, these findings suggest that dependence on visual information is more characteristic of checkworthy judgments. When annotators cannot make decisions based on text alone, the resulting uncertainty makes the claim more likely to be considered worthy of further check.

A closer look at the image-text relationship further clarifies this pattern. Among cases where annotators report that removing the image would change their judgment, misleading image-text relations are substantially more common (28.6%) than in cases where the judgment remains unchanged (12.6%). Conversely, full image-text support is more frequent among judgments that do not change after image removal (59.8% vs. 47.4%).

This indicates that image dependency is closely tied to image-text inconsistency. When the image and text do not clearly align, annotators are more likely to rely on the image to interpret the claim. As a result, removing the image in such cases directly destabilizes the judgment.

Uncertainty about AI manipulation and verification motivation. To examine whether image authenticity and potential perception of AI manipulation affect judgments, Q13 gives three options, suspicion, hard to tell, and no suspicion. We want to analyzes how this uncertainty influences check-worthiness judgments.

The results show that the relationship between whether an image is checkworthy and whether it is suspected of being AI-generated is not linear; that is, greater suspicion does not necessarily lead to a greater desire for verification. In both checkworthy and not checkworthy groups, the distribution of suspicion op-

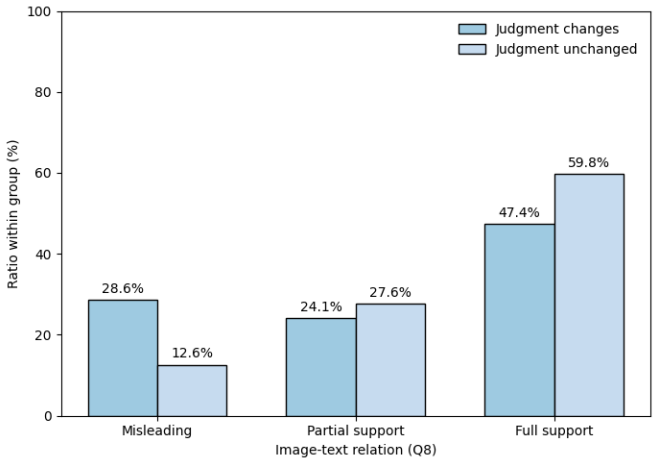

Figure 4: Impact of Image–Text Support on Judgment Stability.

tions is not significantly different, with no suspicion being the most frequent choice in both. However, in the checkworthy group, the proportion of suspect and hard to tell options is somewhat higher, suggesting that when annotators suspect an image is AI-generated, their willingness to check increases, but it is not the only decisive factor.

Active reference to images in reasoning texts. Finally, this study analyzes free text reasoning to examine whether images truly enter annotators' decision logic. According to preset criteria, if the reasoning text contains image-related keywords such as *image, photo, video, shown*, it is considered that annotators actively reference the image in their explanation.

Image Mentioned in Reasoning	Samples	Checkworthy Ratio
No	188	20.7%
Yes	72	68.1%

Table 10: Checkworthy Ratio Based on Image Mention in Reasoning

The result is very significant: when reasoning texts explicitly mention image related text, 68.1% of samples are judged as checkworthy, more than three times the proportion in cases where images are not mentioned. This finding verifies the previous analysis, indicating that once images are included as argumentative evidence, they have a significant impact on check-worthiness judgments.

Combining the results of Q8, Q10, Q13, and reasoning texts, the role of images can be understood as a clear influence chain: images first change annotators'

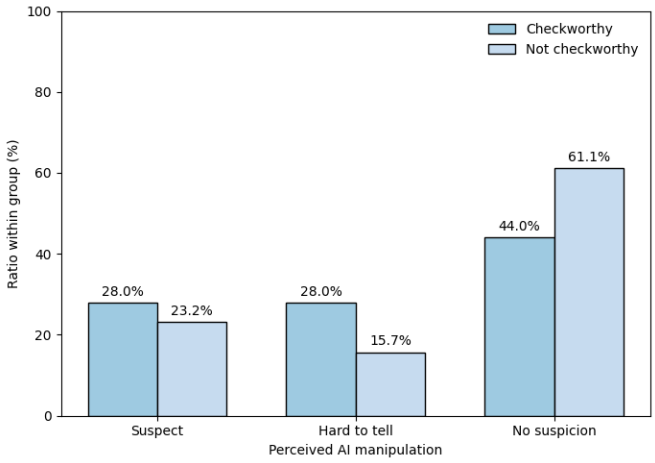

Figure 5: Check-worthiness ratio under different levels of uncertainty about AI-generated or manipulated images (Q13).

intuition about "whether this information is reliable/safe" (sense of risk, uncertainty), then change their tendency of "whether to check," and finally manifest in the choice of checkworthy.

The specific chain is as follows: when annotators believe the image fully supports the text (Q8), they usually feel the information is relatively reliable, therefore less likely to treat it as checkworthy; but once the image only partially supports or is even considered potentially misleading, annotators are more likely to feel "there might be a problem here," and thus more inclined to judge it as checkworthy. At the same time, regarding AI generation or tampering (Q13), the results show it is not "more suspicion leads to more check," suspecting AI generation does indeed increase the likelihood of check to some extent, but it's not the only influencing factor.Finally, Q10 and reasoning texts further illustrate "whether images truly participate in decision making": when annotators explicitly state that removing the image would change their judgment (Q10), or actively cite the image as basis in their reasoning (e.g., mentioning image/photo/shown, etc.), the checkworthy proportion increases significantly. This means that images do not always play a role, but once annotators treat images as key evidence, their promotion of final checkworthy judgments becomes very obvious.

In short, if the image is inconsistency or uncertainty, annotators are more willing to check, so more likely to choose checkworthy; and "removing the image would change judgment / mentioning image in reasoning" indicates images truly enter the core decision process, amplifying this promotion.

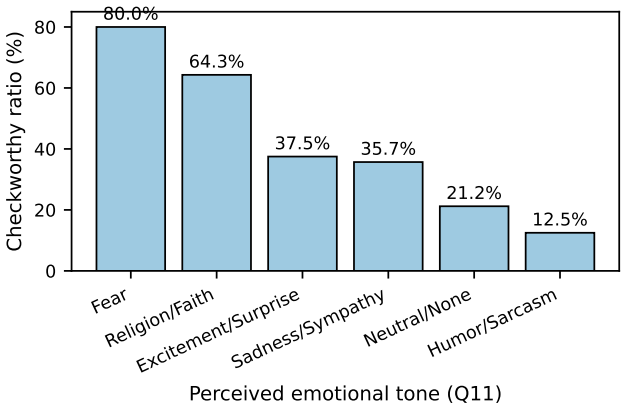

Figure 6: Check-worthiness ratio under different perceived emotional tones (Q11).

5.1.7 The Role of Emotional Factors in Check-worthiness Judgments

In addition to visual cues, emotional responses are also important cognitive factors influencing judgement. Therefore, we analyze the impact of emotions on core decisions from three perspectives: the category of explicit emotions, the intensity of self-perceived emotional influence, and the use of emotional language in reasoning texts.

Check-worthiness Tendencies Across Different Emotion Types. To investigate the influence of emotions contained in the content on annotators' judgments, Question 11 directly asked annotators to identify the emotions present in the content. The Table below shows the proportion of content judged as checkworthy under different emotion categories.

The results indicate that the emotional tone of a claim systematically shapes annotators' check-worthiness judgments, and that this effect is closely related to perceived potential risk. Fear associated claims exhibit the highest proportion of checkworthy judgments (80.0%), followed by content related to religion or faith (64.3%). This suggests that when a claim evokes strong emotions linked to threat, anxiety, or value based conflict, annotators tend to become more vigilant and are more likely to consider the claim checkworthy, even before establishing its factual accuracy. In contrast, emotionally neutral content shows a lower checkworthy ratio (21.2%), while humor or sarcasm corresponds to the lowest proportion overall (12.5%), indicating that content perceived as non serious or emotionally low-stakes is less likely to trigger fact-checking concerns.

These findings suggest that the check-worthiness judgment is driven by the relevant social influence implied by different emotions. Emotions that indicate potential harm, panic or public consequences tend to heighten perceived risks,

thereby increasing the likelihood that claims are considered checkworthy. On the contrary, emotions related to entertainment, satire or emotional neutrality often dilute the perceived impact, leading annotators to have a reduced willingness to verify them.

Self-reported emotional influence and check-worthiness judgments. To further distinguish between the presence of emotions and their influence intensity, the questionnaire asked annotators to rate on a 0–5 scale the degree to which emotions implied in the content influenced their judgment. The results show that as the emotional influence score increases, the proportion judged as checkworthy generally shows an upward trend, especially in the high score range (4–5).

Emotion Influence Score	Samples	Checkworthy Ratio
0	146	24.0%
1	48	33.3%
2	27	29.6%
3	21	33.3%
4	13	46.2%
5	5	60.0%

Table 11: Check-worthiness ratios across self-reported emotion influence scores (Q12).

The Q12 results show a clear positive association between emotional impact scores and check-worthiness judgments. As the reported emotion impact increases, the proportion of claims labeled as checkworthy rises consistently, from 24.0% at score 0 to 46.2% at score 4 and 60.0% at score 5. This monotonic trend indicates that claims perceived as having stronger emotional impact are more likely to be judged as warranting verification, whereas emotionally neutral content is less frequently considered checkworthy.

Emotional language as risk clues. Finally, to capture emotional cues in reasoning texts, this study used the publicly available emotion dictionary NRC Emotion Lexicon to perform word level scanning of free text explanations. Specifically, all reasoning texts were first converted to lowercase and punctuation normalized, then each term was matched with the NRC dictionary to identify its corresponding emotion label (e.g., anger, fear, sadness, joy). If at least one emotion-related word appears in the text, the variable EmotionMention is set to 1; otherwise set to 0.

The results show that samples containing emotional words have a significantly higher proportion judged as checkworthy than samples without emotional words

Emotion Words Present in Reasoning	Samples	Checkworthy Ratio
Yes	70	48.6%
No	190	21.6%

Table 12: Check-worthiness ratios stratified by the presence of emotion-related words in annotators' reasoning.

(48.6% vs. 21.6%). This indicates that emotions not only influence the judgment process but are also explicitly embedded in the reasoning structure through linguistic forms.

Combining Q11, Q12, and the analysis of emotion words in reasoning texts, it can be seen that emotions tangibly change how annotators process information: emotions first affect whether annotators feel the content has risk, needs to be taken seriously, then affect their willingness to check.

The data reveal consistent associations between emotional cues and check-worthiness across three levels of analysis:

- **Emotion Type:** Claims evoking high-risk emotions (e.g., fear, conflict) correlate with higher checkworthy rates, whereas humorous or sarcastic content is more frequently dismissed as low-priority "entertainment."
- **Self-Reported emotion:** Higher self-reported emotional impact scores (4–5) correspond to substantially elevated checkworthy ratios (up to 60.0%), indicating that stronger perceived emotional salience increases annotators' vigilance.
- **Reasoning Text:** Explanations containing emotion-related words show a significantly higher checkworthy proportion (43.5% vs. 25.7%), suggesting that emotions not only influence decisions implicitly but also surface explicitly in justifications.

Together, these patterns indicate that emotional cues are systematically integrated into human check-worthiness judgments.

5.2 Experiment 2: Human and LLMs Comparisons

In Experiment 1, we validated that the ACORN annotation framework can capture human judgments about the check-worthiness of multimodal claims and revealed how image and emotional factors influence this decision process. However, only human annotation results is insufficient to evaluate whether this framework is broadly applicable for LLMs. Therefore, Experiment 2 further introduces

LLMs to compare humans and models on the same set of image-text samples, focusing on the differences between the two in terms of check-worthiness judgment results and reasoning expression styles.

This experiment no longer only focuses on whether the model's judgment is correct, but further explores a deeper question: when both humans and models are asked to explain why a claim is or is not checkworthy, do they show systematic differences in their reasoning logic.

5.2.1 Runtime and Computational Cost Analysis

To evaluate the runtime efficiency and resource overhead of different LLMs in the experiment, we separately counted (1) the total reasoning time (seconds) of each model across the five test sets, (2) the request volume, average token count, and inference cost based on the OpenRouter aggregated API. All models ran under unified reasoning processes and output format constraints (single-round inference, one output per sample), therefore differences in time and cost between different models are mainly attributed to differences in the model's own reasoning efficiency, generation length, and pricing strategies.

Model	**Test1**	**Test2**	**Test3**	**Test4**	**Test5**	**Model Avg.**
Gemini-2.5-Pro	318.08	328.62	269.97	315.52	293.71	305.18
GPT-5.1	218.10	261.82	181.63	275.68	229.07	233.26
Claude-Opus-4.5	158.49	151.71	129.65	154.34	141.88	147.21
Llama-4-Maverick	131.96	107.49	83.30	126.78	97.23	109.35
Grok-4-fast	105.86	115.77	92.50	107.59	97.08	103.76
Test Set Avg.	**186.50**	**193.08**	**151.41**	**195.98**	**171.79**	**179.75**

Table 13: Total inference time per model across test sets (in seconds).

Model	**Requests**	**Total Cost (USD)**	**Avg. Cost /Request**	**Avg. Input Tokens**	**Avg. Output Tokens**
google/gemini-2.5	35	1.066	0.0305	2460	2810
anthropic/claude-4.5	30	0.726	0.0242	1670	625
openai/gpt-5.1	38	0.434	0.0114	1480	860
llama-4-maverick	37	0.0335	0.0009	1975	825
grok-4-fast	40	0.0338	0.00085	1330	1186

Table 14: Cost and token statistics collected from the OpenRouter console.

Combining Table 13 and 14, cost differences are related to pricing strategies on one hand, and generation length on the other. Taking the average output tokens

as an example, Gemini-2.5-Pro has the highest average output tokens (2810), also corresponding to the highest average single cost (0.0305 USD/request) and highest total cost (1.066 USD); relatively speaking, the average single cost for Llama-4-Maverick and Grok-4-fast are both below 0.001 USD/request, with total costs around the 0.03 USD level, showing significantly lower inference overhead. Claude-Opus-4.5 has a lower average output token count (625), but still a high single cost (0.0242 USD/request), suggesting that its cost is mainly influenced by the pricing structure, not just generation length.

5.2.2 Human vs Model Reasoning Differences

Table 15 summarizes overall performance between human annotators and models.

Decision Maker	N	Acc.	Prec. (CW)	Rec. (CW)	F1 (CW)	TP	FP	TN	FN
Human	260	0.56	0.80	0.37	0.51	60	15	84	101
Gemini-2.5-Pro	50	0.88	0.93	0.87	0.90	27	2	17	4
GPT-5.1	50	0.86	0.96	0.81	0.88	25	1	18	6
Claude-Opus-4.5	50	0.88	0.90	0.90	0.90	28	3	16	3
Llama-4-Maverick	50	0.90	0.91	0.93	0.92	29	3	16	2
Grok-4-fast	50	0.84	0.87	0.87	0.87	27	4	15	4

Table 15: Detailed performance metrics and confusion-matrix counts for check-worthiness classification. CW and NCW denote the *checkworthy* and *not check-worthy*.

From Table 15, human annotators reach an overall accuracy of 0.56 on this task. Their precision for the checkworthy class is relatively high (Precision = 0.80). However, their recall remains low (Recall = 0.37).

This pattern shows that human judgments are not always agree with the gold standard. Human annotators often show lower willing to check the claims that are labeled as checkworthy in the gold standard. The confusion matrix reflects this tendency, with more missed checkworthy cases (FN = 101) than not checkworthy cases (FP = 15).

In contrast, the outputs of the five LLMs align more closely with the gold standard. All models achieve higher overall accuracy (0.84-0.90). The models also show much higher recall (Recall = 0.81–0.93). At the same time, the models maintain strong alignment on claims that are labeled as not checkworthy. Overall, the model aligns better with the gold standard, especially in check-worthiness label.

Core Reasoning Differences and Actual Cases. To further explore the source of the differences, this experiment categorizes samples based on whether

humans and models agree with the gold label, by checkworthy status, content domain, and human and model answer status. The sample number distribution for each category is shown in the table below.

Ground Truth	Only Human Right	Only Model Right	Both Right	Both Wrong	Total
Checkworthy	0	18	10	3	31
Not Checkworthy	1	0	17	1	19

Table 16: Distribution of human and model judgments compared with the gold standard.

Domain (Q4)	Only Human Right	Only Model Right	Both Right	Both Wrong	Total
Entertainment & Gossip	0	7	4	1	12
Health & Medicine	1	0	3	1	5
Others	0	4	4	2	10
Politics & Policy	0	2	6	0	8
Science & Environment	0	4	3	0	7
Social Events	0	1	7	0	8
Total	1	18	27	4	50

Table 17: Distribution of judgment outcomes across different content domains.

From the above two tables, it can be observed that humans and models are not always consistent in check-worthiness judgments. Overall, when the gold label is not checkworthy, the two almost always reach consistent judgments, indicating that in low-risk situations, humans and models have a highly consistent understanding of check-worthiness. In contrast, when the gold label is checkworthy, disagreement increases significantly, mainly manifested as the model judging correctly while the human judges incorrectly(compared with gold standard). This phenomenon indicates that human annotators tend to underestimate the necessity for check-worthiness in some situations, resulting in systematic differences.

Further combining the content domain distribution, it can be found that these samples where only the model is correct are mainly concentrated in entertainment, science and environment, and other boundary-blurring content types, while in domains with clearer public impact like politics and social events, humans and models are more likely to form consistent judgments. This suggests that the difference between humans and models does not stem from disagreement over the goal of fact-checking, but more from different ways of weighing potential social impact.

To analyze on which content humans and models can reach agreement, we first capture core reasoning phrases from free texts in samples.

Our approach uses a small set of logic categories to summarize how people justify check-worthiness decisions.

- A LLM proposes 10-12 short and human-readable categories based on the dataset. We review the list and make sure it covers both checkworthy and not-checkworthy cases.

- Next, each reasoning text is mapped to exactly one logic category from this fixed set. The model extracts a short evidence phrase of two to five words copied from the original text. These phrases serve as compact summaries of the main reasoning signal.

- Finally, the study reports results separately for humans and models.

The analysis counts category frequencies and computes percentages. Table 18 shows the final distributions.

Details of model parameters and prompts can be seen in Appendix D.

Reasoning Type (Logic Category)	**GL**	**Human (n)**	**Human (%)**	**Model (n)**	**Model (%)**
Could mislead the public	CW	11	7.90%	85	63.00%
Totally irrelevant content	NCW	45	32.40%	14	10.40%
Just personal feelings	NCW	31	22.30%	4	3.00%
Might cause unnecessary panic	CW	7	5.00%	19	14.10%
Sounds like a joke or satire	NCW	5	3.60%	1	0.70%
Is purely a descriptive statement	NCW	0	0.00%	5	3.70%
Might damage someone's reputation	CW	3	2.20%	1	0.70%
Could promote hate or violence	CW	2	1.40%	0	0.00%
Hard to verify	CW	1	0.70%	0	0.00%
Other / Unclear	Both	34	24.50%	6	4.40%
Total	–	139	100%	135	100%

Table 18: Frequency of key reasoning phrases used by human annotators and models.

Humans and the model are in agreement. In samples where the gold label is checkworthy, humans and models actually share a very intuitive set of expressions: both sides use phrases like *"could mislead the public"*, *"might cause unnecessary panic"*, *"might damage someone's reputation"*, *"could promote hate or violence"* to ground the reason for checkworthy on possible specific social consequences. This shows that when the risk direction of the content is clear enough, there is no obvious disagreement between humans and models in reason types, both tending to use harm–consequence to determine check. Additionally, it can be seen that models highly concentrate on using *"could mislead the public"* in such correct samples, almost using it as the main core reason for checkworthy; humans also use it, but the frequency is lower. In other words, when humans judge something as checkworthy, their expressions are more diverse, such as falling into

"People would get their hopes up for nothing" in the "other" category; while models are more inclined to write social risk reasons more explicitly and stably, which is also one reason why models can better identify samples checkworthy.

Case 1. Claim: *Mexican citizens are required to have government-issued photo id cards in order to vote in federal elections*

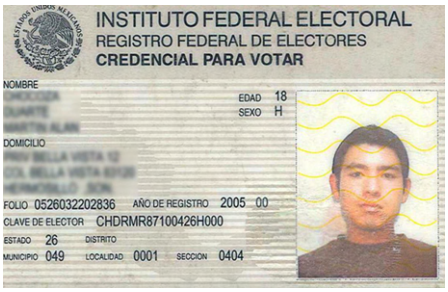

Figure 7: Example1.

- **Gold Standard:** checkworthy.
- **Human:** 4 correct, 1 wrong.
- **Model:** all correct.
- **Reason:** Humans thought it needed verification because it could mislead the public; models gave similar reasons.

In samples where the gold label is not checkworthy, humans and models also have common expression: both sides use phrases like *"irrelevant content"*, *"just personal feelings"*, *"sounds like a joke or satire"*, *"a descriptive statement"* to express the reason for "not checkworthy" as "lack of public impact/lack of verifiable factual core/more like expressing attitude or joke." This shows that both parties can classify clearly harmless content as not checkworthy. Additionally, humans in not checkworthy samples more often write reasons as irrelevant, subjective, joke relatively more subjective judgment words; models also write this way, but the proportion is lower.

Case 2. Claim: *Hundreds of people walking on a very busy street with stars hanging from buildings* (with an image of a busy street).

Figure 8: Example2.

- **Gold Standard:** not checkworthy.
- **Human:** all correct.
- **Model:** all correct.
- **Reason:** Humans thought this was just a statement, irrelevant; models also thought it was just an ordinary statement.

Only Humans Are Wrong. Next, we analyze the Only Human Wrong situation, which has the largest sample size, so it best reflects the systematic differences in reasoning strategies between the two.

We similarly extract core reasoning phrases from free reasoning texts to analyze why humans make large-scale misjudgments while models can make correct judgments.

Reasoning Category	Human (n)	Human (%)	Model (n)	Model (%)
Sounds like a joke or satire	26	27.4%	6	6.7%
Totally irrelevant content	20	21.1%	0	0.0%
Just personal feelings	17	17.9%	1	1.1%
Hard to verify	12	12.6%	10	11.1%
Other / Unclear	6	6.3%	0	0.0%
Could mislead the public	4	4.2%	44	48.9%
Might cause unnecessary panic	3	3.2%	12	13.3%
Might damage someone's reputation	2	2.1%	6	6.7%
Seems scientifically implausible	1	1.1%	8	8.9%
Could promote hate or violence	1	1.1%	1	1.1%
Is a well-known fact	2	2.1%	0	0.0%
Basic common knowledge	1	1.1%	0	0.0%
Is purely a descriptive statement	0	0.0%	2	2.2%

Table 19: Distribution of reasoning categories in samples where only human annotators made incorrect judgments while models were correct.

This table clearly shows that in samples where only humans judge incorrectly, there are significant differences in the focus of reasoning between humans and models: Human annotators mainly rely on: *"Sounds like a joke or satire"*, *"Totally irrelevant content"*, *"Just personal feelings"* three types of reasoning, accounting for **66.4%** of reasoning texts. All of these reflect an attitude of the annotator towards this type of content, they treat the content as entertaining, irrelevant, or subjective, thus negating its verification necessity.

Models' reasoning is highly concentrated on *"Could mislead the public"* (48.9%), supplemented by risk oriented categories like *"Might cause unnecessary panic"*, *"Seems scientifically implausible"*, etc. This indicates that models are more inclined to trigger verification judgments from the perspective of potential social deception and public risk.

Case 3: A photograph shared online in may 2021 shows a "super moon" behind the temple of poseidon at cape soundon, greece.

Figure 9: Example 3.

- **Gold Standard:** checkworthy.
- **Human:** 4 wrong, 2 correct.
- **Model:** 4 correct, 1 wrong.
- **Reason:** Humans thought the content would not cause substantive negative impact; those who were correct also did so out of personal interest, finding it interesting to check. Models thought the image looked unusual, possibly fabricated, and needed verification.

Case 4: A photograph shows a "unicorn puppy" with a tail growing from his forehead.

Figure 10: Example 4.

- **Gold Standard:** checkworthy.
- **Human:** 3 wrong, 2 correct.
- **Model:** all correct.
- **Reason:** Humans thought this was obviously fake and didn't need verification; some people wanted to verify out of personal interest. Models thought this was a biological phenomenon, the image could be fake, so worth verifying.

Case 5: The photograph shows a young nancy Pelosi as the "miss lube rack" of 1959.

Figure 11: Example 5.

- **Human:** 5 wrong, 1 correct.
- **Model:** all correct.

- **Reason:** Most humans thought this was entertaining content, would not affect our lives, but overlooked possible trouble for the person involved; while models classified it as a political issue, thinking it could mislead the public about her past, thus having some political impact.

From these cases, it can be seen that when human annotators judge whether a piece of content is checkworthy, they are often not thinking about whether it is true or false, but whether it is worth checking. When faced with content that looks strange, exaggerated, or even obviously untrue, humans are more inclined to judge based on intuition whether it will have a real-world impact; if a piece of information is considered harmless, interesting, or just entertainment, people often assume it will not bring actual consequences, thus giving up further verification. In contrast, models do not behave like humans; they focus more on whether there are factual anomalies or potential deception. This difference leads to humans showing a lower willingness to check some low-risk but untrue content, whereas models are more likely to trigger checkworthy judgments. In other words, humans are more like making judgments based on experience and personal feelings, while models are executing a more systematic risk checking mechanism.

Both Humans and Models Are Wrong. **Case 6:** A girl playing in the sink with toys and some potentially dangerous household chemicals.

Figure 12: Example 6.

- **Gold Standard:** not checkworthy.
- **Human:** 4 wrong, 2 correct.
- **Model:** 4 wrong, 1 correct.
- **Reason:** Both humans and models thought these chemicals could potentially harm the child; also some samples thought this was just a daily descriptive picture, would not affect the public, so not worth verifying.

Case 7: A photograph shows a short-tailed weasel.

Figure 13: Example 7.

- **Gold Standard:** checkworthy.
- **Human:** 4 wrong, 1 correct.
- **Model:** 3 wrong, 2 correct.
- **Reason:** Most humans and models thought that it would not have practical impact, no need to check; some thought that it was an AI-generated image, false information, necessary to check.

Only Models Are Wrong. Case 8 A photo shows a brothel in Italy and Valencia (Spain) where 86 ... quarantined for 14 days due to the novel coronavirus epidemic.

Figure 14: Example 8.

- **Gold Standard:** not checkworthy.
- **Reason:** Humans thought it was a joke, no need to check; models thought it was related to COVID-19, had spreading risk, could cause panic, needed verification.

From these cases, it can be seen that in actual check-worthiness judgment cases, there are indeed some gray areas, leading to deviations in judgment results, but these disagreements should not be simply understood as right or wrong. On one hand, in the scene of a child with chemicals and the animal species identification picture, both humans and models thought there was potential harm risk or authenticity suspicion, thus more likely to give a judgment of checkworthy, and the gold labels for these samples themselves are not checkworthy. It reflects that the boundary of check-worthiness is not always clear: the same content may be interpreted as having only daily descriptions, lacking clear and verifiable claims, or being risky or involving factual identification, thereby leading to different judgment results.

Similarly, in the pandemic case where models lean more towards worth checking, humans may combine real-world situations to treat it as a joke, think it's definitely fake, will not cause negative impact; models, because the content may bring spreading risk or impact, tend to check. Overall, these cases show that humans subjectively judge the content's impact on reality, while models are more sensitive to potential risk; so when the judgment standard of the task itself has gray areas, disagreement between humans and models is also more likely to occur, so it should not all be attributed to right or wrong problems, but rather the nature of the content itself.

Influence of Other Factors. Based on free reasoning texts from humans and models, we analyzed the core reasoning logic differences between humans and

models. To better explain these differences, we further combine other collected information (Q8-Q13) to analyze the disagreement between humans and models in check-worthiness judgments from the perspective of reasoning chains.

- Q8 Visually misleading: Does the image support the claim?
- Q10 Image dependence: Would your judgement change without the image?
- Q11 Emotion detection: Does the text use emotional language?
- Q13 Authenticity suspicion: Do you suspect this picture was AI-generated?

To quantify qualitative characteristics in the reasoning process, we performed structured coding on the responses of Q8-Q13 in our questionnaire. For each indicator, we defined positive determination rules; Responses that did not meet the rules were uniformly coded as 0.

Dimension	**Question Reference**	**Coding Rule (Value = 1)**
Q8 Visual Misleading	Does the image support the claim?	The *Misleading* option is explicitly selected.
Q10 Image Dependence	Would the judgment change without the image?	The annotator explicitly answers *Yes*.
Q11 Emotion Detection	Does the text use emotional language?	Any response other than *Neutral* is selected.
Q12 Emotion Influence	Did these emotions influence your judgment?	The annotator explicitly indicates that emotions had an influence.
Q13 Authenticity Suspicion	Do you suspect the image was AI-generated?	The annotator explicitly selects an option indicating suspicion.

Table 20: Coding rules for intermediate reasoning factors used in the analysis.

For all percentage type indicators, the denominator is the number of samples in the corresponding group with valid mappable answers for that question, the numerator is the number of samples coded=1. For each group G (e.g., group "Only Human Wrong") and a certain indicator F like image dependence, its statistic $P(F|G)$ is calculated as:

$$P(F|G) = \frac{\sum_{i \in G} I(F_i = 1)}{|G|} \times 100\%$$

where:$|G|$ is the total number of valid sample pairings in that group (Count).$I(\cdot)$ indicates 1 when the answer conforms to the positive rule in the table, otherwise 0.

Considering differences in sample size across different groups, we focus on analyzing the *Only Human Wrong* scenario. This group not only has the largest sample size, but models can consistently make correct judgments in this situation, thus best reflecting the structural differences in reasoning strategies between the two. The data is as follows:

It can be observed that in samples where only humans are wrong, models have a relatively higher dependence on image content and consider the image

Group	Agent	Image support	Image use	Emotion cues	AI suspection
Only Human Wrong	Human	21.40%	48.10%	40.50%	31.20%
	Model	47.60%	56.40%	54.00%	10.00%
Only Model Wrong	Human	38.50%	53.80%	38.50%	53.80%
	Model	30.80%	84.60%	53.80%	7.70%
Both Wrong	Human	41.20%	68.20%	41.20%	40.00%
	Model	25.90%	83.50%	47.10%	4.70%

Table 21: Proportion of samples in which each intermediate factor was flagged.

content more misleading; this is consistent with our analysis conclusion above, that in such samples, humans base their judgment more on the content itself, thinking the content"Sounds like a joke or satire", "Totally irrelevant content", "Just personal feelings", so even if they suspect the image might be fake more than models (31%), they still don't think it's worth checking. For example, in the "super moon"image case, most human annotators thought the content would not cause substantive negative impact, even if untrue, it doesn't matter; models focused on the abnormal presentation of the image, thought it might involve synthesis or deception, thus triggering checkworthy judgment.

On the other hand, the gap between humans and models in capturing emotions in claims is not large, but the way they use emotional signals differs. From human free reasoning texts, it can be seen that humans more often describe their own subjective emotions: I think this is interesting, worth checking / I think this is a joke, absurd, no need to check; while models more often focus on the content itself, that this content may have an inflammatory emotional effect, worth checking.

For example, the following case: *a video shows osama bin laden singing lady gaga's "poker face."*

Figure 15: Example 9.

- **Gold Standard:** checkworthy.

- **Human:** all wrong, thought the content emotion was funny, spoof, no need to verify, I know it is fake.
- **Model:** thought such absurd content must be false, needs verification.

6 Discussion and Limitations

6.1 Discussion

This study compared the decision-making processes of humans and LLMs in determining whether information needs to be checked. We have identified a key difference: LLMs are closer to the gold labels, especially in identifying more potential misinformation. However, this difference is not merely a matter of accuracy; it reveals the fundamental disparity in the judgment strategies of the two.

6.1.1 Humans Weigh Values, Models Scan for Risks

The most obvious divergence between humans and LLMs occurs when dealing with content that seems absurd, humorous or entertaining, such as "Bin Laden singing pop songs" or "Bruce Lee playing table tennis with nunuchs". Human annotators have continuously employed a judgment method based on actual value. Their decision-making process is not only about judging truth or false, but also includes an implicit analysis of pros and cons. They are actually asking:"If this claim is false, will it cause actual harm or significant social impact?" When content is classified as obvious jokes, satire or low-risk entertainment, humans systematically place it at a low priority, this is a reasonable strategy for allocating limited attention and verification resources in the real world.

In contrast, the LLMs in this study mainly served as systematic risk scanners. Due to a lack of understanding of human social contexts, humor and common sense, they have adopted a conservative strategy: marking content based on potential false or misleading signs detected. Therefore, those statements that humans consider insignificant are often marked by LLMs. The reasons of the model focus on potential consequences, such as"it may mislead the public" or"it may damage an individual's reputation". This pattern indicates that LLMS can identify some potential risks, which may be filtered out by human subjective feelings or social experiences.

6.1.2 Redefine The Criteria For Check-worthiness

The gap between human and model judgment strategies requires us to reexamine the criteria for checking-worthiness. Many existing systems adopt expert-

In instruction	Not in instruction
misleading	public
health	cause
panic	social
harm	lead
risks	person
trust	seen
media	dangerous
risk	interesting
government	try
	life

Table 22: Top key words in human reasoning, split by whether the word appears in the annotation instruction.

based gold standard. The LLMs in this study may reflect that the patterns in their training data followed this standard, thus treating more claims as checkworthy.

However, human judgment reflects a more pragmatic and result-oriented standard: giving priority to content that would cause actual harm or social impact. Therefore, merely training models to imitate gold standard might capture content that has no real practical impact on society, wasting verification resources.

6.1.3 The Influence Of Instruction On Human Judgment

To help annotators clearly understand the definition of checkworthy, we used instruction at the beginning of the questionnaire to explain it. To explore whether annotators overly rely on instructions and lack their own judgment, we pulled down the key words used by the annotators within the instructions and the additional reasoning words that did not appear in the instructions. The results are shown in the following table.

Firstly, the instruction clearly emphasizes the consequences, such as health harm, social panic, or damage to trust, which constitute the core reasons for the annotator's judgment. Secondly, there are those that are not explicitly listed in the instruction but appear in the reasoning text, mainly used to assess whether the above potential consequences have the conditions for actual occurrence, such as whether the information is accessible to the public, whether it is likely to be spread, and whether it has a misleading or deceptive nature.

This result indicates that instruction does not completely constrain human judgment logic, but clarifies the core concern of check-worthiness judgment; On this basis, annotators introduce spreading and triggering conditions to conduct realistic evaluations of potential social consequences. This is consistent with our

previous analysis. Humans are more concerned about whether misinformation will actually affect and spread in real world.

6.2 Limitations

Although our research has revealed some findings, we must admit that there are some limitations in the experimental design itself, which may affect our understanding and generalization of the conclusions.

6.2.1 Limited participants

We only invited 26 people to do the annotation. This quantity is somewhat small for drawing a general conclusion. More importantly, most of these participants are from South Africa and have a relatively high proportion of women. People's perception of whether information is dangerous is actually closely related to their cultural background and life experiences. For instance, people from different regions may have different understandings of what constitutes humor or sensitive topics. Therefore, we have observed that the phenomenon where people tend to ignore seemingly funny or absurd content may, to some extent, reflect the habits of this specific group of participants.

6.2.2 Annotation quality

Although we explicitly requires annotators to provide explanatory reasoning in the annotation notes and emphasizes the completeness and specificity of the responses through instruction and examples, in the actual collected free explanatory texts, it can still be observed that some responses are overly brief or relatively perfunctory. This phenomenon is not an isolated case, so it may reflect the common challenge of online open annotation tasks: that is, in the absence of immediate control, some participants tend to complete the task at their lowest cost, thereby reducing the depth of expression of the reasoning text.

This limitation mainly affects this study in terms of the granularity of reasoning and analysis. Although the overall trend and core judgment logic remain clearly visible, brief or perfunctory responses to some extent limit the depiction of more detailed cognitive factors and may also underestimate the diversity of human reasoning.

6.2.3 The experimental environment is different from the real world environment

Our experiment design is relatively simple: participants only saw pictures and text, without any information about the identity of the publisher or the number

of likes and comments. However, this is quite different from the real social media scenarios. In reality, when we judge the reliability of information, we rely heavily on clues such as"who said it" (for example, whether it was from an authoritative media outlet) and the reactions of other netizens. In the experiment, due to the lack of these key clues, participants might be forced to rely more on whether the content itself is reasonable to make judgments, which might amplify the intuitive response of humans when facing humorous content:"This looks fake, no need to check."

6.3 Future Work

Based on the findings and limitations discussed in this study, several possible directions for future research:

6.3.1 From Risk Scanning to Value-Aware Reasoning

Our results highlight a strategic gap: LLMs act as risk scanners, while humans apply value-based filtering. A critical next step is to explore how models can be guided not only to detect potential misinformation but also to reason about the relative importance of a claim within a given social context. This could involve:

- Fine-tuning or prompting LLMs with value-laden narratives or annotated datasets that capture human prioritization logic.
- Developing hybrid architectures where a risk-scanning module (like current LLMs) is coupled with a value-assessment module that estimates potential social impact, humor intent, or personal versus public relevance.
- Exploring whether models can be trained to provide calibrated uncertainty estimates for their check-worthiness judgments, flagging not just "risky" content but also "hard-to-judge" content that need human review.

The goal is to move towards LLM assistants that better align with human pragmatic decision-making, optimizing for both thoroughness and efficiency in resource-constrained fact-checking pipelines.

6.3.2 Cross-Cultural and Demographic Validation

While our annotator pool provided initial insights, it was limited in geographic and cultural diversity (predominantly South African). Future work should recruit annotators from a more globally representative set of regions (e.g., Asia, South America, the Middle East) and varied socio-economic backgrounds. This would enable analysis of how cultural dimensions (e.g., individualism versus collectivism,

power distance) modulate perceptions of harm, humor, and check-worthiness. Such studies could ultimately inform the development of culturally adaptive fact-checking systems.

6.3.3 Improve Annotation Quality

Although setting a minimum word count and other measures can to some extent restrain overly short responses, this approach essentially only increases the output length of the annotator and cannot guarantee the quality of the reasoning itself. In fact, even when explanations are explicitly required, annotators can still complete tasks through repetitive, vague or templated expressions. This indicates that the problem does not mainly lie in "not writing enough", but rather in the excessive burden of expression and organization imposed on annotators by free text reasoning.

The solution is reducing the organizational difficulty of reasoning expression rather than simply raising the output requirements. Future annotation design could consider breaking down the original open-ended reasoning into more granular but clearly structured sub-steps, such as guiding annotators respectively to point out the key reasons on which their judgments are based, potential consequences, and the causal relationship between the two. Even if each part of the response is relatively brief, this structured design can still form a clear and analyzable reasoning chain as a whole. By breaking down the different components of "why", it makes it easier for annotators to provide specific and distinguishable judgment bases, and also offers a more stable and comparable reasoning unit for subsequent analysis.

7 Conclusion

In this study, we aimed to answer the following core questions by constructing the ACORN dataset and systematically comparing human and LLMs performance in multimodal check-worthiness judgment:

For Question 1 (How to design the annotation task): We proposed and validated a questionnaire-based annotation framework that successfully captured humans' check-worthiness judgments for image-text claims and the structured reasoning chains behind them. It explicitly collect annotators' reasoning logic in areas such as risk perception, evidence evaluation, emotional influence, and multimodal interaction, laying the foundation for building explainable fact-checking models.

For Question 2 (Consistency and diversity of human reasoning): The study found that human annotators show high surface agreement in their judgments. Their reasoning logic highly depends on the content domain: for high

risk content (e.g. politics, health), they easily form consensus based on harm consequences; for low risk entertainment content, they generally adopt a subjective filtering strategy, choosing not to check because the content is considered irrelevant, subjective, or a joke, causing no effect in real world.

For Question 3 (Human LLMs reasoning comparison): LLMs demonstrate a higher sensitivity than humans in identifying content with potential risk, thus compensating for cases where humans systematically deprioritize entertaining or absurd content. However, this difference stems from a strategic gap: humans apply a value judgment strategy, consciously filtering out what they deem inconsequential, while execute a systematic risk scanning strategy. Therefore, the human model difference is not primarily about right or wrong, but about the inherent trade off between pragmatic efficiency and conservative thoroughness in the face of uncertainty.

For Question 4 (Challenges posed by synthetic/ manipulated content): The study found that when faced with suspicious content, the key driver for check is not *confirmed false content*, but *uncertain content*. The gray areas of image-text inconsistency and content that is hard to judge for authenticity strongly stimulate check motivation. However, with the development of AI, these uncertain contents will become increasingly realistic. Some contents that are uncertain now may be impossible to distinguish in the future, so the difficulty of checking misinformation itself will increase.

References

[1] Claire Wardle et al. Information disorder: The essential glossary. *Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School*, 2018.

[2] Shahla Nasiri and Armin Hashemzadeh. The evolution of disinformation from fake news propaganda to ai-driven narratives as deepfake. *Journal of Cyberspace Studies*, 9(1):229–250, 2025.

[3] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. In *Proceedings of the 29th international conference on computational linguistics*, pages 6625–6643, 2022.

[4] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[5] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.

[6] Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. *arXiv preprint arXiv:2402.11073*, 2024.

[7] Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288*, 2024.

[8] Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. M4fc: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset. *arXiv preprint arXiv:2510.23508*, 2025.

[9] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

[10] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[11] Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*, 2023.

[12] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital threats: research and practice*, 2(2):1–16, 2021.

[13] Michiel van der Meer, Pavel Korshunov, Sébastien Marcel, and Lonneke van der Plas. Hintsoftruth: A multimodal checkworthiness detection dataset with real and synthetic claims. *arXiv preprint arXiv:2502.11753*, 2025.

[14] Gordon Pennycook, Tyrone D Cannon, and David G Rand. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865, 2018.

[15] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):159, 2024.

[16] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

A Keyword Extraction Rules in Experiment 1

This appendix documents the exact filtering and normalization rules used to extract frequent reasoning terms from annotators' free-text explanations, following the implementation in our analysis script.

A.1 Text preprocessing

- **Lowercasing & punctuation removal.** All text is converted to lowercase and punctuation symbols are replaced by whitespace before tokenization.
- **Tokenization.** Tokens are obtained by whitespace splitting after punctuation removal.
- **Length filter.** Tokens shorter than 3 characters are removed.

A.2 Stopword and task-noise filtering

We remove (i) standard English function words and (ii) task-specific "noise" terms that frequently appear in explanations but do not convey substantive reasoning (e.g., references to the annotation setup). Examples of removed task-noise terms include:

> *statement, photo, picture, image, claim, content, post, looks, seems, people, might, could, may, would, think, believe, say, true, false, question, answer.*

A.3 Manual normalization (lemma mapping)

To reduce surface variation, we manually map semantically similar variants to a shared canonical form. The mapping includes (non-exhaustive examples):

- **Misleadingness:** *misleading, misled, mislead* → *mislead*
- **Deception:** *deceive, deceptive, deception* → *deception*
- **Harm:** *harmful, harming, harm* → *harm*
- **Danger:** *danger* → *dangerous*
- **Verification:** *verification, verify* → *verify*; *checked, check* → *check*
- **Risk:** *risks, risky, risk* → *risk*
- **Consequences:** *consequences, consequence* → *consequence*
- **Manipulation:** *manipulated, manipulation* → *manipulation*
- **Humor:** *funny, humorous, humor* → *humorous*
- **Plural normalization:** *opinions* → *opinion*; *facts* → *fact*

A.4 Lightweight suffix normalization (for out-of-map tokens)

For tokens not covered by the manual mapping above, we apply a lightweight suffix heuristic:

- **Plural stripping:** remove a trailing `-s` (except tokens ending with `-ss`)
- **Gerund stripping:** remove a trailing `-ing`
- **Past tense stripping:** remove a trailing `-ed`

A.5 Label standardization used for grouping

To group explanations by check-worthiness, we standardize labels from Q5:

- Any Q5 string containing ‘‘`not`’’ (case-insensitive) is mapped to *Not Checkworthy*.
- Any Q5 string containing ‘‘`checkworthy`’’ (case-insensitive) is mapped to *Checkworthy*.

B Model Versions in Experiment2

Model Name	OpenRouter API Identifier	Provider
GPT-5.1	openrouter/openai/gpt-5.1	OpenAI
Gemini-2.5-Pro	openrouter/google/gemini-2.5-pro	Google
Claude-Opus-4.5	openrouter/anthropic/claude-opus-4.5	Anthropic
Llama-4-Maverick	openrouter/meta-llama/llama-4-maverick	Meta
Grok-4-fast	openrouter/x-ai/grok-4-fast	xAI

Table A.1: Specific model identifiers used in Experiment 2 via OpenRouter.

C Prompt Template For LLMs Annotation

All models were queried using the same prompt template. The prompt was designed to minimally constrain model behavior while enforcing a consistent output format.

```
You are an expert survey annotator.

HARD RULES:
- You MUST answer Q1 ~ Q16 in ENGLISH only.
- Do NOT use Chinese or any non-English words.
- Output format: Q1: ... Q2: ... ... Q16: ...
  No extra commentary before or after.
```

D Prompt and Model Parameter For Key words Exaction.

D.1 Parameter Configuration.

The extraction process was performed using the OpenAI GPT-4o model (openai/gpt-4o). To ensure deterministic and stable outputs during logic category assignment, the following inference settings were used:

- Temperature: 0.0
- Response format: JSON object (strict schema enforced)
- Batch size: 12 instances per request
- Maximum input length: 550 characters per reasoning text

Output constraint: one category per instance, selected from a predefined codebook

D.2 Prompt

```
You are an expert data coder, you are given a predefined codebook. For each input it

Codebook:The full list of logic categories is provided verbatim.

Important:
1You may output either the full category string or only the category name before par
2Short names will be mapped back to the full codebook entry automatically.

Input format:
You will receive a JSON array. Each item contains:
1idx: the index of the item
2text: the reasoning text provided by the annotator

For each item, you must
1Select exactly one category from the codebook.
2Extract 2{5 words from the text as Evidence.

Evidence must be copied verbatim from the text.
Do not paraphrase.
If the text is blank or too short, output:Category = "Other", Evidence = ""
```