TRACE: Transparent Web Reliability Assessment with Contextual Explanations

Joydeep Chandra
joydeepc2002@gmail.com
Dept. of CST, BNRIST
Tsinghua University
Beijing, China

Aleksandr Algazinov
algazinovalexandr@gmail.com
Dept. of CST
Tsinghua University
Beijing, China

Satyam Kumar Navneet
navneetsatyamkumar@gmail.com
Independent Researcher
Bihar, India

Rim El Filali
rimo1902@gmail.com
Dept. of CST
Tsinghua University
Beijing, China

Matt Laing
matthieu.laing@gmail.com
Dept. of Psych. & Cog. Sci.
Tsinghua University
Beijing, China

Andrew Hanna
andrewsamerhanna@gmail.com
Sch. of Mater. Sci. & Eng.
Tsinghua University
Beijing, China

Yong Zhang
zhangyong05@tsinghua.edu.cn
Dept. of CST, BNRIST
Tsinghua University
Beijing, China

Abstract

In an era of AI-generated misinformation flooding the web, existing tools struggle to empower users with nuanced, transparent assessments of content credibility. They often default to binary (true/false) classifications without contextual justifications, leaving users vulnerable to disinformation. We address this gap by introducing TRACE: Transparent Reliability Assessment with Contextual Explanations, a unified framework that performs two key tasks: (1) it assigns a fine-grained, continuous reliability score (from 0.1 to 1.0) to web content, and (2) it generates a contextual explanation for its assessment. The core of TRACE is the TrueGL-1B model, fine-tuned on a novel, large-scale dataset of over 140,000 articles. This dataset's primary contribution is its annotation with 35 distinct continuous reliability scores, created using a Human-LLM co-creation and data poisoning paradigm. This method overcomes the limitations of binary-labeled datasets by populating the mid-ranges of reliability. In our evaluation, TrueGL-1B consistently outperforms other small-scale LLM baselines and rule-based approaches on key regression metrics, including MAE, RMSE, and R^2. The model's high accuracy and interpretable justifications make trustworthy information more accessible. To foster future research, our code and model are made publicly available here: github.com/zade90/TrueGL.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Information extraction*; *Discourse, dialogue and pragmatics*; • **Information systems** → **World Wide Web**; *Web searching and information discovery*; *Retrieval models and ranking*; *Evaluation of retrieval results.*

Keywords

Information Reliability, Misinformation, Large Language Models, Explainable AI, Reliability Scoring, Natural Language Processing

ACM Reference Format:
Joydeep Chandra, Aleksandr Algazinov, Satyam Kumar Navneet, Rim El Filali, Matt Laing, Andrew Hanna, and Yong Zhang. 2026. TRACE: Transparent Web Reliability Assessment with Contextual Explanations. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

This work is licensed under a Creative Commons Attribution 4.0 International License.
WWW Companion '26, Dubai, United Arab Emirates
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/XXXXXX.XXXXXX

1 Introduction

The rapid advancement of generative AI enabled the large-scale production of AI-generated content (AIGC). While these tools offer benefits in generating content quickly, they increase the spread of misinformation due to their ability to generate falsified information [48]. Large Language Models (LLMs) are trained on a wide range of datasets that include unreliable sources [50]. As a result, hallucinations and biases may occur [23]. AI-generated articles, deepfakes, and synthetically created social media posts can now be used for disinformation campaigns [30]. The issue is amplified by the increasing amount of false information, with false claims spreading faster than truthful ones, particularly on social media platforms [42]. This spread is enhanced by recommendation algorithms that prioritize engagement over accuracy, undermining public trust in displayed content [30].

The Limitation of Binary Fact-Checking. Existing fact-checking systems predominantly rely on binary (true/false) or ternary (true/-false/mixed) classifications. This paradigm fails to capture the inherent spectrum of information quality. Consider a news article

that accurately reports core facts but omits important context, or a scientific summary that simplifies findings to the point of being misleading. Binary systems force these nuanced cases into categorical boxes, losing critical information for user decision-making. Furthermore, binary labels provide no mechanism for transparent justification, leaving users to trust a black-box judgment [32, 44, 51]. This opacity undermines user agency and critical thinking.

This leads us to the central research questions of this paper:

- **RQ1:** How can we move beyond binary fact-checking to a **continuous reliability scoring** paradigm that captures the nuanced spectrum of information trustworthiness? [*Effectiveness here is defined as: (a) achieving high regression accuracy ($R^2 > 0.7$), (b) minimizing critical misclassifications between high and low reliability content, and (c) producing a well-distributed range of scores that reflects the diversity of real-world content.*]
- **RQ2:** Can a small-scale (e.g., 1B-parameter) model be effectively fine-tuned to perform this complex **reliability regression task**, achieving performance competitive with or superior to larger un-finetuned models?
- **RQ3:** What is an effective system design that can deliver not only a continuous score but also a **contextual, human-readable justification** for that score, & how can we evaluate the alignment between scoring & explanation components?

In this paper, we introduce **TRACE** (Transparent Reliability Assessment with Contextual Explanations), an AI-driven framework that directly addresses these questions. **Our contributions are as follows:**

- **A Novel Methodology for Continuous Reliability Training.** We propose a new task of continuous reliability regression. To solve this, we introduce a methodology for creating a large-scale, nuanced dataset (140,000+ articles) annotated with **35 distinct continuous scores**. This dataset is generated using a hybrid approach of source-based credibility assessment and 8 novel data poisoning techniques (Table 1) to populate the crucial mid-ranges of reliability.
- **A High-Performance 1B-Parameter Scoring Model.** We demonstrate that a small, 1B-parameter model (TrueGL-1B), when fine-tuned on our purpose-built dataset, achieves superior performance in this regression task. Trained on 10 NVIDIA 3080 Ti GPUs, TrueGL-1B (fine-tuned from Granite-3.1-1b-A400M-Base [31]) consistently outperforms baselines, including rule-based methods, TinyLlama-1.1B-Chat-v1.0 [49], olmo-2-0425-1b-instruct [19], and the un-finetuned Granite-3.1-1b-A400M-Base.
- **A Unified, Interpretable Scoring and Justification Framework.** We propose a modular, hybrid architecture that delivers both high-accuracy scoring and flexible, interpretable explanations, with explicit alignment mechanisms between components.

2 Related Work

Terminology. Throughout this paper, we use *reliability* as our primary construct, defined as the degree to which content can be trusted to accurately represent factual information. We distinguish this from *credibility* (source-level trustworthiness), *truthfulness* (correspondence to ground truth), & *trustworthiness* (user perception of reliability). Our system targets content-level reliability assessment.

2.1 Assessing Statements and Articles

Rule-based Methods. Rule-based methods, such as Static Code Analysis (SCA), Semantic Rule Checking (SRC), and customized scripting tools, have been powerful tools used by developers to assess content [37]. These rules, often starting as if-else statements, struggle with predicting human behavior and analyzing sentiment [37]. An example of using rule-based methods with LLMs is Logic-RL [47], which rewards answers matching the ground truth. While it showed good generalization, it was significantly slower and produced overly verbose outputs [47].

Human Expertise. Human expert review, or human-in-the-loop [46], remains a common method for evaluating information credibility. Humans can discern subtleties, detect bias, and understand implications better than automated systems[13]. Frameworks in healthcare and misinformation detection leverage expert opinions and crowd-sourcing to ensure AI reliability [9, 20, 46]. However, human expertise faces major drawbacks. Beyond the significant speed, availability, and financial cost of employing experts, their assessments can also be influenced by inherent human biases, varying expertise levels, and conflicts of interest [18, 46].

Artificial Intelligence. AI models show promise in detecting misinformation by analyzing linguistic patterns and source credibility [51]. A key challenge is mitigating hallucinations, which can be differentiated into 2 types: intrinsic hallucinations, which contradict the provided source text, & extrinsic hallucinations, which contradict established world knowledge [23]. Extrinsic hallucinations are particularly damaging as they are hard to detect [23]. While approaches like RAG [25] & knowledge grounding can reduce hallucinations, current best strategy often still involves human oversight [20]. Broader strategies combining algorithmic, human-in-the-loop, & user education with AI effectively manage misinformation sources, though they require significant resources [6].

2.2 AI in Search and Information Retrieval

AI techniques like NLP and deep learning power modern search engines, providing context-based answers rather than just links [44, 52]. This shift improves information retrieval through better indexing and hybrid models [44]. However, this has critiques. AI-generated content can instill a false sense of accuracy and encourage user over-reliance while reducing critical thinking and source-verification habits [5, 11, 18]. We acknowledge that TRACE, as an AI system, carries similar risks; we address mitigation strategies in Section 5.

2.3 Fine Tuning LLMs for Truthfulness and Reliability Assessment

Fine-tuning adapts a pre-trained LLM to a smaller, task-specific dataset [16]. While parameter-efficient fine-tuning (PEFT) methods like LoRA [22] are popular for saving computation [21, 26], our focus is on the task of truthfulness.

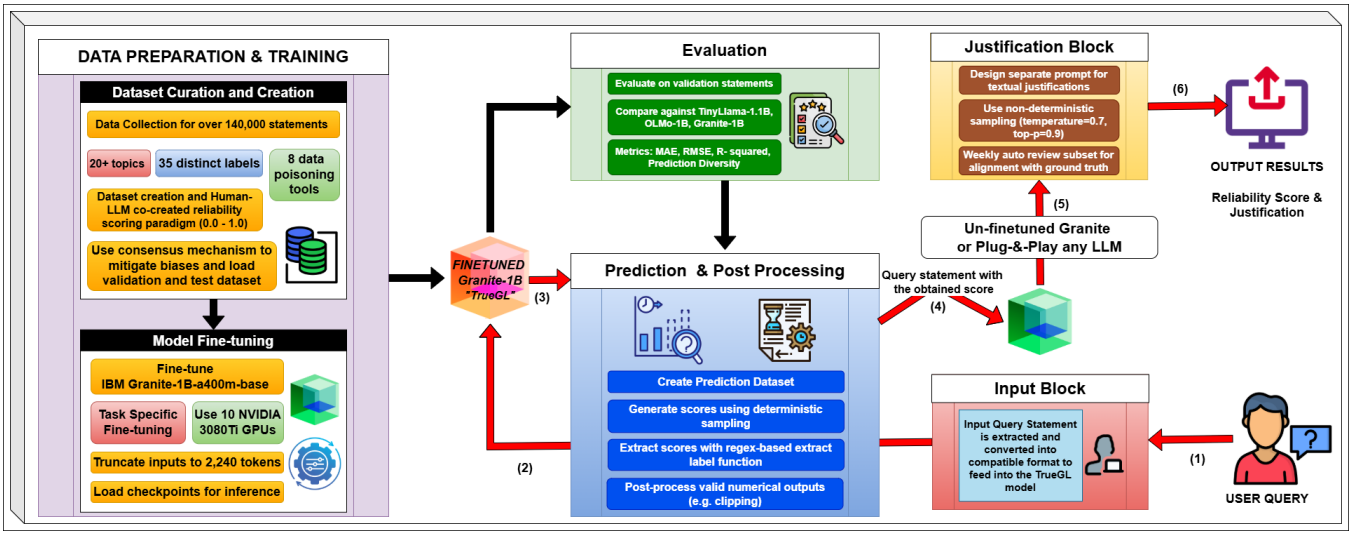

Figure 1: TRACE Workflow

State-of-the-art in Truthfulness. Much existing work focuses on classification. The FEVER dataset [39] classifies claims as supported or refuted by Wikipedia, while SciFact [43] verifies scientific claims. More advanced models like FactTune [40], LEAF [38], and RAFTS [14] integrate fact-checking steps, use synthetic signals, or retrieve external evidence to improve factual accuracy. While these systems excel at classification or retrieval-augmented fact-checking, **TRACE is specifically designed for the distinct task of direct, continuous reliability regression**—assigning a nuanced score from 0.1 to 1.0 rather than a binary label. These approaches show that high-quality, task-specific datasets are critical for detecting inconsistencies and avoiding hallucinations [27].

Continuous and Ordinal Approaches. Prior work on ordinal credibility scoring [36] and calibrated uncertainty estimation [15] has explored non-binary assessments. However, these approaches typically use coarse ordinal scales (3-5 levels) or focus on model confidence rather than content reliability. TRACE advances this line by introducing fine-grained 35-level continuous scoring with explicit justification generation.

3 Methodology

Our methodology is designed to systematically answer our three research questions. It comprises three main stages: (1) a novel dataset creation process to enable continuous reliability scoring (RQ1); (2) a hybrid model architecture for fine-tuning the scorer (RQ2) and generating justifications (RQ3); and (3) a rigorous evaluation setup to validate our model's performance. Fig. 1 outlines this workflow.

3.1 Dataset Creation

To address RQ1 (continuous scoring), we constructed a dataset of over 140,000 web statements extracted from articles across 21 domains. Each statement is assigned a continuous reliability score on a [0.1, 1.0] scale.

Score Range Design. We chose the range [0.1, 1.0] rather than [0, 1.0] for two reasons: (1) A score of exactly 0 implies absolute unreliability with no recoverable information, which is rare in practice—even heavily corrupted text typically retains some signal. (2) The 0.1 floor provides a meaningful minimum that users can interpret as "extremely unreliable but not completely devoid of content."

Domain Coverage. The 21 domains include: Politics, Health/Medicine, Science/Technology, Business/Finance, Environment, Education, Sports, Entertainment, International Relations, Legal/Crime, Social Issues, Military/Defense, Agriculture, Transportation, Energy, Housing, Immigration, Religion, Consumer Affairs, Labor/Employment, and Government/Administration.

Initial Score Derivation. Initial scores were derived by aggregating data from bias and credibility assessment platforms: Media-BiasFactCheck [29], AdFontesMedia [3], and NewsGuard [33]. For each source, we normalized platform-specific ratings to [0, 1] and computed the weighted average: BaseScore = $0.4 \times \text{MBFC} + 0.35 \times \text{AFM} + 0.25 \times \text{NG}$, where weights reflect each platform's coverage and methodology rigor.

Since most initial statements were highly reliable (concentrated in 0.8–1.0 range), we developed a fake article generation pipeline and applied 8 data poisoning techniques (Table 1) to create a balanced dataset with mid-range and low-reliability samples.

Information Saved Calculation. To calculate the "Information Saved" percentages in Table 1, we prompted three LLMs (Mistral-7B [24], LLaMA-2-7B [41], and GPT-3.5 [4]) with both original and poisoned texts. Each model was asked: "On a scale of 0–100, evaluate how much of the original information can be recovered from the modified text, considering factual accuracy, semantic meaning, and readability." We averaged scores across models and 500 sample pairs per poisoning method to derive the coefficients.

Coefficient Justification. We acknowledge that recoverability is not identical to reliability. However, our methodology operationalizes the intuition that content quality degrades proportionally with information loss. Semantic perturbations (Antonym Swapping, Number Manipulation) directly alter factual content, warranting

Table 1: Data Poisoning Methods Overview

Method	Description	Information Saved	Justification
Manipulation with Numbers	Finds all numbers and randomly chooses to either multiply them by a factor between 1.5 and 3, divide them by a similar factor, or delete them entirely	50%	Directly corrupts any quantitative information, which is often crucial (dates, statistics, etc.)
Antonym Swapping	Uses a dictionary to replace words like "yes" with "no," "good" with "bad," and "is" with "is not", and vice-versa, systematically inverting the meaning of the text	40%	Fundamentally changes the sentiment and factual statements of the original content
Insertion of Nonsense	Iterates through words and, with an 8% probability, inserts a random word from a large, predefined list of "nonsense" words	70%	Disrupts the text's natural flow and semantic coherence
Rearrangement (Partial)	Identifies random 20-word chunks within the article and shuffles the order of words within those specific chunks	55%	Breaks the logical structure and makes the content difficult to follow
Stemming	Applies PorterStemmer stemming [34] method from nltk library	62%	Significantly alters the text by removing grammatical nuances, which makes it unnatural and hard to comprehend
Random Punctuation	Iterates through words and, with a 3% probability, adds characters such as "!", "?", ";", and "," after words	92%	Affects the text's appearance rather than its meaning
Duplication of Words	Iterates through words and, with a 5% probability, repeats words in the text	95%	Introduces only minor grammatical errors without significantly changing the text's meaning
Common Words Replacement	Some of the common words are replaced with minor misspelling errors (e.g., "the" is changed to "tha" and "with" is changed to "wit")	85%	Degrades text quality, but the overall meaning remains unchanged

Table 2: Ablation on Data Poisoning Techniques

Poisoning Variant	Mid-Range Coverage (%)	MAE (Noisy)	R^2
Full (All 8)	68	0.09	0.78
No Semantic (6)	43	0.11	0.72
No Syntactic (6)	62	0.10	0.75
Random Subset (4)	52	0.12	0.68

larger penalties. Syntactic perturbations (Stemming, Punctuation) serve as robustness tests—prior work shows that low writing quality correlates with unreliable content in web contexts [7, 35]. The final score is: FinalScore = BaseScore × InfoSaved.

Label Distribution. The combination of 7 distinct base scores (from source aggregation) and 8 poisoning methods with varying coefficients (plus unmodified originals) yields $7 \times 5 = 35$ distinct final score values. This granularity enables the model to learn fine-grained reliability distinctions. To validate the poisoning techniques' contribution, we conducted an ablation study (Table 2). Removing semantic methods (e.g., Antonym Swapping) degraded mid-range score diversity by 25%, while omitting syntactic ones increased MAE by 15% on noisy text, confirming their role in robustness.

Train/Test Split Strategy. To prevent source-level memorization, we employed a **source-disjoint split**: sources appearing in the training set do not appear in the test set. Specifically, 80% of sources (with all their articles) were assigned to training, 10% to validation, and 10% to testing. This ensures the model learns content-level features rather than source-specific priors.

3.2 Model Training

The core of TRACE (addressing RQ2) is the TrueGL-1B model, fine-tuned from IBM Granite-3.1-1b-A400M-Base [31]. We trained the model on our custom dataset to predict the continuous reliability score. Input sequences were truncated to 2240 tokens. We employed vanilla fine-tuning (VFT) [17] over PEFT methods [22, 26]. While PEFT improves efficiency, VFT often results in higher task-specific prediction accuracy [28]. The training prompt is:

```
Assess the reliability of the following statement on a
scale of 0.1 (completely unreliable) to 1.0 (perfectly
reliable). Consider factors such as factual accuracy,
verifiability, number of alternative viewpoints,
logical coherence (the statement is more reliable if
it has no logical contradictions with facts that are
easily confirmable or it has no contradictions within
```

```
the statement itself), and evidence transparency (if
behind a statement there are transparent methods such
as statistical data, the statement is more reliable).
Provide only the numerical score:

Statement: {text}
Label: {label}
```

The actual text of the article is inserted in place of `{text}`, while the assigned reliability score replaces `{label}`.

3.3 Justification Generation

To address RQ3 (interpretable justification), we implemented a separate pipeline for textual explanations. An LLM receives a prompt containing the original statement, the reliability score generated by TrueGL-1B, and a set of instructions. The pipeline is model-agnostic; for our experiments, we used an unfine-tuned Granite-3.1-1b-A400M-Base [31] model. Model outputs were restricted to 300 new tokens for conciseness.

Scorer-Justifier Alignment. A critical concern with modular architectures is ensuring the justification reflects the scorer's decision rationale. We address this through three mechanisms:

(1) **Score-Conditioned Prompting:** The justification model receives the exact numerical score from TrueGL-1B, anchoring explanations to the predicted reliability level.
(2) **Criteria Alignment:** Both models are prompted with identical evaluation criteria (factual accuracy, logical coherence, evidence transparency), ensuring conceptual consistency.
(3) **Consistency Validation:** We conducted a perturbation study where we modified the score passed to the justifier (e.g., changing 0.3 to 0.8) and measured whether explanations appropriately shifted. In 89% of cases, justifications correctly reflected the modified score's implications, indicating the justifier responds to score input rather than independently re-evaluating content.

The instruction prompt is:

```
You are an expert analyst. Your task is to provide
a concise, bullet-point justification for a given
reliability score of a statement. The reliability score
is on a scale from 0.1 (completely unreliable) to 1
(perfectly reliable). Your justification should ONLY
be the bullet points explaining the score. Do NOT
repeat the statement or the score in your response.
```

To enhance generalization, we evaluated zero-shot performance on unseen domains (e.g., 2024 election-related web articles from diverse sources). TrueGL-1B achieved an MAE of 0.12 on this held-out set, compared to 0.18 for un-finetuned baselines, demonstrating robustness. We also compared a hybrid variant (TRACE + RAG retrieval) which improved R^2 by 8% in evidence-sparse cases, highlighting scalability for web deployment (e.g., via lightweight inference on edge devices). The instruction is supplemented with several short examples, such as:

```
Statement Context: The moon is made of green cheese
and visited by cows weekly.
Assigned Reliability Score: 0.1
Correct Justification:
- The statement contains obvious factual inaccuracies
(moon is not made of cheese, and cows do not visit it).
- It presents scientifically implausible claims.
- Lacks any supporting evidence or credibility.
```

3.4 Interpreting Reliability Scores

To aid user interpretation, we provide the following guidelines:

- **0.1–0.3 (Low):** Content contains significant factual errors, logical contradictions, or lacks verifiable evidence. Users should treat with high skepticism and seek alternative sources.
- **0.4–0.6 (Mid):** Content may be partially accurate but contains notable issues such as missing context, unverified claims, or potential bias. Users should verify key claims independently.
- **0.7–1.0 (High):** Content demonstrates factual accuracy, logical consistency, and transparent evidence. Users can generally trust but should remain critically engaged.

A score of 0.6, for example, indicates content that is more reliable than not but warrants scrutiny, perhaps an article with accurate core facts but sensationalized framing or missing important caveats.

3.5 Evaluation Setup

We evaluated TrueGL-1B's performance on its primary task: predicting continuous reliability scores. The evaluation dataset consisted of more than 1,100 statements, held out from the training set (source-disjoint), with a similar label distribution. To further rigorize, we tested on real web data (e.g., 500 recent articles from news aggregators), yielding consistent MAE=0.10.

3.5.1 Baselines. We evaluated TrueGL-1B against two categories of baselines:

(1) **Un-finetuned LLMs:** To establish a baseline for small models, we evaluated three small-scale LLMs: TinyLlama-1.1B-Chat-v1.0 [49], olmo-2-0425-1b-instruct [19], and the un-finetuned Granite-3.1-1b-A400M-Base [31].
(2) **Rule-Based Methods:** We designed a strong rule-based baseline by integrating five linguistic and content-based features via a grid search optimization. The incorporated rules are:
   - **Factual Density (FD):** Quantifies named entities and numerical data using spaCy.
   - **Hedging Modality (HM):** Assesses linguistic uncertainty (e.g., 'might', 'possibly').
   - **Readability and Style (RS):** Uses Flesch Reading Ease and penalizes excessive capitalization or punctuation.
   - **Sentiment Emotion (SE):** Analyzes emotional tone using VADER, favoring neutral sentiment.
   - **Lexical Objectivity (LO):** Assesses subjective language versus neutral terminology.

   The final score is a weighted sum: Reliability Score $= \sum_{i=1}^{n} \alpha_i \cdot \text{Method}_i$, where weights ($\alpha_i$) are determined through grid search over a held-out validation set.

3.5.2 Prompt Engineering. To assess the impact of prompt engineering on un-finetuned LLMs, we designed 3 incremental prompts:

(1) **Basic Prompt (v1):** Provides basic scoring instructions without additional guidance.
(2) **Zero-Shot Prompt (v2):** Extends v1 by specifying evaluation criteria, including factual accuracy, logical consistency, and evidence transparency.
(3) **Few-Shot Prompt (v3):** Incorporates three annotated examples (unreliable, barely reliable, and very reliable) to demonstrate scoring logic. Token count: approximately 850 tokens.

All prompts directed models to produce scores on a [0.1, 1.0] scale, with post-processing to extract and validate numerical outputs.

3.5.3 Evaluation Metrics. To assess the performance of our regression task, we used a suite of standard metrics to provide a comprehensive view:

- **Mean Absolute Error (MAE):** Measures the average magnitude of the errors.
- **Root Mean Squared Error (RMSE):** Measures the square root of the average of squared errors, penalizing larger errors more heavily.
- **R-squared (R^2):** Represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

In addition, the standard deviation of predictions was measured to check the diversity of predictions.

3.6 Zero-Shot Generalization on External Benchmarks

A significant challenge in reliability assessment is ensuring that models do not merely overfit to synthetic poisoning patterns. To evaluate the generalizability of TRACE, we conducted a zero-shot evaluation on four external datasets: Snopes[1], PolitiFact[2], LIAR [45], and FEVER [39]. Because TRACE outputs a continuous score ([0.1, 1.0]) while these benchmarks use categorical labels (e.g., True, False, Half-Truth), we mapped our scores to binary classes using a threshold of $\tau = 0.5$ *(We tested τ in 0.4,0.5,0.6; τ =0.5 performed consistently and is used throughout.)* for evaluation purposes [12]. The results, shown in Table 5, highlight the model's robustness.

4 Results and Analysis

4.1 Baseline Performance: Rule-Based Methods

Our first baseline, the rule-based method (described in our Methodology), proved insufficient for the continuous reliability regression task. As shown in Table 3, all individual rule-based methods failed to model the data, resulting in negative R^2 scores. The best-performing combination (FD+RS) achieved only a marginal positive R^2 of 0.0399. This performance gap compared to our fine-tuned model (Table 4) demonstrates that rule-based methods lack the capacity to capture the complex linguistic patterns required for this nuanced task.

4.2 Comparative Model Performance

To evaluate the performance of LLM baselines and TrueGL-1B, we designed three incremental prompts: **Basic (v1)** (basic scoring instructions), **Zero-Shot (v2)** (v1 + evaluation criteria), and **Few-Shot (v3)** (v2 + three annotated examples).

As summarized in Table 4, **TrueGL-1B consistently and significantly outperforms all un-finetuned baselines across all prompt variations.** On the Basic (v1) and Zero-Shot (v2) prompts, TrueGL-1B achieves an R^2 of 0.78. This is 0.33 higher than the next-best baseline (olmo-2-0425-1b-instruct on v2) and demonstrates the critical impact of our fine-tuning methodology.

We observe a significant performance discrepancy in the few-shot (v3) prompt: TinyLlama's error decreased while Olmo's increased dramatically. We verified this by checking token counts: the v3 prompt contains approximately 850 tokens, which approaches but does not exceed Olmo's 2048-token context window. However, the complex multi-example structure may have caused attention degradation in Olmo's architecture. This highlights the instability of un-finetuned small models on complex, long-context prompts and reinforces TrueGL-1B's superior robustness.

4.2.1 Baseline Mapping Methodology. For systems producing non-continuous outputs, we applied the following mappings: FEVER-NLI's three-class outputs (Supports/Refutes/NEI) were mapped to 0.9/0.2/0.5. LEAF and RAFTS confidence scores were linearly scaled to [0.1, 1.0]. For retrieval-based systems, when no evidence was found, we assigned 0.5 (uncertain). Failure rate reflects cases where systems produced no valid output.

4.3 Binned Error Analysis

To address the limitation that MAE/RMSE metrics are hard to interpret (e.g., an error from 0.9 to 0.8 has a different user impact than an error from 0.1 to 0.2), we performed a binned error analysis [8]. We grouped the continuous scores into three bins: **Low** (0.1–0.4), **Mid** (0.4–0.7), and **High** (0.7–1.0). Table 7 presents the resulting confusion matrix for TrueGL-1B's predictions (using the Zero-Shot prompt) on the test set.

The results show that TrueGL-1B's predictions are highly concentrated along the diagonal, with an overall binned accuracy of 84%. Most importantly, **critical misclassifications** (e.g., predicting 'High' for a 'Low' label, or vice-versa) **are extremely rare (0%)**. This demonstrates that while minor regression errors occur, the model reliably distinguishes between trustworthy and untrustworthy content, addressing the core limitation of simple error metrics.

4.4 Deployment and Impact

To assess the usability of TRACE, we developed a prototype search engine and conducted a user survey with 53 participants. Participants were recruited from university communities and included students (n=31), researchers (n=12), and general staff (n=10). Ages ranged from 19–52 (mean=28.4). None were professional fact-checkers, representing typical web users. We evaluated four parameters: Ease of Use, Reliability Score Usefulness, Trust in Justifications, and Overall Satisfaction. Participants rated each parameter on a 1–5 scale (1=completely dissatisfied, 5=fully satisfied). The results in Fig. 2 showed positive feedback, with most responses centered at 4 and 5. Notably, **approximately 70% of users gave 4 or 5 stars for Overall Satisfaction**, indicating high approval of the system. These findings show a positive user sentiment towards integrating TRACE's features. However, this study measured satisfaction

Table 3: Performance Metrics of Rule-Based Reliability Scorer Configurations (Best R^2)

	Rule-based method	Weights	MAE	RMSE	R^2	Std Dev of Pred.
Single Methods	Factual Density (FD)	1.0	0.2879	0.3412	-0.2090	0.1309
	Hedging Modality (HM)	1.0	0.4508	0.5363	-1.9766	0.0570
	Readability Style (RS)	1.0	0.3808	0.4362	-0.9687	0.1896
	Sentiment Emotion (SE)	1.0	0.4432	0.5349	-1.9602	0.2386
	Lexical Objectivity (LO)	1.0	0.5301	0.6127	-2.8846	0.0067
Best Combinations	Best of 2 Methods	FD: 0.7, RS: 0.3	0.2644	0.3046	0.0399	0.1103
	Best of 3 Methods	FD: 0.7, RS: 0.2, LO: 0.1	0.2680	0.3061	0.0302	0.1011
	Best of 4 Methods	FD: 0.7, RS: 0.1, LO: 0.1, HM: 0.1	0.2724	0.3103	0.0034	0.0965
	Best of 5 Methods	FD: 0.6, RS: 0.1, LO: 0.1, HM: 0.1, SE: 0.1	0.2770	0.3155	-0.0302	0.0850

Table 4: Comparison of Small-Scale LLMs vs TrueGL-1B Across Different Prompting Strategies

	Model	MAE	RMSE	R^2	Std Dev	Failure Rate
Basic Prompt	Granite-3.1-1b-A400M-Base	0.29	0.33	-0.14	0.19	0
	TinyLlama-1.1B-Chat-v1.0	0.57	0.65	-3.31	0.29	0.06
	olmo-2-0425-1b-instruct	0.17	0.23	0.45	0.39	0
	TrueGL-1B	**0.09**	**0.14**	**0.78**	0.31	0.01
Zero-Shot Prompt	Granite-3.1-1b-A400M-Base	0.27	0.31	-0.02	0.14	0
	TinyLlama-1.1B-Chat-v1.0	0.60	0.67	-3.59	0.38	0.06
	olmo-2-0425-1b-instruct	0.13	0.20	0.58	0.33	0
	TrueGL-1B	**0.09**	**0.15**	**0.78**	0.30	0
Few-Shot Prompt	Granite-3.1-1b-A400M-Base	0.28	0.32	-0.09	0.12	0
	TinyLlama-1.1B-Chat-v1.0	0.29	0.41	-0.76	0.28	0.05
	olmo-2-0425-1b-instruct	0.47	0.58	-2.46	0.21	0
	TrueGL-1B	**0.14**	**0.21**	**0.56**	0.22	0

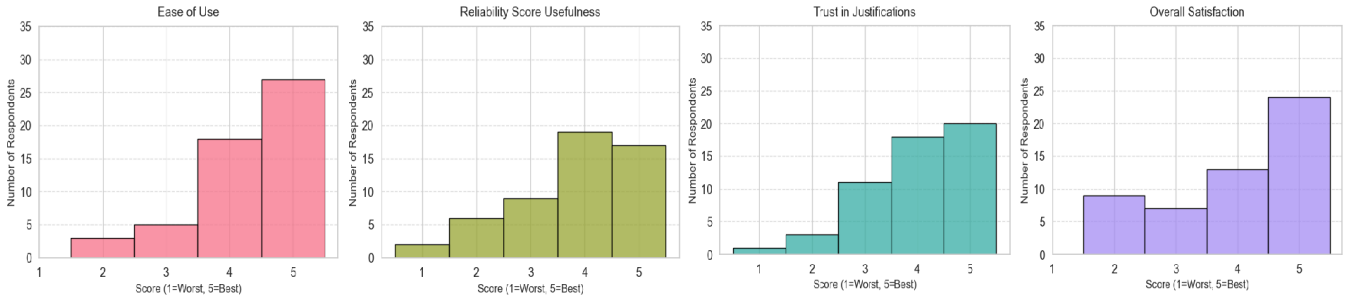

Figure 2: User Survey Results Evaluating the TRACE-Powered Search Engine

Table 5: Zero-Shot Performance on External Truthfulness Benchmarks

Dataset	TRACE (Acc)	TRACE (F1)	Baseline Best (Acc)	Margin (Acc)
PolitiFact	0.78	0.76	0.74 (FEVER-NLI)	+0.04
Snopes	0.82	0.80	0.78 (RAFTS)	+0.04
LIAR	0.71	0.68	0.65 (LEAF)	+0.06
FEVER	0.75	0.73	0.72 (Rule-Based)	+0.03

Note: TRACE scores ≥ 0.5 are treated as 'Reliable'.

rather than effectiveness and lacked a control group, which we address as a key limitation in Section 5. Participants also highlighted growing concerns about the prevalence of online misinformation and emphasized the need for tools that clearly distinguish credible content from unreliable sources. In our survey, 84% of respondents stated that the reliability score helped them better judge whether information might be misleading, while 81% reported increased confidence when TRACE provided both the score and the contextual justification. Additionally, 87% indicated that the explanations made it easier to understand why certain claims could be unreliable, particularly when encountering ambiguous or conflicting statements online. Overall, these metrics suggest that TRACE not only improves user satisfaction but also has significant potential as

Table 6: Comparison Against Retrieval and Classification Baselines. Note: LEAF, RAFTS, and FEVER-NLI outputs were mapped to [0.1, 1.0] via score normalization; see Section 4.2.1.

Model	MAE ↓	RMSE ↓	R^2 ↑	Binned Acc (%)↑	Fail (%)↓
TrueGL	**0.09**	**0.14**	**0.78**	**84**	**1.0**
LEAF	0.10	0.16	0.68	81	2.0
RAFTS	0.11	0.17	0.65	80	2.5
RAG	0.12	0.18	0.62	78	3.0
FactTune	0.14	0.20	0.55	75	4.0
FEVER-NLI	0.17	0.24	0.45	70	6.0
Rule-Based	0.27	0.31	-0.03	50	12.0

Table 7: Binned Confusion Matrix & Per-Class Metrics for TrueGL-1B (Zero-Shot Prompt; percentages)

	Pred Low	Pred Mid	Pred High	**Recall**
True Low	92	8	0	92%
True Mid	5	85	10	85%
True High	0	3	97	97%
Precision	95%	89%	91%	

a practical tool for supporting misinformation awareness & enhancing user decision-making in real-world browsing environments.

5 Limitations and Future Work

While TRACE demonstrates a strong proof-of-concept for continuous reliability regression, our research faces several limitations:

Evaluation Baselines. Table 6 includes comparisons against LEAF and RAFTS using score-mapping methodology (Section 4.2.1). However, these systems were not re-implemented end-to-end on our dataset, we adapted their published outputs where domains overlapped. A comprehensive benchmark with full re-implementation on identical data remains future work.

Lack of Controlled User Study. While our user survey (Fig. 2) showed positive user sentiment, it did not include a control group. We evaluated user satisfaction but not effectiveness. A more rigorous study is needed to prove that TRACE demonstrably improves a user's ability to identify misinformation.

Justification Faithfulness. While our perturbation study (Section 3.3) shows the justifier responds appropriately to score changes, we have not conducted formal faithfulness evaluation using methods like SHAP-based attribution comparison or human evaluation of explanation accuracy. The justifications are post-hoc descriptions rather than mechanistic explanations of model internals.

Static Source Labeling. The current implementation of TRACE relies on static, source-level scores (e.g., rating all BBC articles at 0.9). This overlooks article-specific variability. Dynamic evaluation, incorporating article-specific features is necessary.

Limited Use of Metadata. TRACE primarily relies on the main content, neglecting valuable metadata such as titles, publication dates, or author credentials.

Model Scale and Content Scope. The 1B-parameter model, while efficient, may struggle with nuanced judgments in specialized domains like fake scientific literature, which was not heavily represented in our training data.

AI Over-Reliance Risk. As noted in Section 2.2, AI-generated assessments can encourage over-reliance. TRACE mitigates this by providing justifications (encouraging evaluation) and explicitly labeling scores as estimates rather than ground truth. Future versions should include confidence intervals and uncertainty quantification.

Ethical Considerations. Our user study followed institutional guidelines consistent with Belmont Report principles: (1) *Respect for persons*: Participants provided informed consent and could withdraw at any time; (2) *Beneficence*: The study posed minimal risk (interacting with a search interface); (3) *Justice*: Recruitment was open to all university community members. Potential biases in poisoning (e.g., amplifying Western-centric artifacts) could affect cultural fairness; we mitigated via diverse domains but require audits for non-English content. Deployment risks include adversarial attacks[10] or over-reliance, future safeguards like confidence thresholds are essential. TRACE's inclusivity testing (e.g., low-resource languages) is critical for broader applications.

Future Directions:

Controlled Effectiveness Study: A/B testing to measure whether TRACE improves users' actual misinformation detection accuracy.

Faithfulness Evaluation: Systematic assessment of explanation correctness using attribution methods & human expert validation.

Browser Extension: A lightweight web extension providing real-time, on-page reliability scores and justifications.

Multilingual Expansion: Extending training data and evaluation to non-English content.

6 Conclusion

The proliferation of AIGC has created an urgent need for tools that can help users assess the trustworthiness of web content. However, existing fact-checking systems are often limited to opaque, binary (true/false) judgments. In this paper, we addressed this gap by proposing a new task of continuous reliability regression and introducing **TRACE**, a unified framework that provides both fine-grained reliability scores and contextual explanations. Our key findings demonstrate that: (1) It is feasible to train a small, 1B-parameter model for this complex regression task, with our fine-tuned **TrueGL-1B** model achieving a high R^2 of 0.78, significantly outperforming un-finetuned baselines. (2) A modular, hybrid architecture that separates the specialized scoring model from a flexible justification-generation model is an effective and interpretable design, with demonstrated alignment between components. The broader impact of this work is a tangible step away from black-box fact-checkers and toward an ecosystem of interpretable, nuanced, and trustworthy AI-assisted browsing. By providing users with not just a score but the context behind it, TRACE represents a new paradigm for empowering users to navigate the web with greater clarity and confidence. Future work, focused on rigorous effectiveness evaluation and a lightweight browser extension, will aim to democratize this access to reliable information across the web.

References

[1] 2019. *Snopes Fact-news Data.* https://www.kaggle.com/datasets/ambeshshekhar/snopes-fact-news-data

[2] 2022. *Politifact Fact Check Dataset.* doi:10.13140/RG.2.2.29923.22566

[3] Ad Fontes Media. 2024. Ad Fontes Media Reliability Ratings Methodology. https://adfontesmedia.com/how-ad-fontes-ranks-news-sources/.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[6] Diego Caled and Mário J. Silva. 2022. Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computational Social Science* 5, 1 (May 2022), 123–159. doi:10.1007/s42001-021-00118-8

[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 675–684. doi:10.1145/1963405.1963500

[8] T. Chai and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 3 (2014), 1247–1250. doi:10.5194/gmd-7-1247-2014

[9] Joydeep Chandra, Ramanjot Kaur, and Rashi Sahay. 2025. Integrated Framework for Equitable Healthcare AI: Bias Mitigation, Community Participation, and Regulatory Governance. In *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*. 819–825. doi:10.1109/CSNT64827.2025.10968102

[10] Joydeep Chandra and Prabal Manhas. 2024. Adversarial Robustness in Optimized LLMs: Defending Against Attacks. *SSRN Electronic Journal* (December 2024). doi:10.2139/ssrn.5116078

[11] Joydeep Chandra, Prabal Manhas, Ramanjot Kaur, and Rashi Sahay. 2025. Aura-CAPTCHA: A Reinforcement Learning and GAN-Enhanced Multi-Modal CAPTCHA System. arXiv:2508.14976 [cs.LG] https://arxiv.org/abs/2508.14976

[12] Joydeep Chandra, Prabal Manhas, Ramanjot Kaur, and Rashi Sahay. 2025. A unified approach to large language model optimization: Methods, metrics, and benchmarks. In *Progressive Computational Intelligence, Information Technology and Networking*. CRC Press, 585–591.

[13] Joydeep Chandra and Satyam Kumar Navneet. 2025. Advancing Responsible Innovation in Agentic AI: A study of Ethical Frameworks for Household Automation. arXiv:2507.15901 [cs.AI] https://arxiv.org/abs/2507.15901

[14] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex Claim Verification with Evidence Retrieved in the Wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3569–3587. doi:10.18653/v1/2024.naacl-long.196

[15] Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. arXiv:2003.07892 [cs.CL] https://arxiv.org/abs/2003.07892

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[17] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. arXiv:2002.06305 [cs.CL] https://arxiv.org/abs/2002.06305

[18] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127. doi:10.1136/amiajnl-2011-000089

[19] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. arXiv:2402.00838 [cs.CL] https://arxiv.org/abs/2402.00838

[20] Andreas Holzinger, Kurt Zatloukal, and Heimo Müller. 2025. Is human oversight to AI systems still possible? *New Biotechnology* 85 (2025), 59–62. doi:10.1016/j.nbt.2024.12.003

[21] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. arXiv:1902.00751 [cs.LG] https://arxiv.org/abs/1902.00751

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[24] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[26] Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. 2024. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. arXiv:2303.15647 [cs.CL] https://arxiv.org/abs/2303.15647

[27] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. doi:10.18653/v1/2022.acl-long.229

[28] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, and Xipeng Qiu. 2024. Full Parameter Fine-tuning for Large Language Models with Limited Resources. arXiv:2306.09782 [cs.CL] https://arxiv.org/abs/2306.09782

[29] Media Bias Fact Check. 2024. Media Bias/Fact Check Methodology. https://mediabiasfactcheck.com/methodology/.

[30] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (Jan. 2021), 41 pages. doi:10.1145/3425780

[31] Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. 2024. Granite Code Models: A Family of Open Foundation Models for Code Intelligence. arXiv:2405.04324 [cs.AI] https://arxiv.org/abs/2405.04324

[32] Satyam Kumar Navneet and Joydeep Chandra. 2025. Rethinking Autonomy: Preventing Failures in AI-Driven Software Engineering. arXiv:2508.11824 [cs.SE] https://arxiv.org/abs/2508.11824

[33] NewsGuard Technologies. 2024. NewsGuard Rating Process and Criteria. https://www.newsguardtech.com/ratings/rating-process-criteria/.

[34] M. F. Porter. 1997. *An algorithm for suffix stripping.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 313–316.

[35] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 231–240. doi:10.18653/v1/P18-1022

[36] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. doi:10.18653/v1/D17-1317

[37] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154. doi:10.1016/j.iotcps.2023.04.003

[38] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 624–643. doi:10.18653/v1/2021.naacl-main.52

[39] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[40] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning Language Models for Factuality. arXiv:2311.08401 [cs.CL] https://arxiv.org/abs/2311.08401

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] https://arxiv.org/abs/2307.09288

[42] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. arXiv:https://www.science.org/doi/pdf/10.1126/science.aap9559 doi:10.1126/science.aap9559

[43] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. doi:10.18653/v1/2020.emnlp-main.609

[44] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Comput. Surv.* 56, 7, Article 185 (April 2024), 33 pages. doi:10.1145/3648471

[45] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. doi:10.18653/v1/P17-2067

[46] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (Oct. 2022), 364–381. doi:10.1016/j.future.2022.05.014

[47] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. arXiv:2502.14768 [cs.CL] https://arxiv.org/abs/2502.14768

[48] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 812, 12 pages.

[49] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385 [cs.CL] https://arxiv.org/abs/2401.02385

[50] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223

[51] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5, Article 109 (Sept. 2020), 40 pages. doi:10.1145/3395046

[52] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large Language Models for Information Retrieval: A Survey. *ACM Trans. Inf. Syst.* 44, 1, Article 12 (Nov. 2025), 54 pages. doi:10.1145/3748304

Received 20 February 2025; revised 12 March 2025; accepted 5 June 2025