Multi-dimensional Hybrid Fact-Checking: Integrating Structured and Unstructured Information to Promote LLMs

Yanqing Liu, Zhong Qian*, Peifeng Li, Qiaoming Zhu
School of Computer Science and Technology, *Soochow University*, Suzhou, China
20235227053@stu.suda.edu.cn, {qianzhong, pfli, qmzhu}@suda.edu.cn

***Abstract*—Dialogue-based fact-checking (DialFC) presents distinct challenges due to contextual dependencies, informal phrasing, and the entanglement of claims within multi-turn exchanges, which conventional verification paradigms struggle to capture. Existing approaches face significant limitations: they often lack explicit reasoning modeling, underutilize structured knowledge, and remain vulnerable to hallucination and factual drift in Large Language Models (LLMs). To address these issues, we propose Multi-Dimensional Hybrid Fact-Checking (MDH-FC), a unified framework that fuses semantic reasoning with structured knowledge modeling. Specifically, our approach integrates (1) a multi-perspective Chain-of-Thought (CoT) prompting strategy to generate interpretable reasoning traces; (2) a graph-based Claim–Evidence Entity Relationship (ERE) module utilizing Graph Convolutional Networks (GCNs) to capture structural consistency; and (3) a Mixture-of-Experts (MoE) network to dynamically filter and fuse heterogeneous reasoning pathways. Comprehensive experiments on the DialFact and HEALTHVER benchmarks demonstrate that MDH-FC effectively mitigates hallucinations and substantially outperforms state-of-the-art baselines in complex conversational settings.**

***Index Terms*—fact-checking, structured knowledge, mixture-of-experts, semantic reasoning**

I. Introduction

Dialogue-based fact-checking (DialFC) verifies claims within conversational contexts by leveraging dialogue history and external knowledge [1]. Unlike traditional verification of isolated claims against static knowledge [2], DialFC operates on claims embedded in dynamic, multi-turn exchanges.

As illustrated in Figure 1, a dialogue-based claim's veracity often depends on implicit context, coreference resolution, and informal phrasing. Consequently, models must resolve complex pragmatic dependencies beyond surface-level semantics to determine if a claim is SUPPORTED (S), REFUTED (R), or represents NOT ENOUGH INFORMATION (NEI).

Despite advancements in evidence retrieval, the core verification stage in dialogue settings remains fragile [3]. Existing methodologies exhibit three persistent limitations. First, they often lack explicit modeling of reasoning processes, relying on opaque end-to-end generation or simple retrieval [4], which limits interpretability and robustness. Second, they underutilize structured evidence; while graph-based methods exist [5], they are rarely integrated into conversational verification, leaving models prone to entity-level errors. Third, hallucination remains a common challenge. Current mitigation strategies

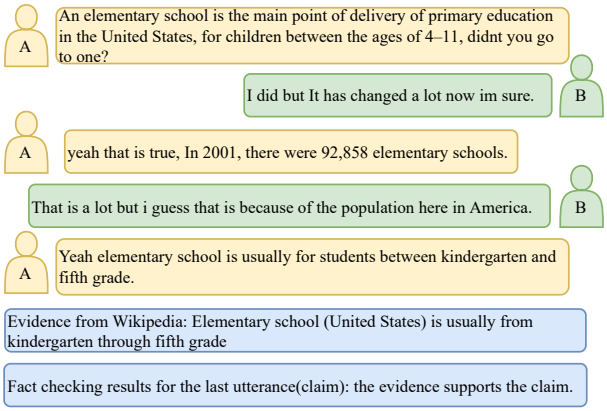


Fig. 1: Example of Dialogue-based fact-checking (DialFC)

typically focus on post-hoc correction or self-verification [6], [7], rather than explicitly enforcing factual consistency and structural constraints during the inference process.

To address these deficiencies, we propose Multi-Dimensional Hybrid Fact-Checking (MDH-FC), a modular framework that fuses structured grounding with logical reasoning. MDH-FC operationalizes a multi-perspective Chain-of-Thought (CoT) strategy to decompose verification into distinct analytical angles (e.g., causal, comparative). To manage the resulting information diversity, a specialized Mixture-of-Experts (MoE) module dynamically aggregates these reasoning traces, filtering noise and amplifying high-confidence signals. Complementing this, a Graph Convolutional Network (GCN) module explicitly models entity-relation consistency between the claim and evidence, providing an orthogonal structural signal. A final stacking fusion mechanism integrates these semantic and structural representations to produce robust verification outcomes. Our code will be available at https://anonymous.4open.science/r/DialFC-D060.

Our primary contributions are: **Novel Hybrid Framework:** We propose MDH-FC, uniquely coupling a multi-perspective CoT-MoE reasoning module with a GCN-based structural analyzer to dynamically filter inconsistent logical traces in parameter-constrained LLMs. **Explicit Hallucination Mitigation:** As our core methodological contribution, we introduce a unified fusion strategy that anchors the LLM's semantic gen-

eration to latent entity–relation graphs, rigorously enforcing cross-dimensional consistency. **SOTA Results on DialFC:** Experiments on the DialFC benchmark show that MDH-FC significantly outperforms existing baselines, establishing a new state-of-the-art (SOTA) in multi-turn dialogue verification.

II. Related Work

Fact-Checking (FC): Traditional automated FC verifies self-contained claims via document retrieval, sentence selection, and claim verification [3], [8]. Large-scale datasets such as FEVER [2] enabled early progress in neural matching and entailment, while recent work emphasizes efficiency and iterative refinement, e.g., FIRE's retrieval–verification loop [9].

Fact-Checking in Dialogue (DialFC): DialFC is challenged by colloquial language, coreference, and unclear claim boundaries. Introduced with the DialFact benchmark [1], later work proposed conversational retrieval strategies [10], multi-task instruction tuning for robustness [11], and dense retrieval reformulations to align colloquial claims with evidence [12].

Fact-Checking Using LLMs: LLM-based FC has rapidly advanced: GPT-4 shows strong performance with sufficient context [13], while reliability is improved via factuality fine-tuning [4] and evidence retrieval [14]. This paradigm extends to dialogue verification [15], with reasoning-focused methods such as CorXFACT [16] and MiniCheck [17], and domain-specific systems like GraphRAG for healthcare [5].

III. Methodology

To address the hallucination, opacity, and lack of structural grounding in DialFC, we propose MDH-FC. As illustrated in Figure 2 and Figure 3, our framework integrates four core components: (1) a multi-perspective CoT prompting module (Section III-A) for diverse reasoning generation; (2) a MoE module (Section III-B) for selective reasoning aggregation; (3) a graph-based Entity Relationship (ERE) module (Section III-C) for structural verification; and (4) a unified Fusion Strategy (Section III-E) that integrates these semantic and structural signals for robust inference.

A. Multi-Perspective Reasoning CoT

We leverage LLMs for dialogue summarization and CoT generation, utilizing their multi-hop inference capabilities [18]. While standard CoT enhances interpretability, it remains prone to single-perspective hallucinations [19]. Thus, we propose a multi-perspective prompting strategy.

1) Prompting Strategy: Our prompt directs the LLM to function as a multi-perspective analyst. As detailed in Figure 4, the user prompt enforces a structured output schema. First, to mitigate the noise introduced by irrelevant chat history, we use a summarization step to extract key dialogue content. Next, the model generates reasoning traces from $n = 6$ distinct perspectives (numerical, causal, comparative, sentiment, temporal, and spatial), selected based on the taxonomy in [20] and triggered by specific lexical cues (Table I). Within each perspective, a two-step "First... Second..." reasoning structure clarifies implicit assumptions. Finally, a synthesis step consolidates these viewpoints into a unified judgment. We incorporate three in-context exemplars to stabilize generation and align outputs with the desired verification logic.

TABLE I: Reasoning Perspective Prompt Words

Reasoning Perspective	Key Words Examples
Number and Ranking Reasoning	first, top three, ranked, one of
Causal Reasoning	because, lead to, result in
Comparison Reasoning	more than, better, worse
Sentiment Reasoning	praised, angry, support
Temporal Reasoning	in 2020, before, after, then
Location or Spatial Reasoning	in the US, near, far, inside, outside

2) Multi-perspective CoT: This framework operationalizes the multi-perspective strategy to capture contradictory or complementary evidence often overlooked by single-path reasoning. Unlike approaches relying on truncated dialogue history [1], we incorporate the full dialogue d, augmented by a model-generated summary, to establish a robust evidential foundation. The two-step inference process, comprising evidence summarization followed by veracious judgment, promotes traceability and mitigates error propagation.

Formally, given an input triple (d, c, e) comprising the dialogue (d), claim (c), and evidence (e), the pipeline elicits a structured set of semantic artifacts: (1) the content summary M; (2) the original single-path reasoning CoT_o; (3) the multi-perspective traces CoT_m with corresponding verifications V_m; and (4) the final synthesized reasoning CoT_s and judgment V_s. These diverse outputs provide the requisite semantic features for the downstream fusion modules.

B. MoE for Structured Reasoning Integration

To aggregate heterogeneous reasoning traces, we implement a MoE module (Figure 3a) designed for feature selection and noise filtering. Let $\mathcal{P} = \{p_1, ..., p_n\}$ denote the set of $n = 6$ reasoning perspectives. For each perspective i, we extract the hidden state $x_i \in \mathbb{R}^d$ of its reasoning trace. The module comprises n expert networks $\{E_i\}_{i=1}^n$, each implemented as a specialized Feed-Forward Network (FFN) to project the perspective-specific features x_i into a unified semantic space.

A gating network G assigns importance weights based on the global context of the claim c and dialogue summary M. The gating score g is computed as:

$$g = \text{Softmax}(W_g \cdot [H_c; H_M] + b_g) \tag{1}$$

where H_c and H_M are the embeddings of the claim and summary, and W_g is a learnable weight matrix. The final aggregated representation P_{moe} is the weighted sum of expert outputs:

$$P_{moe} = \sum_{i=1}^{n} g_i \cdot E_i(x_i) \tag{2}$$

By explicitly routing the representation x_i to its corresponding expert E_i, the gating mechanism g_i acts as a dynamic filter. This prevents reasoning collapse by amplifying relevant reasoning paths while suppressing perspectives that yield low confidence or irrelevant information.

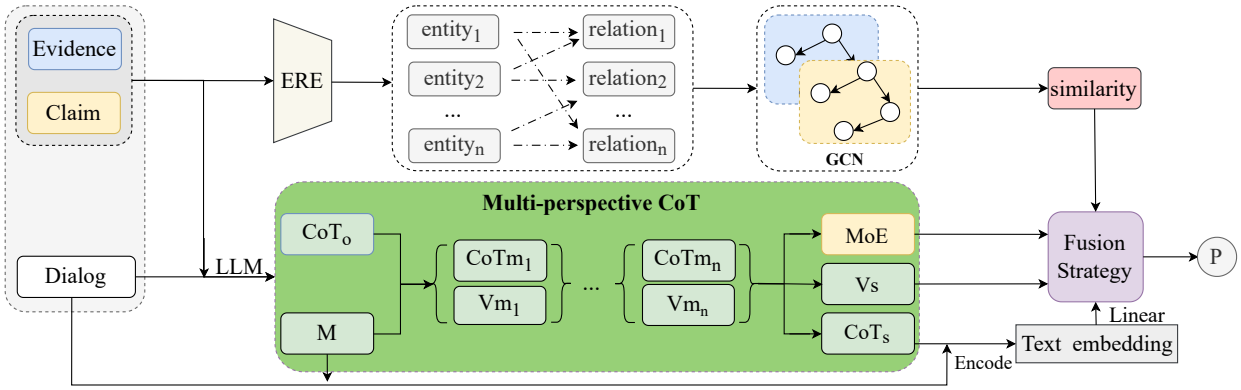
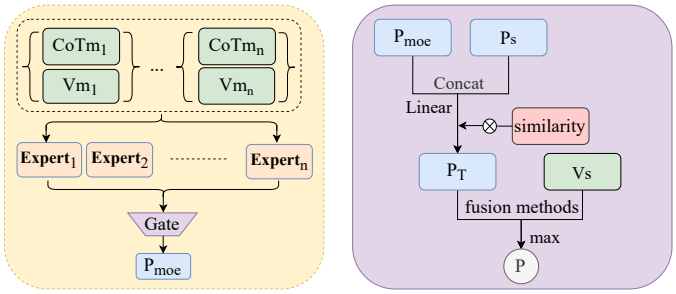


Fig. 2: Overview of the MDH-FC framework. The architecture comprises: (1) a multi-perspective CoT prompting module (Section III-A) producing a summary (M) and verification rationales ($CoT_{(.)}$ and $V_{(.)}$); (2) a graph-based Claim-Evidence Entity Relationship (ERE) module utilizing a GCN (Section III-C); and (3) a MoE fusion mechanism (Section III-B). (4) a unified Fusion Strategy (Section III-E)The final output P represents the predicted veracity label.



(a) MoE (b) Fusion Strategy

Fig. 3: Illustration of major modules in Figure 2. (a) The MoE module (Section III-B) fuses expert representations P_{moe} with the semantic classifier P_s to produce final prediction P, enhancing adaptability and robustness. (b) The Fusion Strategy (Section III-E) integrates P_s, GCN-derived similarity signals, and P_{moe} with the LLM's direct prediction V_s to produce a robust fact verification result.

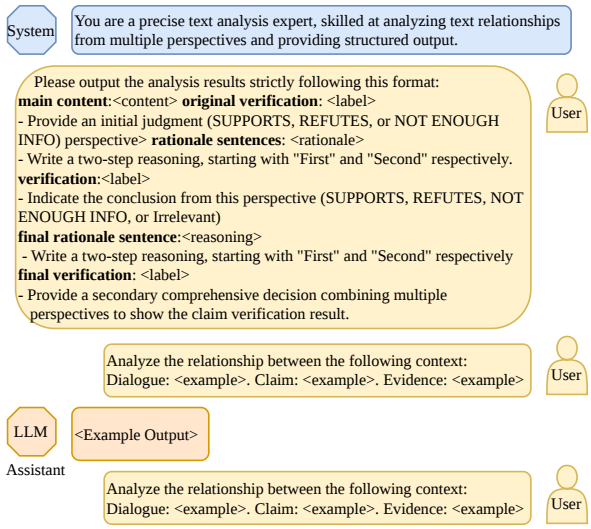


Fig. 4: Multi-Perspective CoT Generation Prompt

C. Claim-Evidence Entity Relationship Modeling

To address the limitations of LLMs in modeling explicit structural dependencies [21], we integrate a graph-based augmentation module. Distinct from approaches utilizing complex graph reasoning networks, our method focuses on metric alignment. We extract entity-relation triples from both the claim and evidence to construct their respective structured graphs, which are then encoded via GCNs to capture topological features.

We employ a contrastive learning objective to align these representations. Let G_c, G_e, and G_n denote the graph embeddings for the claim, positive evidence, and a randomly sampled negative context, respectively. We define the cosine similarities for the positive and negative pairs as $s^+ = sim(G_c, G_e)$ and $s^- = sim(G_c, G_n)$, where $sim(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$. The contrastive loss $\mathcal{L}_{cont}$ optimizes the embedding space by maximizing the relative similarity of valid pairs:

$$\mathcal{L}_{cont} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \exp(s^-/\tau)} \quad (3)$$

where τ is a temperature hyperparameter. The resulting similarity score $S_{struc} = s^+$ serves as a quantifiable measure of structural alignment, providing an explicit feature for the final fusion layer orthogonal to semantic phrasing.

D. Semantic Representation

To mitigate LLM hallucinations, we employ an encoder-only architecture optimized for robust semantic classification rather than generative fluency. We concatenate the dialogue history d, summary M, claim c, retrieved evidence, and synthesized reasoning trace CoT_s into a unified input. Exploiting bidirectional attention, the encoder evaluates the logical alignment between the reasoning steps and the claim, yielding an intermediate probability distribution P_s that captures deep entailment features for the subsequent fusion stage.

E. Final Prediction Fusion

The final verification decision (Figure 3b) integrates four components: (1) the semantic prediction P_s; (2) the structural

TABLE II: Data composition of DialFact and HEALTHVER. S for Supports, R for Refutes, Neu for Neutral in HEALTHVER, NEI for NOT ENOUGH INFORMATION in DialFact.

Dataset	S	R	Neu/NEI
DialFact	7281	7298	7666
HEALTHVER	4986	3227	6117

similarity S_{struc} and contrastive loss $\mathcal{L}_{cont}$; (3) the MoE aggregation P_{moe}; and (4) the LLM's direct judgment V_s.

Training: We employ a joint loss to ensure convergence of the semantic, structural, and expert modules:

$$\mathcal{L} = h_s\mathcal{L}_s + h_g\mathcal{L}_{cont} + h_{moe}\mathcal{L}_{moe} \tag{4}$$

where $\mathcal{L}_s$ and $\mathcal{L}_{moe}$ denote the cross-entropy losses for the semantic classifier and MoE module, respectively, and $h_{(\cdot)}$ are scalar hyperparameters balancing the loss components.

Inference: To reconcile the hallucination-prone semantic prediction (P_s and P_{moe}) with the sparse but precise structural alignment (S_{struc}), we derive an integrated representation P_T. Specifically, the semantic and expert outputs are fused via a linear projection and modulated by the structural score:

$$P_T = \text{Linear}([P_s; P_{moe}]) \cdot (1 + \mathcal{H}_g \cdot S_{struc}) \tag{5}$$

where $\mathcal{H}_g$ is a scaling factor for the graph-based signal. Finally, to produce the prediction P, we employ a Stacking strategy to optimally combine P_T with the LLM judgment V_s:

$$P = \mathcal{F}_{stacking}(P_T, V_s) \tag{6}$$

Empirical experiments indicate that this meta-learning approach outperforms simple baselines (e.g., voting or weighted averaging), which fail to capture the non-linear dependencies between semantic nuances and structural validity, thereby yielding robust verification results.

IV. Experiment

A. Dataset and Experimental Settings

We evaluate our framework on two benchmarks representing distinct conversational complexities: DialFact [1] (balanced), which captures complex multi-turn interactions, and HEALTHVER [22] (imbalanced), which focuses on simple single-turn claim verification (details in Table II). Regarding evidence sources, DialFact relies on Wikipedia passages, whereas HEALTHVER utilizes web search snippets. To ensure rigorous evaluation, we report both Accuracy (Acc) and Macro-F1 (F1). For implementation, we utilize RoBERTa [23] as the semantic encoder and LUKE [24] for entity extraction. We use the AdamW optimizer with a learning rate of $2e^{-5}$ for the encoder and $1e^{-4}$ for the MoE gating network.

B. Baselines

We compare MDH-FC against a diverse set of baselines categorized into standard fine-tuning approaches and recent LLM-based methods.

Fine-tuned Baselines: For DialFact, we evaluate DB-Few [11] (BART0 fine-tuned on InstructDial), Aug-WoW [1] (Colloquial model with concatenated context), and Auto-FC [10] (FEVER-trained models adapted for dialogue). For HEALTHVER, we include SciBERT and T5-base [22], and LI4 [25], employing RoBERTa-large with interactive gating.

TABLE III: Performance on DialFact and HEALTHVER

DialFact			HEALTHVER		
Model	**Acc**	**F1**	**Model**	**Acc**	**F1**
Qwen	42.7	43.0	Qwen	42.0	41.8
LLaMA	47.1	47.0	LLaMA	44.1	43.5
GLM	41.0	41.0	GLM	40.6	40.1
FactDetect	56.6	56.5	FactDetect	76.0	75.8
DB-Few	43.0	43.0	SciBERT	78.1	77.2
Aug-WoW	51.6	51.3	T5-base	80.7	79.6
ERE-GCN	60.7	60.7	ERE-GCN	85.6	85.7
Auto-FC	63.6	63.5	LI4	85.4	84.5
GPT-4-pipeline	72.7	72.6	GPT-4-pipeline	86.5	86.3
Qwen-MDH-FC	**76.0**	**75.9**	**Qwen-MDH-FC**	**89.5**	**89.0**

LLM-based Baselines: We deploy existing LLM-based DialFC frameworks across both datasets. We utilize FactDetect [26], which leverages Llama2-13B to distill factual statements, and the GPT-4-pipeline [15], a strong few-shot CoT baseline. To isolate the contribution of our architecture, we also report the performance of base open-source models (**Qwen2.5-7B, LLaMA3.1-8B, GLM4-9B**) using standard prompting. We report the best-performing configuration (Qwen-MDH-FC) in the main results.

C. Results and Analysis

As reported in Table III, our proposed MDH-FC framework establishes a new SOTA across both benchmarks. It achieves **76.0%** accuracy on the complex, multi-turn DialFact (surpassing GPT-4 by 3.3%) and **89.5%** on the simple, single-turn HEALTHVER, outperforming specialized baselines. These improvements indicate that our architecture provides robust gains over standard CoT prompting and distillation methods.

1) Efficacy of Structural Modeling: Table III highlights the efficacy of the standalone ERE-GCN module. On HEALTHVER, it achieves 85.6% accuracy, comparable to the GPT-4 pipeline (86.5%) and superior to FactDetect (76.0%). This performance stems from the single-turn nature of the dataset where verification relies on direct entity alignment. In this context, explicit structural matching captures veracity signals efficiently. Conversely, on DialFact, the ERE-GCN accuracy decreases to 60.7%. Although structural signals alone are insufficient for complex dialogues, this performance still exceeds baselines like Aug-WoW (51.6%). This confirms structural modeling provides necessary grounding for entity consistency often missed by purely semantic approaches.

2) Comparative Analysis against Baselines: The results highlight distinct baseline failure modes that MDH-FC successfully mitigates:

Limitations of Evidence Distillation (FactDetect): While FactDetect (using LLaMA-13B) achieves a competitive 76.0% on the simpler HEALTHVER, its performance degrades to 56.6% on DialFact. This 19.4% deficit relative to MDH-FC indicates that static evidence distillation is insufficient for

TABLE IV: Ablation study on DialFact and HealthVer. "Origin" denotes base LLM performance. "+CoT" includes multi-perspective prompting. "+MoE" adds expert routing module. "+GCN" adds the graph-based entity relation module. "MDH-FC" represents the full framework with stacking fusion.

DialFact (Acc / F1)

Configuration	Qwen	LLaMA	GLM
Origin	42.7 / 41.9	47.1 / 44.6	41.0 / 40.1
+CoT	63.2 / 61.0	62.8 / 60.4	58.2 / 56.0
+CoT+MOE	68.8 / 67.1	67.6 / 65.9	66.4 / 64.6
+CoT+GCN	69.8 / 68.2	70.0 / 68.5	65.0 / 63.4
+CoT+GCN+MOE	72.2 / 70.6	72.1 / 70.4	69.8 / 68.0
MDH-FC	**76.0 / 75.9**	**74.5 / 72.9**	**71.3 / 69.6**

HealthVer (Acc / F1)

Configuration	Qwen	LLaMA	GLM
Origin	41.2 / 41.8	43.6 / 44.1	40.1 / 40.6
+CoT	47.9 / 48.5	54.7 / 55.3	51.1 / 51.7
+CoT+MOE	85.6 / 86.3	85.4 / 86.2	85.5 / 86.3
+CoT+GCN	85.8 / 86.2	86.3 / 86.9	85.1 / 85.7
+CoT+GCN+MOE	86.7 / 87.4	86.8 / 87.4	86.5 / 87.2
MDH-FC	**89.5 / 89.0**	**87.6 / 88.2**	**87.5 / 88.2**

scattered multi-turn contexts. Conversely, MDH-FC's MoE module dynamically integrates dialogue history, proving essential for complex conversational reasoning.

MDH-FC vs. GPT-4 Pipeline: Comparing our Qwen-MDH-FC against the vastly larger GPT-4 pipeline (72.7% Acc on DialFact) underscores the efficacy of architectural constraints. By outperforming GPT-4 by 3.3%, MDH-FC demonstrates that relying solely on massive parametric knowledge is suboptimal for rigorous fact-checking. While unconstrained LLMs remain susceptible to probabilistic logical drift, MDH-FC anchors predictions with explicit ERE-GCN scores, enforcing structural consistency to filter flawed reasoning traces.

3) Effectiveness of Component Integration: The results confirm the necessity of the proposed integration strategy. On DialFact, the full MDH-FC framework (76.0%) significantly outperforms both the standalone structural module (ERE-GCN at 60.7%) and strong semantic baselines (GPT-4 at 72.7%). This performance gain validates the framework's hierarchical design. First, the MoE module aggregates diverse CoT perspectives to produce a stabilized reasoning representation. Second, the final fusion mechanism integrates this hybrid output (combining MoE and GCN signals) with the LLM's direct verification judgment. This two-stage process ensures the system effectively leverages entity consistency and refined reasoning features to correct errors in the LLM's direct prediction, covering the limitations of individual modalities.

D. Ablation Study

Table IV details the systematic evaluation. We leverage these results to quantitatively address the claims of hallucination reduction and interpretability.

Hallucination Reduction. A key contribution of this work is mitigating hallucination. The "Origin" rows serve as a quantitative baseline for this phenomenon: without constraints, base LLMs achieve only 42-47% accuracy on DialFact, frequently hallucinating contradictions or missing context. The introduction of Multi-Perspective CoT (+CoT) reduces this error rate significantly (e.g., Qwen +20.5% Acc). Furthermore, the full MDH-FC framework (Qwen) achieves 76.0% accuracy. The **33.3% gap** between the hallucination-prone Origin model and MDH-FC serves as a direct quantitative indicator of improved factual consistency and suppressed hallucination.

Interpretability via MoE. While interpretability is often qualitative, the performance delta provided by the MoE module offers a quantitative proxy for reasoning reliability. On DialFact, adding the MoE (+CoT+MoE) improves Qwen's accuracy from 63.2% to 68.8%. This 5.6% gain confirms that the gating mechanism is not merely passing information but actively filtering noise and selecting valid reasoning paths. By suppressing irrelevant perspectives (as verified in manual inspection), the MoE ensures that the final decision is derived from high-confidence logic, thereby quantifying the improvement in the interpretability of the inference process.

Contribution of Structured Signals (GCN). Integrating the GCN module (+CoT+GCN) mitigates the shortcomings of purely semantic reasoning. This effect is most pronounced on HEALTHVER, where augmenting the LLaMA backbone increases F1 from 55.3% to 86.9%. Aligning with the standalone GCN results (Table IV), this gain confirms that entity-relation matching provides a high-fidelity verification signal in precision-critical domains. On DialFact, the improvements are smaller yet consistent, proving that structural cues supply essential grounding even amidst complex semantic contexts.

Synergy and Final Fusion. The full integration (+CoT+GCN+MoE) consistently outperforms isolated modules, confirming the synergy between semantic reasoning and structural grounding. The final stacking fusion (MDH-FC) secures the remaining performance gain (e.g., +3.8% Acc for Qwen on DialFact) by optimizing the balance between the LLM's direct judgment (V_s) and the hybrid representation (P_T) to rectify residual errors.

E. Case study

Figure 5 illustrates how MDH-FC mitigates reasoning drift through component synergy. The Original LLM incorrectly infers that pitch-accent" negates the explicit evidence of simple phonotactics," resulting in an erroneous *Refutes* prediction.

The proposed framework initiates the verification process with the **Main Content** summary. This component distinguishes the claim's objective linguistic facts from the user's subjective background and effectively filters contextual noise. Subsequently, the **Multi-Perspective CoT** decomposes the reasoning process. While the *Comparison* perspective correctly identifies the direct textual alignment (*Supports*), the *Causal* perspective reproduces the baseline's hallucination by inferring complexity from phonological features (*Refutes*).

To resolve this conflict, the **ERE-GCN** module provides a decisive structural signal. It computes a high graph similarity (0.92) and confirms that the entity "simple phonotactics" is explicitly grounded in the evidence graph. Guided by this

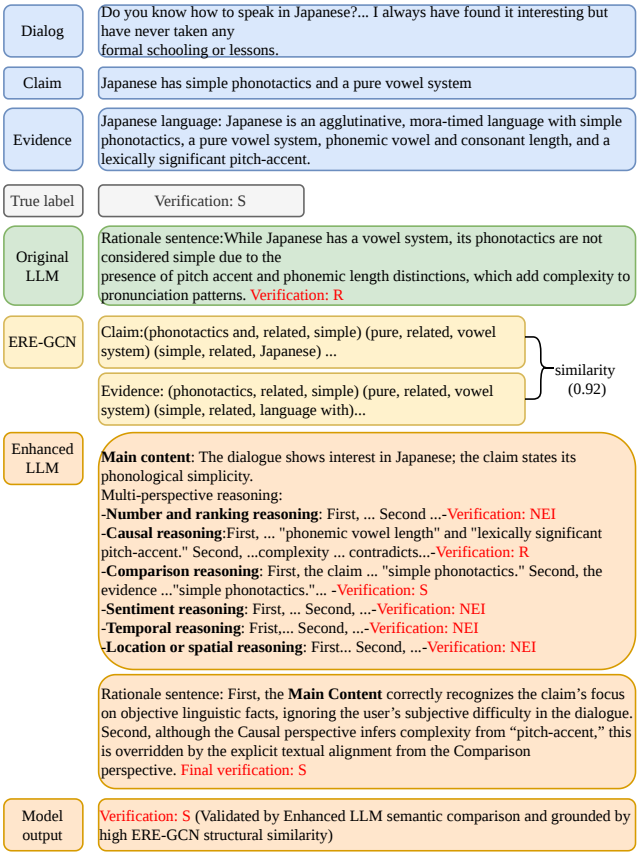


Fig. 5: Case study

objective structural anchor, the **Fusion** mechanism overrides the erroneous Causal inference in favor of the Comparison perspective. This correction demonstrates the framework's capacity to suppress semantic hallucinations by enforcing consistency through structured entity relationship modeling.

V. Conclusion

We introduced MDH-FC, a modular framework for DialFC that integrates multi-perspective CoT prompting, GCN-based relational encoding, and MoE-driven expert specialization. Designed to augment LLM reliability, this hybrid architecture systematically addresses key challenges such as reasoning opacity, structural grounding deficits, and hallucination. Empirical evaluations on the DialFact and HEALTHVER benchmarks demonstrate that MDH-FC achieves SOTA performance. Beyond accuracy gains, the framework enhances transparency by providing auditable reasoning traces and explicit structural verification signals.

Acknowledgment

The authors would like to thank all the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177, 62006167 and 62376181), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Nos. 24KJB520036 and 25KJA521001), Key R & D Plan of Jiangsu Province (BE2021048), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] P. Gupta, C.-S. Wu, W. Liu, and C. Xiong, "DialFact: A benchmark for fact-checking in dialogue," in *ACL*, 2022.
[2] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *NAACL*, 2018.
[3] S. Yang, X. Yuan, T. Gan, and Y. Wu, "A survey of automatic fact verification research," in *WCCCT*, 2024.
[4] K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn, "Fine-tuning language models for factuality," in *ICLR*, 2024.
[5] C. N. Hang, P.-D. Yu, and C. W. Tan, "Trumorgpt: Graph-based retrieval-augmented large language model for fact-checking," *IEEE Transactions on Artificial Intelligence*, 2025.
[6] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan *et al.*, "RARR: Researching and revising what language models say, using language models," in *ACL*, 2023.
[7] Y. Zha, Y. Yang, R. Li, and Z. Hu, "AlignScore: Evaluating factual consistency with a unified alignment function," in *ACL*, 2023.
[8] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *TACL*, 2022.
[9] Z. Xie, R. Xing, Y. Wang, J. Geng, H. Iqbal, D. Sahnan *et al.*, "FIRE: Fact-checking with iterative retrieval and verification," in *Findings of NAACL*, 2025.
[10] E. Chamoun, M. Saeidi, and A. Vlachos, "Automated fact-checking in dialogue: Are specialized models needed?" in *EMNLP*, 2023.
[11] P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. Bigham, "InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning," in *EMNLP*, 2022.
[12] S. Han, F. P. Gomez, T. Vu, Z. Li, D. Cer, H. Zeng *et al.*, "Ateb: Evaluating and improving advanced nlp tasks for text embedding models," 2025. [Online]. Available: https://arxiv.org/abs/2502.16766
[13] D. Quelle and A. Bovet, "The perils and promises of fact-checking with large language models," *Frontiers in Artificial Intelligence*, 2024.
[14] T. Cheung and K. Lam, "Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking," in *APSIPA ASC*, 2023.
[15] A. Rangapur and A. Rangapur, "The battle of llms: A comparative study in conversational qa tasks," 2024. [Online]. Available: https://arxiv.org/abs/2405.18344
[16] X. Tan, B. Zou, and A. T. Aw, "Improving explainable fact-checking with claim-evidence correlations," in *COLING*, 2025.
[17] L. Tang, P. Laban, and G. Durrett, "Minicheck: Efficient fact-checking of llms on grounding documents," 2024.
[18] G. Yona, R. Aharoni, and M. Geva, "Can large language models faithfully express their intrinsic uncertainty in words?" in *EMNLP*, 2024.
[19] S. Das, S. Saha, and R. Srihari, "Diving deep into modes of fact hallucinations in dialogue systems," in *Findings of EMNLP*, 2022.
[20] L. Cui, Y. Wu, S. Liu, Y. Zhang, and M. Zhou, "MuTual: A dataset for multi-turn dialogue reasoning," in *ACL*, 2020.
[21] Y. Zhang, L. Kong, S. Tian, H. Fei, C. Xiang, H. Wang *et al.*, "Multi-view counterfactual contrastive learning for fact-checking fake news detection," in *ICMR*, 2024.
[22] M. Sarrouti, A. Ben Abacha, Y. Mrabet, and D. Demner-Fushman, "Evidence-based fact-checking of health-related claims," in *Findings of EMNLP*, 2021.
[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692
[24] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *EMNLP*, 2020.
[25] X. Zhang, W. Chang, Z. Guoping, and S. Xiaohong, "LI4: Label-infused iterative information interacting based fact verification in question-answering dialogue," in *LREC*, 2024.
[26] N. Jafari and J. Allan, "Robust claim verification through fact detection," *ArXiv*, vol. abs/2407.18367, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271516103