Mihail Popescu · University of Missouri · popescum@missouri.edu

Introduction

Large language models (LLMs) are increasingly used for information extraction, biomedical concept normalization, summarization, question answering, and decision support.

Despite strong performance, they remain difficult to deploy reliably due to their output variability across repeated queries, disagreement across models, and instability under prompt perturbation. One approach to improve reliability is to use **ensemble methods**.

This paper develops a hierarchical uncertainty framework that integrates **Dempster–Shafer (DS) evidence theory** with **Interval Type-3 fuzzy sets** to model structured, multi-layer uncertainty in LLM ensembles.

1 Within-model variability
A single model may give different answers for the same repeated prompt.

2 Between-model disagreement
Different LLMs encode different priors, training corpora, and reasoning behaviours.

3 Prompt-level meta-uncertainty
Prompt phrasing alters the semantic framing and stability of the answer.

Problem Formulation

Consider a proposition ***H*** (e.g., "term t exists in UMLS" or "gene A up-regulates gene B"). Let L_i, i = 1…M, denote an ensemble of M LLMs; let P_r, r = 1…R, denote prompt R configurations.

Under each model–prompt condition, the proposition is queried ***k*** times, producing labels from **A = {Yes, No, Unsure}**

Goal: Produce a structured uncertainty representation for the truth of *H* that captures:

1. Within-model response variation over *k* repeated trials
2. Between-model disagreement under a fixed prompt P_r
3. Between-prompt meta-uncertainty about the resulting belief interval

Framework Overview

LLM Responses
k queries per model–prompt pair → $\{n_y, n_o, n_u\}$

DS Mass Functions
$m_{ir}(Y/N/U)$ per model i, prompt r

DS Fusion (per prompt)
Combine across M models → $I_r = [Bel_r, Pl_r]$

Reliability Weights
W_r = interval width → $R_r = 1 - W_r$ → w_r

Type-3 Aggregation
Weighted avg across R prompts → $I^n(Y)$

Accept / Reject
Narrow I^n → Accept | Wide I^n → Reject

Evidence Representation from LLM Responses

For a fixed model L_i, prompt P_r, let n_y, n_o, n_u be the counts of Yes, No, Unsure in k repeated trials. The DS mass function is:

$$m_{ir}(Y) = n_y/k, \quad m_{ir}(N) = n_o/k, \quad m_{ir}(U) = n_u/k$$

This separates **support**, **contradiction**, and **ignorance** directly. Evidence across models is combined iteratively with Dempster's rule to obtain a fused mass function m^*_r. The fused belief and plausibility for proposition Y are:

$$Bel_r(Y) = m^*_r(Y), \quad Pl_r(Y) = m^*_r(Y) + m^*_r(U)$$

Prompt P_r produces a **belief–plausibility interval** equivalent to a Type-2 fuzzy membership:

$$I_r(Y) = [\, Bel_r(Y),\ Pl_r(Y) \,]$$

Stacking these intervals across R prompts creates the higher-order (Type-3) uncertainty object — the core of our framework.

Type-3 Fusion Operator

Definition 1 (Weighted Type-3 Fusion Operator)

Given prompt-level intervals $I_r(Y) = [Bel_r(Y), Pl_r(Y)]$ and weights w_r:

$$I^n(Y) = [\, \Sigma\, w_r\, Bel_r(Y),\ \Sigma\, w_r\, Pl_r(Y) \,]$$

Reliability Weighting

Width: $W_r = Pl_r(Y) - Bel_r(Y)$
Reliability: $R_r = 1 - W_r$
Norm. weight: $w_r = R_r / \Sigma R_j$

Narrower intervals (more stable prompts) receive higher weight. A case is **accepted** when the Type-3 interval is sufficiently narrow; otherwise **rejected (abstain)**.

Experiments & Results

Three local models were evaluated $M = 3$ LLMs: **MedGemma, LLaMA 3.2, Mistral 3B**. For each item, R = **4 prompt formulations** × k = **10 repeated queries** were used. DS masses were fused per prompt, then aggregated with the Interval Type-3 operator.

Table I: UMLS Concept Verification

Method	Acc	Prec	Rec	F1	Cvg	Rej
MedGemma	0.90	0.83	1.00	0.90	1.00	0
LLaMA 3.2	0.43	0.46	0.82	0.59	1.00	0
Mistral 3B	0.98	0.99	0.97	0.98	1.00	0
DS+T3 (ours)	1.00	1.00	1.00	1.00	0.84	31

Table II: Gene–Gene Interaction Verification

Method	Acc	Prec	Rec	F1	Cvg	Rej
MedGemma	0.58	0.54	0.99	0.70	1.00	0
LLaMA 3.2	0.49	0.49	1.00	0.66	1.00	0
Mistral 3B	0.49	0.49	1.00	0.66	1.00	0
DS+T3 (ours)	1.00	1.00	1.00	1.00	0.07	182

Cross-Dataset Interpretation

UMLS Verification

Coverage: 84.5%
31 abstentions

High-coverage reliability enhancer

Framework accepts most items, abstains on prompt-sensitive or conflict-heavy cases.

Gene–Gene Interaction

Coverage: 7.6%
182 abstentions

High-precision uncertainty filter

All models over-predict. Framework becomes extremely conservative to preserve perfect precision.

Conclusions

Hierarchical fusion: DS mass functions capture within-model and between-model uncertainty; Type-3 intervals capture between-prompt meta-uncertainty.

Adaptive behavior: The framework operates as a high-coverage enhancer (UMLS) or high-precision filter (gene–gene) depending on task difficulty — automatically.

Perfect classification: Both tasks achieve Acc/Prec/Rec/F1 = 1.000 on accepted items, at the cost of selective abstention.

Key insight: Higher-order evidence fusion should be judged by the appropriateness of its confidence structure, not solely by raw accuracy.

References

[1] Banerjee et al. "LLMs Will Always Hallucinate." ArXiv 2409.05746, 2024.
[2] G. Shafer, A Mathematical Theory of Evidence. Princeton Univ. Press, 1976.
[3] J. M. Mendel, Uncertain Rule-Based Fuzzy Systems. Prentice Hall, 2001.
[4] Mendel & John, "Type-2 fuzzy sets made simple," IEEE Trans. Fuzzy Syst., 2002.
[7] P. Smets, "Combination of evidence in the TBM," IEEE TPAMI, 1990.
[8] Sellergren et al., "MedGemma Technical Report," arXiv 2507.05201, 2025.
[9] Grattafiori et al., "The Llama 3 Herd of Models," arXiv 2407.21783, 2024.
[10] Jiang et al., "Mistral 7B," arXiv 2310.06825, 2023.