Semantic–Epistemic Inflation Ratio Framework for LLM Truth Evaluation

This document compiles the conceptual and quantitative framework developed from the observation that the ratio between Epistemic Truth (ET) and Semantic Truth (ST) scores may itself function as a discriminator between semantically true and semantically false outputs generated by Large Language Models (LLMs). The key observation was: **“In Table 3, if you take the ratio of credible (epistemic) to factual (semantic), the ratios above 1.6 are false semantically, while the ratios below 1.6 are true semantically (except for one point).”** This observation motivated a transformed geometric interpretation of hallucination and semantic calibration.

Condition	Semantic Truth (ST)	Epistemic Truth (ET)	ET/ST Ratio
BBC True	2.62	4.81	1.84
BBC Fake	1.19	3.27	2.75
CNN Newer True	3.19	4.93	1.55
CNN Newer Fake	3.01	4.87	1.62
CNN Older True	4.94	4.98	1.01
CNN Older Fake	1.19	3.63	3.05
Historical True	4.76	4.88	1.03
Historical Fake	1.02	3.06	3.00

Methodological Transformation

The revised framework introduces the Epistemic Inflation Ratio: R = ET / ST where ET represents the Epistemic Truth Score and ST represents the Semantic Truth Score. This transformation converts the original truth-evaluation problem into a geometric calibration problem. In the original framework, fabricated texts frequently received moderately high semantic and epistemic scores simultaneously, making classification ambiguous. However, plotting Semantic Truth against the ET/ST ratio reveals a strong separation between semantically true and semantically false outputs. Authentic texts tend to cluster near ET/ST ≈ 1, indicating alignment between factual correspondence and perceived credibility. Fabricated texts tend to exhibit ET/ST >> 1, indicating that the text appears substantially more credible than it is factually grounded. This reframes hallucination not merely as low semantic truth, but as disproportionate epistemic inflation relative to semantic grounding.

Quadrant Interpretation

Using the transformed coordinate system, the framework can be reformulated into a quadrant-based diagnostic space. The horizontal axis represents Semantic Truth (ST), while the vertical axis represents the Epistemic Inflation Ratio (ET/ST). Quadrant I: High ST, Low ET/ST → aligned and trustworthy outputs.
Quadrant II: Low ST, High ET/ST → hallucinations, deceptive plausibility, and fluent misinformation.
Quadrant III: Low ST, Low ET/ST → obvious falsehoods or recognized nonsense.
Quadrant IV: High ST, High ET/ST → novel, unfamiliar, or underrecognized truths. This transformed space produces substantially clearer separation between authentic and fabricated texts than the original ST-versus-ET space.

Conceptual Novelty

The novelty of this framework lies in converting LLM truth evaluation from a scalar scoring problem into a geometric inflation problem. Rather than asking only whether a model assigns high or low truth scores, the framework evaluates whether the model’s epistemic confidence is appropriately proportional to its semantic grounding. The resulting

methodology provides a potentially generalizable metric for: - hallucination detection, - calibration analysis, - semantic–epistemic divergence measurement, - misinformation identification, - and evaluation of overconfident LLM outputs. This suggests that hallucinations are fundamentally characterized not simply by semantic falsehood, but by epistemic inflation under semantic collapse.

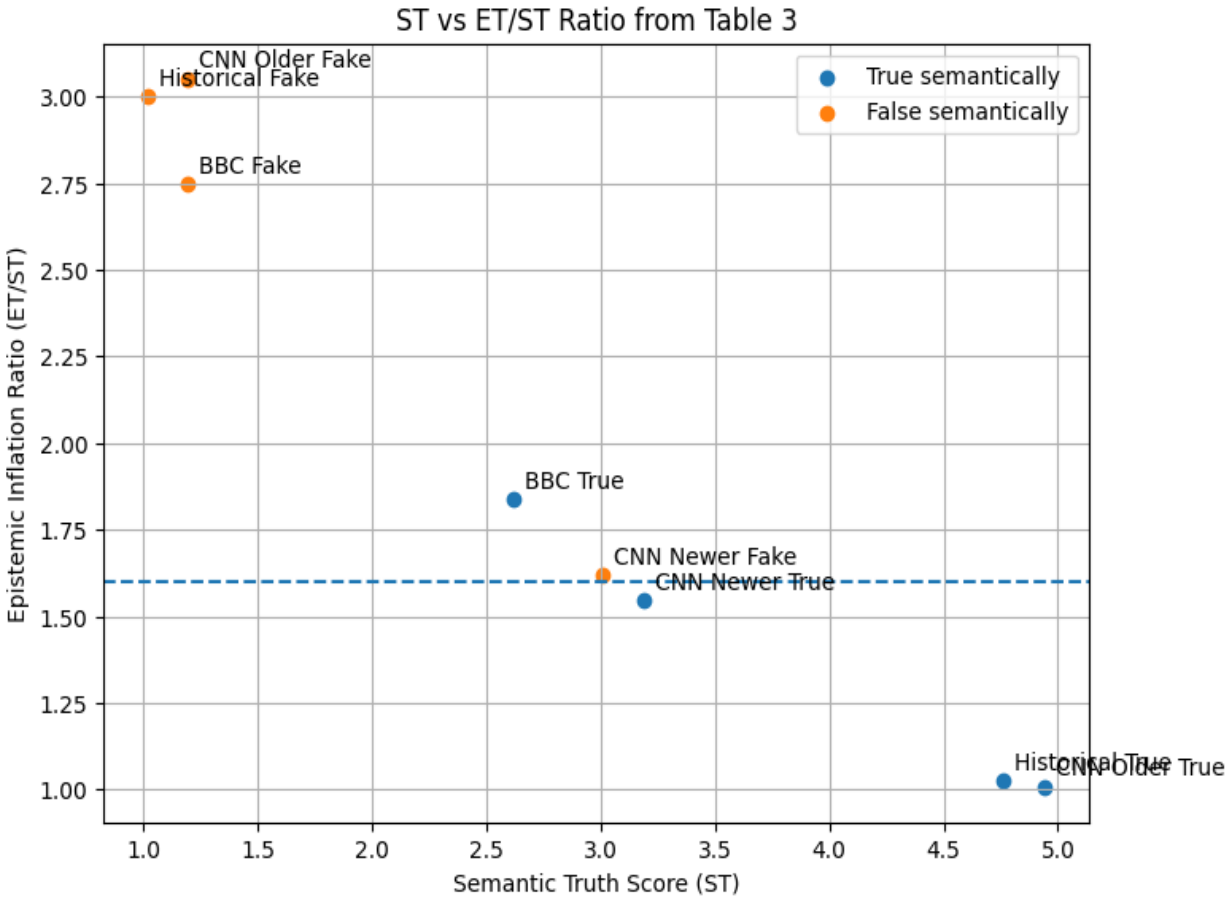

Figure 1. Original ST versus ET/ST ratio plot.

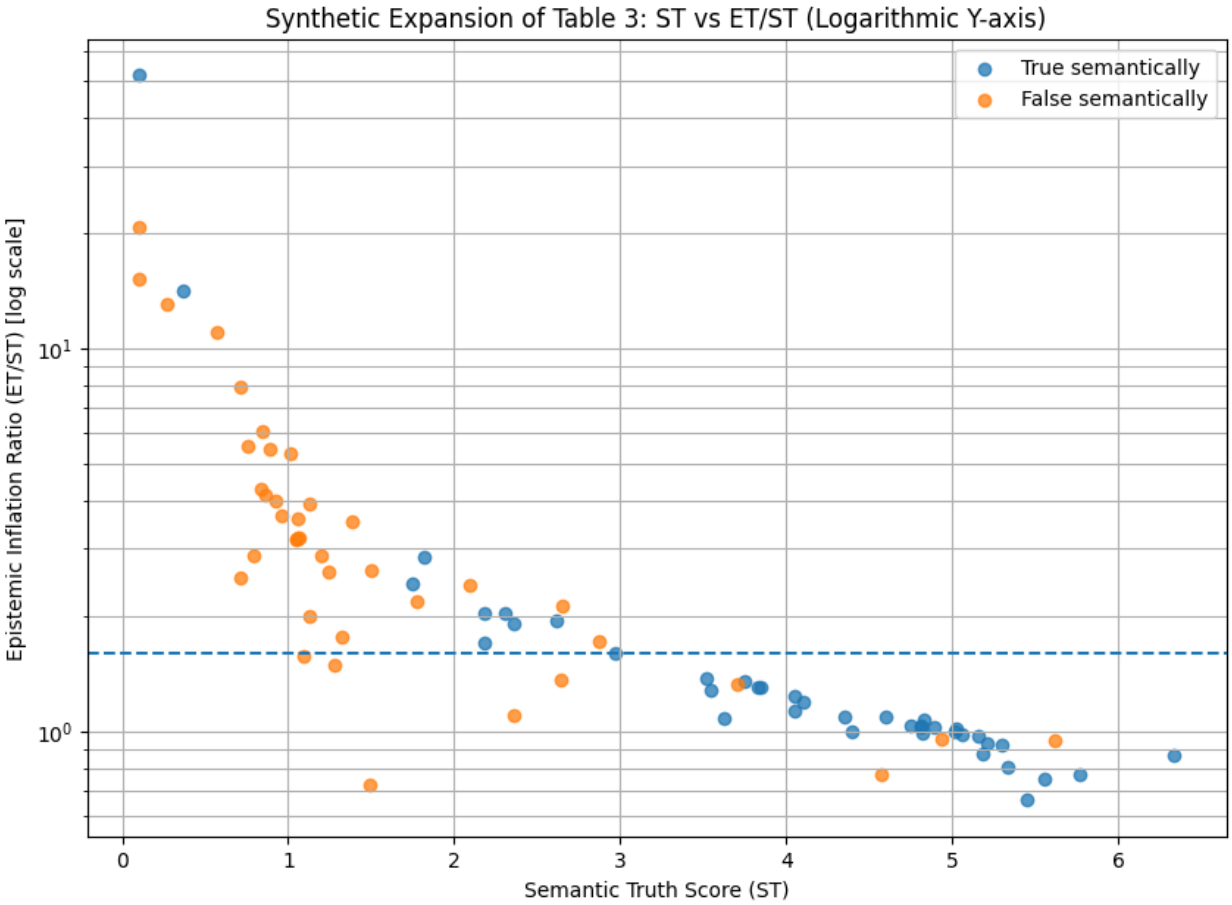

Figure 2. Logarithmic ST versus ET/ST ratio plot.

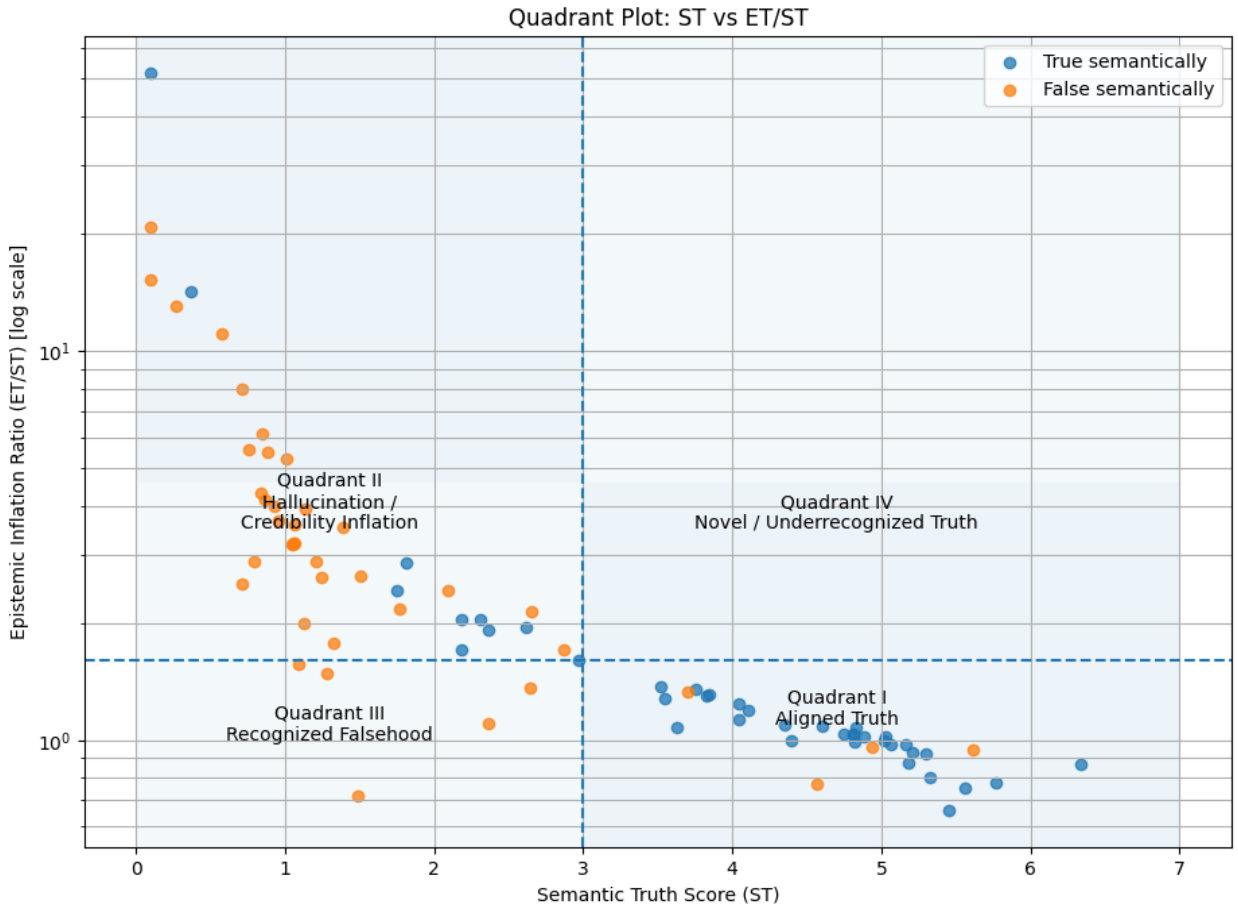

Figure 3. Quadrant interpretation of ST versus ET/ST space.

Conclusion

The ET/ST transformation reveals that semantically false LLM outputs tend to exhibit elevated epistemic inflation ratios, while semantically true outputs remain close to semantic–epistemic alignment. This creates a geometric structure in which hallucinations appear as upward inflation trajectories under conditions of semantic weakness. The framework therefore provides both a conceptual and quantitative basis for understanding hallucinations as overconfident semantic–epistemic divergence rather than merely binary factual failure.