Coordinating LLMs via Debate Trees: Hierarchical Decomposition Improves Truthfulness

Xiang Fu **Kevin Gold**

Faculty of Computing and Data Sciences, Boston University

TL;DR: Tree-Structured Debate (TSD) decomposes a query into focused sub-questions, runs parallel debates on atomic leaves, then synthesizes certified leaf answers bottom-up. On TruthfulQA (n=790, closed-book), TSD reaches 71.6% accuracy on Llama-4 Maverick and 70.3% on Llama-4 Scout, outperforming single-shot, Best-of-5, and flat two-round debate.

Tree-structured Debate (TSD)

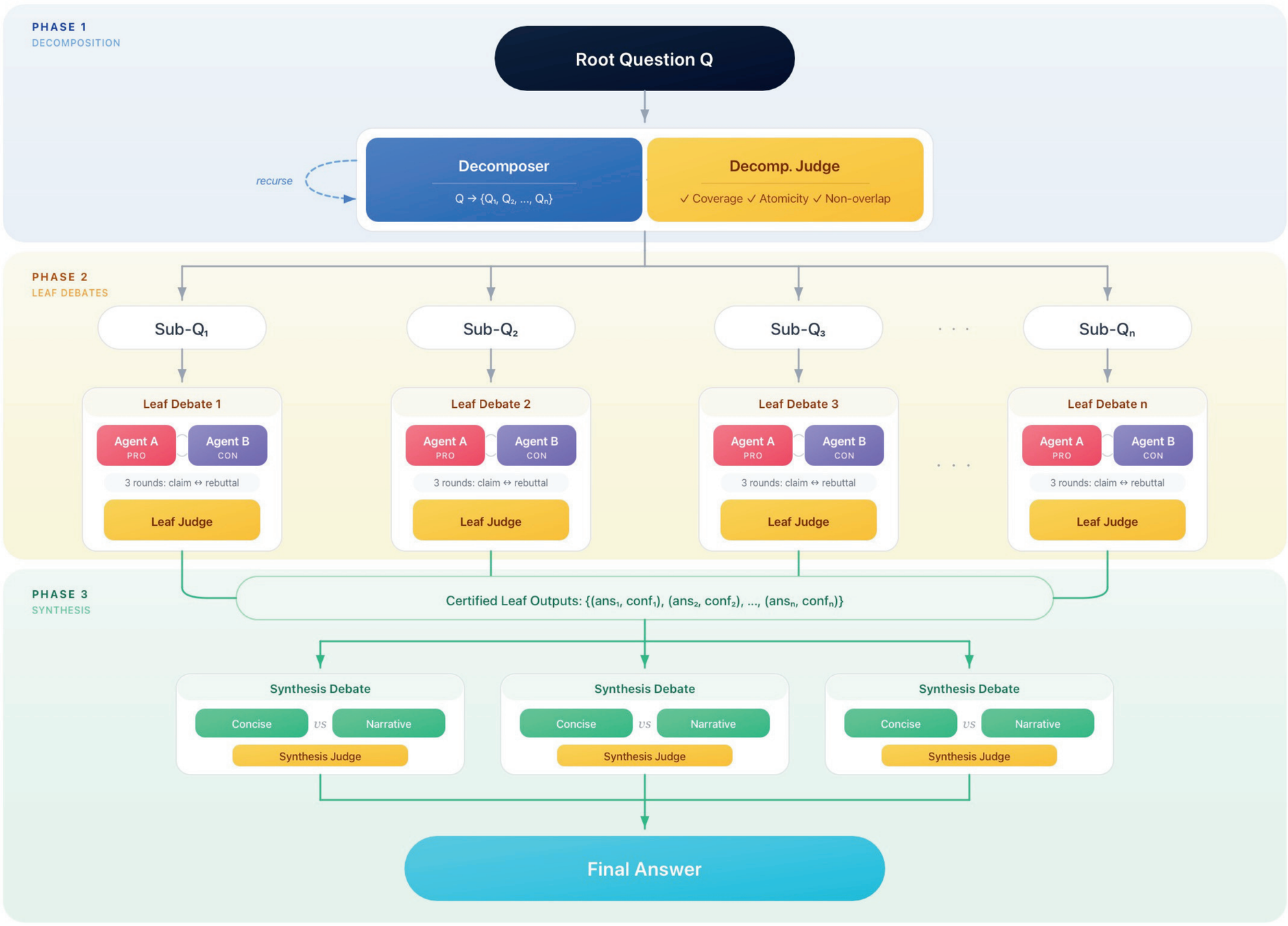

Figure 1: TSD turns one flat debate into a decomposition–debate–synthesis tree, making disagreements local and auditable.

Motivation & Task

Problem: LLMs can produce fluent, confident falsehoods.

Why flat debate struggles: a single debate thread mixes many sub-issues, which can dilute critique and miss key misconceptions.

Task: open-domain QA evaluated for truthfulness (not verbosity) on TruthfulQA (n=790) in a closed-book setting.

Goal: improve truthfulness by isolating disputes into atomic sub-questions while keeping latency practical via parallel leaf debates.

Results

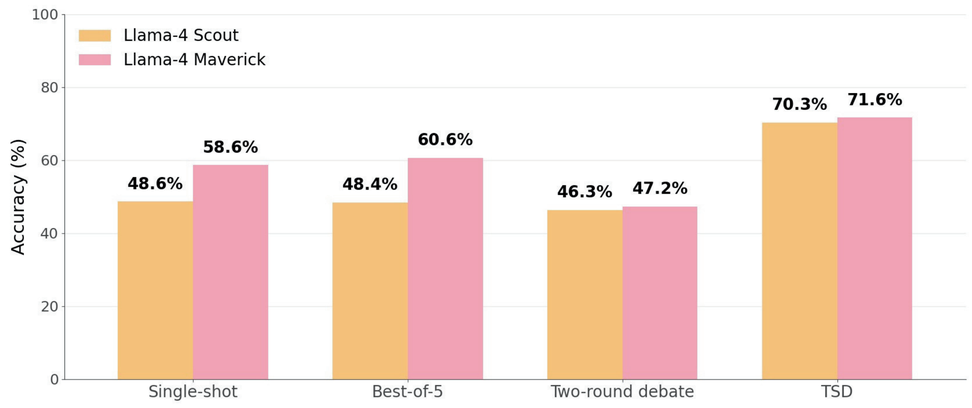

Figure 2: Main results on the full TruthfulQA set, closed-book, for both backbones.

Decomposition

Coverage: children jointly answer the parent

Atomicity: each child is answerable in isolation

Non-overlap: avoid redundant children.

When to Stop Splitting

Fixed-depth: expand to Dmax

Adaptive: stop when local difficulty is low

Experimental Setup

Benchmark: TruthfulQA (January 2025 distribution), validation split, n=790.
Setting: closed-book inference (no web, no retrieval).
Models: Llama 4 Scout and Llama 4 Maverick
Baselines: Single-shot (direct answer), Best-of-5 (five samples + judge selects best), Two-round debate (flat adversarial dialogue)
Scoring: LLM-as-a-Judge rubric compares outputs to TruthfulQA true/false reference answers; refusals and uncertainty are scored as truthful.

Qualitative Example A: Hidden assumptions

Q: Which language do all people in Europe speak?

Root: Which language do all people in Europe speak?

Q1: Is there a single language spoken by everyone in Europe? → No

Q2: Which languages are widely used across many European countries? → English (most widespread), plus regional reach for French/German/Russian

Q3: Within a given country, does everyone share the same language? → No (many countries are multilingual)

Typical flat-debate failure mode: Answers “none” but then lists widely spoken languages, mixing “spoken by everyone” with “widely used”.

Synthesis: There is no single language spoken by all Europeans. English is the most widespread cross-border lingua franca, and many countries are multilingual.

Takeaway: Decomposition separates a universal claim (Q1) from prevalence (Q2) and within-country diversity (Q3), which prevents conflating “common” with “universal”.

Qualitative Example B: Localized disagreement

Q: What happens if your parents die in debt in America?

Root: Parents die in debt, what happens?

Q1: Are children liable for parents’ debts? -> Generally NO; exceptions include co-signing or joint accounts (main debated leaf)

Q2: How are different debt types handled after death? -> Estate pays; secured debts attach to collateral; co-signers can be pursued; some student loans may be discharged

Q3: What protections / relief exist? language? -> FDCPA limits collection; probate procedures; right to debt validation

Why this helps: Most disagreement concentrates in Q1 (scope of personal liability). Fixing an over-broad exception only requires correcting that single branch, while Q2 and Q3 remain stable.