A pipeline for disinformation annotation using complementary detectors

Martin Pavlíček[1,2,*], Tomáš Filip[1,2] and Petr Sosík[1]

[1]Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, 30. dubna 22, 701 03 Ostrava, Czech Republic
[2]Department of Informatics and Computers, Faculty of Science, University of Ostrava, 30. dubna 22, 702 00 Ostrava, Czech Republic

Abstract

A false claim can take many surface forms: sensational clickbait, a sober fabrication wrong about a single fact, fluent machine-generated text, or a paraphrase of a lie debunked years ago. Each form exposes a different signal, and a detector tuned to one form can miss another. We describe the Cognitive Disinformation Annotation (CDA) pipeline, which rests on two design choices. First, we select detectors for different failure modes, not for standalone accuracy alone. Second, the pipeline produces an *annotation* rather than a binary label: for each claim it returns a verdict (supported, refuted, or not enough information), a confidence, cited evidence, and a short rationale. The pipeline splits text into atomic claims, filters them by check-worthiness, matches them against a closed-world store of known false claims, scores the survivors with five detectors (a stylometric gradient-boosting model, a fine-tuned semantic encoder, a few-shot large language model, a retrieval-augmented judge over a curated knowledge base, and an adversarial multi-agent tribunal for hard cases), and combines the signals. We evaluate four detectors on the same balanced 300-claim FEVER slice. Standalone accuracies are modest (0.40–0.55). Even so, at least one detector is correct on 91% of claims, the best two-detector pair combines the two weakest members (0.82), and each detector gets some claims right that no other detector does. A trained stacking arbiter turns this diversity into a 5.7-point gain over the best single detector (0.61 vs. 0.55), where naive majority voting does not. These preliminary numbers support, but do not yet confirm, the idea that failure-mode diversity is a useful design axis for disinformation annotation.

Keywords

disinformation detection, fact verification, ensemble methods, retrieval-augmented generation, multi-agent debate, annotation pipeline

CEUR-WS.org/Vol-4323/paper28.pdf

1. Introduction

Automated disinformation work is usually framed as one classification problem: given a text, decide whether it is true or false. This framing hides a difficulty. The label "false" covers different surface patterns. Clickbait shouts in capitals and exclamation marks. A state-sponsored fabrication reads as grammatical and sober, and is wrong about one date. An AI-generated article has no stylistic tell. A recycled conspiracy theory paraphrases a lie that fact-checkers debunked years ago. Each form carries a *different* signal, and a detector built for one form misses the others.

Our working hypothesis is that robust annotation benefits from *orthogonal* detectors: methods chosen because their errors need not coincide, not only because each is accurate alone. If the hypothesis holds, a small experiment should already show three things. Standalone accuracies stay modest. The union of correct answers exceeds the best single detector. And each detector handles some claims that no other detector does.

A second choice shapes the design: the system *annotates* a claim rather than detecting it. For each claim the pipeline emits a record, not a bit. The record holds a three-class verdict (supported, refuted, or not enough information, NEI), a confidence, cited evidence with a source-authority tier, a natural-language rationale, and, for escalated claims, a dossier that names which sub-claims survived scrutiny. Accountability motivates this. A verdict that a journalist or moderator can audit and overturn keeps a human in the loop instead of automating censorship. "Not enough information" becomes a valid

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia
*Corresponding author.
martin.pavlicek@osu.cz (M. Pavlíček); tomas.filip@osu.cz (T. Filip); petr.sosik@osu.cz (P. Sosík)
0000-0003-1429-2668 (M. Pavlíček); 0009-0001-4386-0620 (T. Filip); 0000-0001-7624-3816 (P. Sosík)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

annotation rather than a detector failure, a distinction a binary detector cannot express. A single detector detects a signal; the pipeline annotates the claim.

Contributions.

1. We describe an end-to-end annotation architecture (§3.1) organised by dual-process cognition [1]: cheap System-1 detectors on every claim, System-2 evidence-based reasoning where warranted, a closed-world signature cache in front, and an adversarial tribunal behind for conditional escalation.
2. We implement five heterogeneous detectors and their infrastructure (§3.2): a curated knowledge base with lexical, dense, and graph stores, kept separate from a human-governed store of known false claims.
3. We report a shared-slice evaluation (§4): four detectors on the same 300 FEVER claims, with standalone accuracy, cross-detector complementarity (oracle union, unique catches, disagreement), and qualitative findings from running the stages end to end.

We claim no ranking among detectors and no deployable accuracy. The preliminary results serve only to inspect complementarity and to surface failure modes before a larger study.

2. Related work and state of the art

State of the art in automated fact-checking. Surveys consolidate automated fact-checking into a pipeline of claim detection, evidence retrieval, verdict prediction, and justification production [2], and map how large language models (LLMs) have since reshaped every stage [3]. The dominant current paradigm is LLM-based decompose-then-verify: FacTool and FActScore split text into atomic claims and verify each against retrieved evidence [4, 5], ProgramFC compiles complex claims into executable reasoning programs [6], and Self-RAG trains the model itself to decide when to retrieve and to critique its own output [7]. The frontier remains active: DEFAME, a modular pipeline that plans tool use and drafts structured fact-check reports, recently set the state of the art on uni- and multimodal benchmarks [8], while the 2025 AVeriTeC shared task shifted the emphasis to open-weights, reproducible systems that verify real-world claims in under a minute on a single GPU [9]. These systems optimise a single strongest verification path. Our contribution is orthogonal: we compose heterogeneous detectors whose failure modes need not coincide, and expose the composition as an auditable annotation rather than a bare label.

Fact extraction and verification pipelines. Automated disinformation analysis is often framed as a pipelined classification problem. The FEVER benchmark [10] established atomic claim verification and split the task into evidence retrieval and veracity prediction. Real-world claims rarely match FEVER's single-hop structure. LIAR [11] and ISOT [12] added short political statements and full-length fake-news articles, which stress different linguistic structures and style-based deception; ISOT remains a common substrate for ensemble-style detectors [13]. FEVEROUS [14] extended verification to tables and unstructured text, and AVeriTeC [15] introduced real-world claims whose evidence comes from the open web, which raises the bar for retrieval in agentic RAG systems.

Complex claims and many-hop verification. Fabrications increasingly rely on partial truths or need multi-step reasoning to debunk. HoVer [16] introduced many-hop fact extraction that traces evidence across several Wikipedia articles. MuSiQue [17] showed that standard multi-hop benchmarks admit disconnected reasoning shortcuts, and proposed a bottom-up question-composition procedure. Our planned dependency-graph decomposition targets these compositional claim structures.

Knowledge-grounded verification and multi-agent debate. To reduce hallucination and over-confidence in parametric language models, recent systems ground reasoning in external knowledge with retrieval-augmented generation (RAG). Even with strong retrieval, a single-agent judge struggles with conflicting evidence. Multi-agent debate suggests that adversarial interaction, in which agents

debate the evidence rather than answer in one pass, can improve factuality and logical consistency and reduce single-model blind spots [18, 19]. Our own findings (§4.5) show that debate also introduces its own failure modes. Our tribunal fixes explicit prosecution and defense roles and separates evidence verification from debate synthesis, which limits the opinion trivialisation seen in unconstrained debate.

Ensemble orthogonality and check-worthiness. Our central hypothesis is that diversity of failure modes can matter as much as the strength of individual detectors. This is the classical condition for an ensemble to beat its members: errors must be frequent enough to matter and uncorrelated enough to cancel [20]. Stylometric and transformer-based encoders perform well on isolated datasets, yet they can share vulnerabilities under distribution shift. Combining detectors with different views of a claim is a practical way to cover one view's blind spot with another view's signal. Because scoring every web claim is expensive, check-worthiness estimation [21] acts as a triage gate for claims that are factual, novel, or consequential enough to justify downstream processing.

3. Methodology

The methodology has three layers: the pipeline architecture that routes a claim from raw text to an auditable annotation (§3.1), the five detectors with their knowledge infrastructure (§3.2), and the combination layer that turns their heterogeneous signals into one verdict (§3.3). The implementation is Python 3.12; every generative stage calls one shared LLM client against locally served open-weights models (Ollama, Gemma 4 in this run) at temperature 0 unless a stage overrides it, with a prompt-side JSON contract and a tolerant parse-and-repair layer that returns an explicit failure rather than raising.

3.1. Pipeline architecture

The runtime flow (Figure 1) places cheap filters in front of expensive reasoning.

Decomposition. A prompt-programmed LLM splits text into atomic, self-contained claims in two passes. The generation pass uses a fixed system prompt and returns a strict JSON schema; the prompt contains an atomic-claim definition, four heuristics (atomicity, coreference resolution, factual integrity, exclusion of rhetoric), and one worked exemplar, and is kept verbatim across calls so the serving layer can cache it. The validate-and-refine pass re-presents each claim against its source passage, once and without retry, and returns a grounded/not-grounded judgement plus an optionally rewritten self-contained sentence; an ungrounded claim is dropped, and a parse failure keeps the claim unchanged. For each claim the stage emits a standalone sentence and a best-effort subject–predicate–object (SPO) triple. It favours precision over recall and drops borderline material rather than pass an unsupported claim; Section 4.5 reports the cost of this choice.

Signature cache (before the gate). Each claim is first matched against a closed-world registry of *confirmed* false claims. This follows an antivirus model: identify a lie once, then catch it cheaply afterwards. An exact or normalised match (whitespace collapse, lower-casing, stripping of surrounding quotes and terminal punctuation) against the SQLite registry settles the claim at confidence 1.0 with no model calls. Only human-confirmed entries match; a human governs the registry through the states candidate, confirmed, and retired. The cache runs before the check-worthiness gate. A semantic variant of the store also acts later as a detector signal: claims and signatures are embedded with a sentence encoder (all-MiniLM-L6-v2) into two vector collections, known lies and known truths, and the signal fires only when cosine similarity to a known lie reaches 0.85 or the contrastive margin (lie similarity minus truth similarity) leaves a ± 0.05 dead zone; otherwise it abstains. The registry stays separate from the knowledge base, so a lie is never retrieved as evidence for itself.

Check-worthiness gate. A rubric-prompted LLM scores each claim's checkability in $[0, 1]$ with a short justification. The rubric asks whether the claim is a factual assertion rather than opinion or prediction, is verifiable against reliable evidence, could mislead public discourse if false, and concerns public interest. A single threshold ($\tau = 0.5$, deliberately uncalibrated at this stage) routes the claim to *accept* or *discard*; a malformed score follows a configurable policy whose default fails open: the claim proceeds, flagged for human review. A wrong drop silently removes a claim from the pipeline, whereas a wrong admit costs only one detector pass.

Detector battery and combination. The detectors of §3.2 score each accepted claim; §3.3 details how their heterogeneous signals combine into one verdict. A detector abstains by returning nothing when its store is absent or its stage fails. Abstention is the absence of a signal, not evidence for NEI.

Escalation and document verdict. A claim escalates to the tribunal on either of two triggers: combined confidence below 0.6 (a conservative, deliberately uncalibrated default), or disagreement, defined as more than one distinct non-NEI verdict among the detectors. An escalated claim carries the RAG stage's retrieved evidence into the hearing, and the tribunal's agreement score replaces the combined confidence. The document verdict is refutation-asymmetric: any refuted claim refutes the document, otherwise any supported claim supports it, otherwise the verdict is NEI. One well-supported lie among many true claims is how disinformation is often packaged.

3.2. The detectors and their infrastructure

Each detector is a standalone module. Its only obligation is to emit a verdict, a confidence, and a feature vector keyed by claim id. Adding a detector means adding one module; removing one means dropping its features. The five span different views of a claim.

GBM: stylometry. A LightGBM classifier (300 trees, depth ≤ 6, learning rate 0.05, leaf count 31, feature and row subsampling 0.8, seed 42) over 38 hand-crafted stylometric features in eight families: lexical richness (8), syntactic complexity (7), readability (4), sentiment (2), cue words and discourse markers (5), part-of-speech distribution (6), information-theoretic (3), and numeric and entity signals (3). It is trained on FEVER claim text mapped to the three-class space with ordinal class weights, under five-fold stratified cross-validation. It uses *how* a claim is written, not what it says. By design it is the cheapest and most limited member, useful only when its correct answers are unique. Feature importances make it interpretable.

BERT: semantics. A fine-tuned DistilRoBERTa regressor (`distilroberta-base`): attention-masked mean pooling of the final layer, through dropout (0.1) and a linear head with sigmoid, onto a scalar in $[0, 1]$ (supported $= 0$, NEI $= 0.5$, refuted $= 1$). Fine-tuning runs four epochs at batch size 32, learning rate 2×10^{-5}, weight decay 0.01, 10% linear warm-up, and maximum length 128 tokens ($\sim$90 minutes on one RTX 4090), co-training FEVER claims with ISOT full-length articles mapped to the same scalar target; the continuous score is binned to a verdict at report time. It encodes what a claim *means*, but it cannot verify a date. It measures plausibility, not truth.

LLM: parametric knowledge. A frozen instruction LLM (Gemma 4, served locally at temperature 0.1 with a 128-token output budget), prompted few-shot with three fixed exemplars that pin the JSON output contract, judges claims from its own world knowledge. It is the only detector that explains its verdict in words. It is also overconfident, a property we measure in §4.5.

RAG: evidence. An agentic retrieval-augmented judge over a curated knowledge base. Hybrid retrieval fuses lexical BM25 and dense cosine rankings (768-d `nomic-embed-text` embeddings) by reciprocal rank (k=60) over the two *text* rankings only, returning the top 8 chunks; a sufficiency-gated loop then

issues one refined follow-up query per hop, up to three hops. A knowledge-graph check runs once per claim outside the loop and sorts matched SPO triples into confirmations and contradictions. An LLM judge reads the assembled evidence, and four signals fuse into the verdict under fixed weights (LLM judge 0.35, graph 0.30, evidence similarity 0.20, entity coverage 0.15); the weight of an absent signal is redistributed proportionally rather than counted as neutral evidence, and the fused score maps to supported below 0.38, refuted above 0.62, and NEI between. It is the only detector that cites sources and the only one that improves as the knowledge base grows.

TRIBUNAL: adversarial deliberation (escalation only). A seven-role, four-phase hearing (triage, briefing, debate, adjudication). A router estimates claim coherence and sets the debate depth: 0 rounds above coherence 0.85, 2 rounds between 0.5 and 0.85, 4 rounds below; a malformed routing reply defaults to 2 rounds, never to the fast path. An advocate decomposes the claim and defends it. An investigator files a counter-analysis from model knowledge. A prosecutor attacks along four named axes (source reliability, logical consistency, completeness, bias) over the routed rounds while a ledger tracks each sub-claim through the states active, contested, conceded, defended, grounded, and dropped. A grounding verifier credits only evidence-traceable survivors. A synthesis writer, which never reads the debate because the code enforces that boundary, issues the verdict with an auditable dossier and a full conversation transcript. The hearing's agreement score (0.7 verdict agreement + 0.3 claim survival) replaces the combined confidence; below 0.5 the claim is flagged for human review.

Infrastructure. The truth knowledge base ingests authoritative text once, offline, into three stores that share one provenance spine: a lexical index (SQLite FTS5/BM25), a dense vector store over sentence-windowed chunks (two sentences with one-sentence overlap, character spans kept for citation), and an SPO knowledge graph with a pluggable SQLite or FalkorDB backend, populated by LLM triple extraction. Every downstream verdict can therefore cite "source X, sentence Y." The disinformation-signature store is a strictly separate closed-world mirror. This separation is a security decision, not an implementation detail, because the evidence store must not be contaminable by the material it is used to refute.

3.3. Combining heterogeneous signals

The detectors are deliberately heterogeneous (a feature-based classifier, a fine-tuned encoder, two LLM judges, and a similarity signal), so their interaction is constrained to one narrow interface. Every detector, whatever its internals, must emit a verdict in {supported, NEI, refuted}, a confidence in $[0, 1]$, and a feature vector, keyed by claim id; signals join by strict inner join on that id, and detectors never read each other's output. Heterogeneity is thus contained at the contract boundary, which is also what makes members replaceable (§6.2).

The runtime vote. The deployed combiner is a confidence-weighted vote: each non-abstaining detector casts its confidence as the weight of its verdict, the verdict with the largest total weight wins, and the winner's share of the total becomes the combined confidence; if every detector abstains the claim is NEI at confidence zero. There are no fixed per-detector weights: a detector's influence on a claim is exactly the confidence it reports on that claim. We run this untrained rule at runtime deliberately: it is transparent and auditable, it cannot leak label information, and it is the honest baseline a trained combiner must beat. It is also, measurably, the current weak point: on the shared slice the vote scores 0.517, *below* the best single detector (0.550). Averaging heterogeneous confidences hurts when one member is systematically overconfident (§4.5).

The trained arbiter. The upgrade path is a stacking arbiter: a gradient-boosted three-class model over the detectors' exported feature vectors (feature blocks wider than 32 dimensions are compressed by PCA; label-bearing columns are removed by a blocklist), trained with five-fold stratified out-of-fold predictions and attributed per signal block with SHAP. The offline ablation of §4.3, a stacker over the

Table 1
Per-detector results on the shared 300-claim slice (provisional). Accuracy (↑) is exact three-class *label* match on the paired join (§4.1). The last column gives the failure signature the architecture predicts for each detector.

Detector	Acc.	Cost	Observed signature
GBM	0.403	ms, CPU	predicts SUP for 89%; style $\neq$ truth
RAG	0.463	$\sim$10 s	NEI-biased; graph store
LLM	0.490	$\sim$5 s	43% of errors stated at $\geq$90% conf.
BERT	0.550	ms, GPU	best single here; trained also on long-form news
oracle (any of 4)	**0.910**	—	upper bound, not a deployed combiner

detectors' verdicts and scores, reaches 0.607 on the same slice where the vote scores 0.517. The runtime pipeline still runs the untrained vote; wiring the trained arbiter in is future work (§5.2, §6.1).

Disagreement as a signal. The vote consumes detector *agreement*; the escalation trigger consumes detector *spread*. A claim on which detectors return more than one distinct non-NEI verdict escalates to the tribunal regardless of the vote's confidence, so disagreement between heterogeneous members is treated as the cue that a claim deserves deliberation rather than as noise to average away.

Reproducibility. Every detector scores the same frozen evaluation spine: a SHA-256-pinned sample of FEVER dev claims, balanced across the three classes, drawn at a fixed seed. Every experiment archives per-claim records for all detectors, the configuration, and a run manifest to a versioned run folder, and every number reported here regenerates from those folders by script. The code, evaluation samples, and archived results are available from the authors on request.

4. Experiments

4.1. Setup

We drew a balanced slice of 300 claims (100 per class, fixed seed) from a frozen FEVER-derived evaluation set. Four detectors scored the *same* 300 claims: GBM (trained on FEVER train, ternary mode), BERT (co-trained on FEVER and a long-form news corpus, scores mapped to the nearest regression target), LLM (few-shot, local `gemma4:e4b`), and RAG (a frozen knowledge-base snapshot of $\sim$1,300 Wikipedia pages and 12,000 chunks, with a local judge). The tribunal is escalation-only and an order of magnitude more expensive, so we evaluated it separately (§4.5). We exercised decomposition and the worthiness gate end to end on a sub-slice. We normalised all verdicts to the shared three-class space before scoring and joined on claim id.

Metric. We report label accuracy: exact three-class match on the paired join. FEVER's protocol distinguishes this from the stricter FEVER score, which also requires the submitted evidence to cover the gold set [10]. We use the weaker metric deliberately: three of the four detectors (stylometric, semantic, parametric) produce no evidence by construction, so an evidence-conditioned score is undefined for them, and label accuracy is the only metric all members share, which the complementarity analysis requires. The annotation record of the evidence-producing path retains its citations, so evidence-aware scoring of the RAG and tribunal arms over the archived runs is possible, and the roadmap schedules it (§6.1). All oracle unions are label-only and therefore optimistic in FEVER-score terms.

4.2. Standalone detectors: modest in this run

Table 1 shows modest standalone performance. The best single detector reaches 0.550. GBM, blind to content, often predicts the majority surface form. BERT is the best of the four here, but its stronger

Table 2
Cross-detector complementarity on the shared slice (provisional). Brackets are 10k-sample bootstrap 95% CIs. A "unique catch" is a claim that only one detector got right; the trained arbiter and majority vote are computed on the same 300 claims.

Claims where $\geq$2 detectors disagree	288/300 (96%)
Claims all four detectors miss	27/300 (9%)
Oracle: at least one correct	0.910 [0.877, 0.940]
Best single detector (BERT)	0.550 [0.493, 0.607]
Majority vote (4 detectors)	0.517 [0.460, 0.573]
Trained meta-GBM (5-fold OOF)	**0.607** [0.550, 0.660]
Best two-detector union	0.817 (GBM+RAG)
Unique catches (GBM/BERT/LLM/RAG)	16 / 12 / 10 / 30

performance on long-form news in the same training run did not transfer to short atomic claims. LLM is overconfident. RAG reached overall accuracy 0.463 (Table 1). but its per-class SUPPORTED recall was 0.00 for a concrete implementation reason: its knowledge-graph signal was empty. A follow-up run that populated the graph confirms the diagnosis: per-class SUPPORTED recall recovered from 0.00 to 0.94, but overall accuracy held roughly flat (0.463 to 0.477), because the graph now *confirms* without *contradicting*: it stays neutral on 69% of claims and rarely fires a refutation. Populating the store fixed the dead signal and localised the next bottleneck as contradiction detection and entity resolution, not graph coverage.

4.3. Complementarity: an upper-bound diagnostic

The oracle is an upper bound, not a deployed combiner; the question is how much of its headroom a real arbiter recovers. Four observations follow. First, the gap between the oracle (0.910) and the best single detector (0.550) is large: a perfect arbiter would gain 36 accuracy points. This bounds the prize; it does not show any combiner claims it. Second, a *trained* stacking arbiter (meta-GBM over the four detectors' predictions and confidences, 5-fold out-of-fold) reaches 0.607, 5.7 points above the best single member and about 16% of the oracle headroom. Its 95% CI ([0.550, 0.660]) has a lower bound at best-single's point estimate, so at N=300 the gain is positive but not decisively separated: we read it as evidence the headroom is *partly* learnable, corroborated below by an independent pilot, not as a settled ensemble accuracy. Third, the gain needs a trained combiner: equal-weight *majority vote* scores 0.517, *below* the best single detector, because three weaker members outvote the strongest. Complementarity is real but not free. Fourth, the best two-detector union pairs the two weakest members, GBM (0.403) and RAG (0.463), for 0.817, and every detector has unique catches, including 30 for RAG, so selecting detectors by standalone accuracy alone would discard useful ones. Disagreement at 96% is common enough to serve as a routing signal, though whether escalation improves final accuracy is a separate question the full pipeline must measure.

4.4. Corroboration from an earlier pilot at larger scale

The same complementarity signature appeared, independently, in a pilot run of this system's *previous implementation generation* (four analogous signals, N=999 paired FEVER claims). The pilot is our own earlier, previously unpublished work: it was conducted by the authors as part of the development leading to this paper. We include it as historical corroboration, not as a result of the present system.

Three points carry over (Table 3). The gap between the oracle and the best single member was again large (80.1% vs. 56.3%). Every signal, contributed unique catches. And the most complementary pair again involved the weakest member. Crucially, the pilot independently reproduces the arbiter result: its *trained* stacking arbiter beat the best single member by 2.2 accuracy points out of the 24-point oracle headroom, the same direction as the current slice's 5.7-point gain but at three times the sample size and on a different implementation. Part of the union is learnably recoverable in both generations, which is

Table 3
The earlier pilot (N=999 paired claims, previous implementation generation). It shows the same qualitative signature as Table 2, including a *trained* stacking arbiter that beats best-single at three times the scale.

Per-signal accuracy (GBM / BERT / LLM / RAG)	0.400 / 0.563 / 0.461 / 0.080
Oracle: at least one correct	800/999 (80.1%)
Unique catches (GBM / BERT / LLM / RAG)	50 / 108 / 125 / 24
Most complementary pair (Jaccard of correct sets)	LLM×RAG, 0.059
Trained meta-GBM (5-fold OOF) vs. best single	**0.585** vs. 0.563

more informative than either measurement alone.

4.5. What running end-to-end taught us

Four findings surfaced only because the stages ran together, on real data, with full logging, and each maps to a named design property. Appendix A reproduces verbatim log specimens for every stage, including the exchanges quoted below.

Decomposition drops some claims silently. On the first end-to-end sub-slice, a false claim ("Phoenix, Arizona is in southwestern Mexico") produced *zero* claims out of decomposition. The validator, asked to check grounding, appears to treat "grounded in the source text" as "true in the world," and dropped a refutable lie before any detector saw it. The cost of the precision-first posture is real, not hypothetical.

The gate discards some labelled claims. The check-worthiness gate discarded gold-labelled SUPPORTED claims ("Bank of America has machines") as not check-worthy. These are defensible triage decisions, but they cost benchmark recall, which motivates the gate-versus-no-gate ablation in the re-runs.

LLM overconfidence is visible in the logs. The mean stated confidence was 82%, and 67 of 156 errors were stated at $\geq$90% confidence. Never letting the LLM arbitrate alone is therefore a conservative design choice, not only an architectural preference.

The tribunal produces useful uncertainty, not accuracy. A companion study evaluates the escalation tribunal in isolation on a registered N=600 hard-case slice with pinned gold evidence, against a single-pass judge under fixed model and decoding. Two findings bear on the pipeline. The tribunal's disagreement signal ranks its errors better than the judge's confidence ranks its own (ΔAUROC $+0.074$; the agreement score correlates with hearing correctness at $r = 0.63$). Its agreement-based review flag is therefore informative, which is what an escalation layer needs. The tribunal also shows an *over-concession* failure: it argues correct refutations away under prosecutorial pressure and collapses REFUTES to NEI on 53 of the 600 hearings. A concession-resistant defense protocol roughly halves this, and lifts refutation accuracy on the affected subset from 0.30 to 0.55. Evidence-less hearings, as occur in this pipeline before the RAG stage supplies evidence, drift to NEI and trip the collapsed-hearing flag. We report these tribunal results as a companion study and do not fold them into the shared-slice numbers above.

5. Discussion

5.1. Failure modes and the annotation principle

Every detector can be fooled, each in a characteristic way (Appendix A): stylometry by a sober register, semantics by plausible-but-empty claims, parametric knowledge by evidence-relative labels, retrieval by

its own fusion arithmetic, and deliberation by social pressure. We do not claim any member is robust; we build on an explicit account of each one's blind spot, which becomes a documented property that the arbiter can discount, the escalation trigger can route around, and the annotation record exposes.

These blind spots are not hypothetical: the shakedown exhibited each one as a measured signature. The stylometric detector, facing FEVER's uniformly sober register, collapsed to predicting SUPPORTED for 89% of claims. The parametric detector stated 43% of its errors at 90% confidence or higher. The retrieval detector's graph signal skewed its fusion toward NEI, and the tribunal argued correct refutations away under prosecutorial pressure. That the architecture predicted the *direction* of each failure before the run is the strongest evidence we have for its organising principle; that every failure was visible afterwards in the archived signals is what the annotation layer is for. Because every verdict carries its signals, confidence, and rationale, a reviewer sees what the system concluded *and* which view produced it, and a wrong verdict stays auditable. Robustness here is a property of the composition and its transparency, not of any single component.

The output this transparency most needs to catch is a confident SUPPORTED on a lie, the false negative that lets a fabrication through. We treat false-negative analysis as its own workstream. Every run archives per-claim records for all detectors, so the "all four missed it" subset (27/300) is queryable, and the signal viewer and tribunal transcripts make each miss explainable: which signal was silent, which sub-claim conceded, which evidence never retrieved. The silent decomposition drop of §4.5 came from this branch. Planned runs add designed-weakness slices (fluent rewrites of false claims) to probe the false-negative surface deliberately.

5.2. Limitations

These results are a shakedown, not an evaluation. At $N=300$ we make no calibration claim, and every interval is wide. One backbone model per stage confounds detector identity with model choice. FEVER's short, single-hop, Wikipedia-grounded claims limit external validity. The trained-arbiter number is an *offline* 5-fold stacker over exported predictions; the combiner wired into the running pipeline is still an untrained weighted vote, and a production stacking model on the registered exports remains future work. The tribunal's contribution to the full pipeline, as opposed to its solo behaviour, is not yet measured. No detector ordering should be inferred from Table 1, and we assert none.

6. Future work

6.1. A stronger evaluation

The path to a stronger evaluation is planned and partly mechanised. (i) Pre-registered re-runs on a hard-case-biased slice ($N=600$, calibration set ≥ 500) with paired statistics (McNemar's test on paired verdicts, bootstrap confidence intervals over claims) are running. (ii) A fully populated knowledge graph for the evaluation corpus (LLM triple extraction over all 12,253 chunks) and the repaired fusion rule together re-measure the RAG detector. (iii) The offline combiner ablation reported here (stack 0.607 versus vote 0.517 versus best-single 0.550) moves into the running pipeline and to registered scale, replacing the untrained weighted vote with the trained stacker. (iv) A noisy-retrieval regime replaces gold evidence, where escalation may show measurable headroom. (v) Gate-versus-no-gate and decomposition ablations address the qualitative findings. (vi) Evidence-aware scoring in the style of the FEVER score (label correct *and* cited evidence covering the gold set) for the evidence-producing RAG and tribunal arms, computed over the archived per-claim records. Every number in this paper regenerates from archived run folders by script.

The same programme widens the data along two axes. The shakedown uses one benchmark family: FEVER for short atomic claims and ISOT for the trainable detectors. For external validity, LIAR (short political claims), FEVEROUS (table and multi-sentence evidence), and AVeriTeC (web evidence with question–answer justifications) each break a different FEVER simplification: register, evidence modality, and the benchmark-to-real-world gap. For complex claims (§6.2), HoVer and MuSiQue-style multi-hop

suites test the decompose–resolve–aggregate path. We freeze sampling and prompts before each run and report per-dataset results separately, since we expect components to rank differently across these regimes.

6.2. Extending the architecture

Four extensions follow, in increasing order of ambition: stronger members, a more robust signature store, wider claim structures, and closing the loop from detection back to the store.

Stronger members. Each component has an upgrade path that leaves the architecture intact. Better prompts: the concession-resistance A/B shows a prompt-level fix is measurable (0.30→0.55 on the affected subset). Better models: every LLM stage runs through one shared client with per-stage, and per-role in the tribunal, model selection, so a new backbone is a configuration change. Better methods: each detector sits behind a claim-id contract, so a supervised gate, an entailment-based contradiction checker, or a graph-native retriever can replace its counterpart and face the same test, whether it adds unique catches rather than only accuracy.

Robust signature lookup. The exact-match cache is the most conservative component, firing only on verbatim or normalised recycling of confirmed lies: high precision, low paraphrase recall by design. Coverage grows in graded steps, each already staged. The semantic contrastive signal (similarity to known lies versus known truths) participates as a detector today. Write-through promotion of high-confidence semantic hits into the fast tier is gated on measured paraphrase catch-rates. And SPO-triple keys are feasible because decomposition already emits a subject–predicate–object triple for every claim: the key is the normalised triple itself, so the same structural lie in entirely new words (new framing, new adjectives, same subject, predicate, and object) matches without any embedding call, and a match merely promotes a candidate for human confirmation. One constraint holds throughout: only human-confirmed signatures populate the cache, because an attacker-writable cache would be worse than none.

Complex claims. Real disinformation is rarely atomic. Partial truth is the norm, and the combinator matters as much as the facts: "A and B" fails if either fails, while "A because B" can be false even when A and B are both true. The planned extension generalises decomposition to a dependency graph with an explicit logical form, verifies each leaf through the existing battery, and aggregates under three-valued logic, so one REFUTES sinks a conjunction and NEI is contagious. A compound verdict can then report *which* sub-claim failed, which a reviewer needs for a half-true statement. This layer is design, not result, and we mark it as such.

The ingest loop. The long-term goal is a partly automated loop that turns detections into cache hits: a claim refuted with high confidence, or by a reviewer, is decomposed, matched, investigated, and, only after human confirmation, ingested into the signature store, so a repeated lie is cheaper later. The confirmation step is the governance boundary that keeps the store trustworthy; the loop automates the work around it (candidate generation, deduplication, evidence assembly) and leaves confirmation to a person.

7. Conclusion

What the pipeline is. We presented the Cognitive Disinformation Annotation (CDA) pipeline, which turns free text into an auditable, evidence-grounded annotation of each atomic claim. It decomposes a claim, checks it against a closed-world cache of confirmed lies, filters it by check-worthiness, and scores the survivor in parallel with five detectors (stylometric, semantic, parametric-LLM, retrieval-augmented, and signature-based). It combines the signals into a verdict (supported, refuted, or not

enough information) with a confidence, cited evidence, and a rationale, escalates contested claims to an adversarial tribunal, and forms the document verdict by refutation asymmetry. The organising principle is the diversity of detector failure modes; the organising output is an annotation, not a binary label.

Why we build it this way. Disinformation has no single surface, so a detector built for one guise is blind to the others, and we compose detectors whose errors need not coincide. Because these systems inform contestable decisions about what people may say and see, we treat the task as annotation rather than automated judgement: every verdict carries the evidence, confidence, and reasoning needed to audit and overturn it, and "not enough information" is a first-class outcome. The engineering follows the same logic. Disinformation detection is not one problem but a set of coupled subproblems: segmentation into claims, triage, verification from several views, combination, deliberation, and memory of confirmed lies. We build each as a replaceable module behind a common contract, and we optimise for the system's ability to improve, as members are swapped and the knowledge base grows, rather than for one number on one benchmark. Robustness comes from the composition and its transparency, not from any single model.

What we have shown, and what remains. Our evidence is a preliminary check, not a finished evaluation. On a shared 300-claim slice, four weak-to-modest detectors (0.40–0.55) reached an oracle union of 91%, the most complementary pair combined the two weakest members, and every detector contributed unique correct cases; an earlier generation showed the same signature at $N{=}999$ with a trained arbiter that recovered part of the gap. These are oracle unions, not deployed accuracy. Running the stages end to end also exposed concrete costs: silent decomposition drops, gate over-triage, LLM overconfidence, a retrieval-fusion bug, and tribunal over-concession.

Funding

This article was produced with the financial support of the Operational Programme Just Transition, and of the project Biography of Fake News with a Touch of AI: Dangerous Phenomenon through the Prism of Modern Human Sciences (project no. CZ.02.01.01/00/23_025/0008724) under the Operational Programme Jan Ámos Komenský, and of the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project (project no. CZ.10.03.01/00/22_003/0000048) via the Operational Programme Just Transition.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT GPT-5.6, Antropic Claude Opus 5, Overleaf AI and DeepL in order to do following: Text Translation, Paraphrase and reword, Improve writing style, Grammar and spelling check, Formatting assistance. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

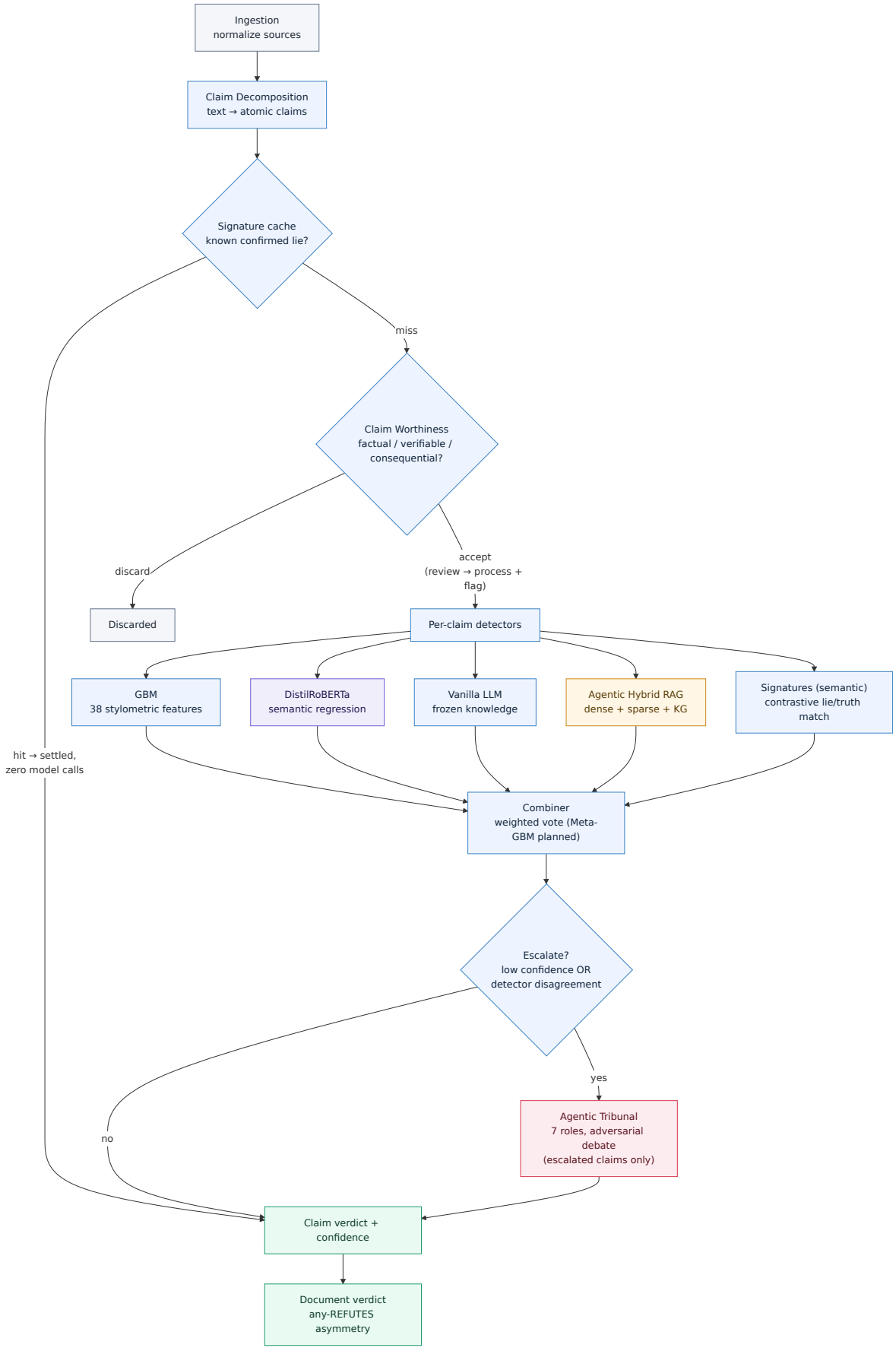


Figure 1: The CDA pipeline. Free text is split into atomic claims. Each claim passes the closed-world signature cache, the check-worthiness gate, the detector battery, and verdict combination. Low confidence or detector disagreement escalates the claim to the adversarial tribunal. The document verdict is refutation-asymmetric: any refuted claim refutes the document.

References

[1] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux, 2011.

[2] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022). doi:10.1162/tacl_a_00454.

[3] A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, I. Augenstein, Claim verification in the age of large language models: A survey, arXiv preprint arXiv:2408.14317 (2024).

[4] I.-C. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, P. Liu, FacTool: Factuality detection in generative AI – a tool augmented framework for multi-task and multi-domain scenarios, arXiv preprint arXiv:2307.13528 (2023).

[5] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained atomic evaluation of factual precision in long form text generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[6] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, P. Nakov, Fact-checking complex claims with program-guided reasoning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023.

[7] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, Self-RAG: Learning to retrieve, generate, and critique through self-reflection, in: The Twelfth International Conference on Learning Representations (ICLR), 2024.

[8] T. Braun, M. Rothermel, M. Rohrbach, A. Rohrbach, DEFAME: Dynamic evidence-based fact-checking with multimodal experts, in: Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025.

[9] M. Akhtar, R. Aly, Y. Chen, Z. Deng, M. Schlichtkrull, C. Whitehouse, A. Vlachos, The 2nd automated verification of textual claims (AVeriTeC) shared task: Open-weights, reproducible and efficient systems, in: Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), 2025.

[10] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, 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), 2018. doi:10.18653/v1/n18-1074.

[11] W. Y. Wang, "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), 2017. doi:10.18653/v1/p17-2067.

[12] H. Ahmed, I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC), 2017.

[13] A. M. Ali, F. A. Ghaleb, B. A. S. Al-Rimy, F. J. Alsolami, A. I. Khan, Deep ensemble fake news detection model using sequential deep learning technique, Sensors 22 (2022). doi:10.3390/s22186970.

[14] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), 2021. doi:10.18653/v1/2021.fever-1.1.

[15] M. Schlichtkrull, Z. Guo, A. Vlachos, AVeriTeC: A dataset for real-world claim verification with evidence from the web, Advances in Neural Information Processing Systems (Datasets and Benchmarks Track) (2023). doi:10.48550/arxiv.2305.13117.

[16] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, M. Bansal, HoVer: A dataset for many-hop fact extraction and claim verification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020. doi:10.18653/v1/2020.findings-emnlp.309.

[17] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal, MuSiQue: Multihop questions via single-hop question composition, Transactions of the Association for Computational Linguistics 10 (2022). doi:10.1162/tacl_a_00475.

[18] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).

[19] C. Zheng, D. Liang, W. Zhang, X.-Y. Wei, T.-S. Chua, Q. Li, A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning, arXiv preprint arXiv:2403.14972 (2024). doi:10.48550/arxiv.2403.14972.

[20] T. G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems (MCS 2000), LNCS 1857, 2000.

[21] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

A. Worked examples from the shakedown logs

Each example below is reproduced *verbatim* (lightly truncated) from the archived run logs of the preliminary evaluation, one specimen set per stage. Claims are FEVER development claims, gold labels are in SMALL CAPS, and system outputs are in quotation marks. The archive paths and the scripts that regenerate them accompany the artifact.

A.1. Decomposition and the check-worthiness gate

The silent drop (the cost of precision over recall). Input claim, gold REFUTES:

> "Phoenix, Arizona is in southwestern Mexico."
>
> **Decomposition output:** *no claims emitted.* The validate-and-refine pass rejected its own generated claim, so the lie never reached the gate or any detector. The validator, asked to check that a claim is *grounded in the source text*, appears to also apply a *world-truth* judgement.

Gate over-triage. Two gold-SUPPORTS claims that the worthiness gate discarded, with its own stated reasons:

> "Bank of America has machines." — score $0.10 \rightarrow$ *discard*: "This is a trivial, easily verifiable statement that lacks public interest or potential for significant misinformation."
>
> "The Beach was adapted for film by John Hodge." — score $0.30 \rightarrow$ *discard*: "…verifiable but lacks immediate public interest or potential for widespread public misinformation."

Both are defensible triage decisions, and both cost benchmark recall, which is why the gate-versus-no-gate ablation is pre-registered.

A.2. GBM (stylometry)

> **Correct, confident** (gold REFUTES, $p_{\text{disinfo}} = 0.946$): "The Wonder Years was only a documentary." The absolutising "only" is exactly the cue-word surface the feature set targets.
>
> **Wrong, confident** (gold NEI, predicted REFUTES, $p = 0.899$): "There are only Chinese people in the Republic of Macedonia." The same "only" cue fires on a claim whose truth status the style cannot settle.
>
> **Fooled by sobriety** (gold REFUTES, predicted SUPPORTS, $p = 0.166$): "Aaron Burr killed Alexander Hamilton in Seaside Heights, New Jersey." A false claim in an encyclopedic register. Style is not truth; this is the detector's designed blind spot, not a bug.

A.3. DistilRoBERTa (semantics)

> **Correct** (gold REFUTES, score $0.981 \rightarrow$ bin *disinformation*): "Daenerys Targaryen is a real person." The encoder captures the semantic implausibility directly.
>
> **Wrong** (gold REFUTES, score $0.013 \rightarrow$ bin *credible*): "A Milli is by a person from a country." FEVER's abstraction games produce claims that are semantically vacuous but plausible, and meaning alone cannot refute them.

A.4. Vanilla LLM (parametric knowledge)

> **Correct, confident** (gold REFUTES, confidence 98%): "Uranium was unable to meet the qualifications to be considered a chemical element." — "Uranium is definitively classified as a chemical element on the periodic table."
>
> **Wrong at 95%, or is it?** (gold NEI, predicted REFUTES): "Sacre-Coeur, Paris is a double obelisk." — "The Sacré-Cœur Basilica is a large Roman Catholic church, not a double obelisk."

The second example is an important footnote to the overconfidence statistic. The model is *right about the world* and wrong about the benchmark label, because FEVER's NEI means "the provided wiki pages do not settle it," not "unknowable." Parametric knowledge and evidence-relative labels disagree by construction on such claims, one reason the LLM is a voice and never the arbiter.

A.5. Agentic RAG (evidence)

> **Correct refutation with citation** (gold REFUTES, fused score $0.74 \rightarrow$ REFUTED): "The Washington Wizards have won zero division titles." Judge (conf. 0.9): "Evidence [1] explicitly states that the Wizards have won 'seven division titles' (listing years including 1971, 1972, 1973, 1974, 1975, 1979…)."
>
> **The dead-graph drift** (gold SUPPORTS, judge SUPPORTED at conf. 0.9, fused verdict NEI at score 0.46): "The Stanford prison experiment was funded by an organization that coordinates…the science and technology programs of the U.S. Navy…" Judge: "Evidence [7] directly states that the Stanford prison experiment was funded by the U.S. Office of Naval Research…"; signal vector $\{\text{evidence}\,1.0,\ \text{graph}\,0.5,\ \text{llm}\,0.1,\ \text{coverage}\,0.5\}$.

The second specimen shows the fusion arithmetic overriding a correct, evidence-grounded judge. With the knowledge-graph store unpopulated, the neutral 0.5 graph signal (weight 0.30) floored the fused score above the SUPPORTED band. The fix, that an absent signal abstains and its weight redistributes, follows the pipeline's own stated rule.

A.6. Tribunal (adversarial deliberation)

Gold REFUTES: "Hot Right Now is from the album Kryptonite." The advocate opens *correctly*:

> **Advocate** (verdict REFUTES): sub-claim A1 = "Hot Right Now is a single by DJ Fresh," cited to evidence "[1] 'Hot Right Now' is a single by British drum and bass producer DJ Fresh."
>
> **Prosecutor** [axis: source reliability]: "Source [1] is not identified (e.g., no author, publication, or date), making its reliability impossible to assess; a single unverified reference is insufficient to support the claim."
>
> **Advocate** (concedes A1): "The claim relies solely on source [1], which is not identified…thus the sub-claim is unsupported."
>
> **Outcome:** all three sub-claims conceded, nothing survives to grounding, and synthesis returns NEI. The agreement score of 0.233 flags the hearing for human review.

This is the over-concession failure in one exchange. The prosecutor attacks the *formatting* of what is in fact the pinned gold evidence, and the advocate abandons a correct, evidence-backed refutation under pressure rather than under shown error. Transcripts like this one motivated the concession-resistant defense protocol (§4.5).