

Verify, Then Argue: Claim Verification in Investment Agent Debates

Joyce Ann Clarize Galang 

UVC Partners
Technical University
of Munich
j.galang@tum.de

Ahmed Mady 

Technical University
of Munich
aabdelghany.mady@tum.de

Georg Groh 

Technical University
of Munich
grohg@mytum.de

Abstract

Multi-agent debate systems incorporate retrieval-based verification to ground agent reasoning with evidence. However, this approach assumes the availability of data and the informative intent of the source text. These conditions are frequently unmet in venture capital, particularly in early-stage startup evaluation. In this industry, evidence is predominantly private or unavailable, and pitch decks use persuasive rhetoric that prior research shows can shift agent conclusions. To measure how agents navigate these conditions, we evaluate how claim verification alters multi-agent debate using a stratified sample of deals from a real venture fund. We find that verification changes the agent’s final investment decision but not its reasoning. Furthermore, agents are insensitive to which claim a verification label is attached to. Reassigning labels to the wrong claims lowers the decision to the same extent. This indicates that the primary bottleneck in automated screening is fundamental claim verifiability rather than retrieval quality.

1 Introduction

Multi-agent debate systems show great promise for improving the factuality and reasoning accuracy of Large Language Models (LLMs) (Du et al., 2023; Liang et al., 2024). However, baseline multi-agent systems assume input factuality and good-faith participation, leaving agents to debate over potentially misleading or adversarial information (Xiong et al., 2025; Kraidia et al., 2026). Researchers have worked to mitigate this by incorporating retrieval of external evidence, grounding claims in retrieved evidence rather than the models’ internal knowledge (Han et al., 2026; Ma et al., 2025).

While retrieval-based methods are generally effective at mitigating misleading debate input, they face limitations in venture capital (VC), the funding of early-stage, high-risk startups. In this indus-

try, claim verification breaks down in two ways. First, accessible evidence is scarce. Startup data like financial projections, product information, and track records are either unavailable or private (Kaplan and Lerner, 2016; Retterath and Braun, 2020). Second, pitch decks are built to persuade, not just inform. Without hard data, investors rely heavily on initial impressions (Huang and Pearce, 2015), and decks are engineered to shape exactly those impressions (Chen et al., 2009; Clark, 2008; Williams et al., 2020) through crafted narratives and loaded language (Guzmán-Solano et al., 2025). LLM agents are themselves swayed by such framing and confident tone, which can shift their conclusions toward incorrect answers (Kraidia et al., 2026). How much utility evidence retrieval retains under these conditions has not been measured, and that is what we set out to quantify.

To investigate this, we test how claim verification affects debate agents on a stratified sample of a real venture fund’s deal pitch decks. We work in a debate architecture deliberately. On decision accuracy alone, debate matches single-pass screeners, but it yields ranked decisions and natural-language explanations a screening team can act on (Bae et al., 2026), and it preserves divergent positions rather than prematurely collapsing them (Liang et al., 2024). We address three questions.

(RQ1) *Does verifying claims before the debate change how agents reason and invest?*

(RQ2) *Is a changed decision driven by the evidence’s content or presence?*

(RQ3) *How much of the document can be checked at all, and when a claim is not confirmed, is it undocumented in public or does retrieval miss it?*

Contributions. We make three contributions. First, we show that in this low-evidence setting verification changes what the agent decides but not what it reasons over, because agents respond no dif-

Ground Truth (Example): SpaceX is valued at **\$800B** at the end of 2025.

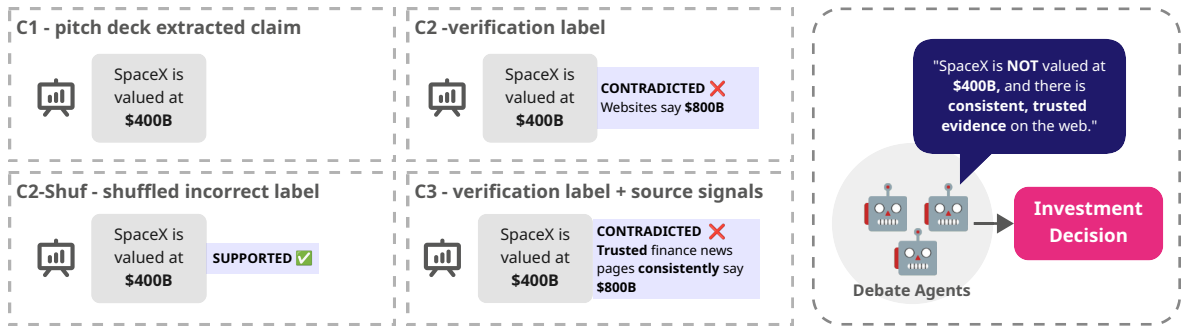


Figure 1: **Verify-then-argue on an illustrative claim.** A claim is passed to a fixed multi-agent debate under different framings of the evidence block: the bare extracted claim (C1), its real web verdict (C2), a *shuffled* verdict (C2-shuf), and the verdict plus source reliability and consistency signals (C3). Inputs are identical except for the evidence block, so any change in the invest-or-pass decision isolates whether the agent responds to the verdict’s content or merely its presence. The raw-text condition (C0) is omitted for space.

ferently when verdicts are reassigned to the wrong claims, and the limiting factor is how little of a deck can be checked rather than a retrieval gap a better search tool would close. Second, we contribute an evaluation design that makes this measurable,¹ combining planted false claims, a five-condition evidence ladder whose shuffled-label control separates content from presence, and a propagation measure that traces a false claim into the reasons a decision rests on. Third, we run the first stress test of verify-then-argue in early-stage VC, where most facts about a private company cannot be checked on the open web.

2 Background and Related Work

Venture capital investment VC firms invest in early-stage, high-risk startups in exchange for equity, expecting most to fail and returns to concentrate in a few winners (Cochrane, 2005), which makes selecting which companies to back a firm’s primary driver of performance (Gompers et al., 2020). This selection is also a problem of scale, as firms screen thousands of companies for every few they fund, with the pitch deck serving as the primary artifact reviewed at the initial stage. To manage this volume, VC firms increasingly adopt LLM-based systems that parse heterogeneous, multi-modal pitch materials and reason over them to triage deals (Galang et al., 2027).

LLM approaches for startup investment Most LLM-based approaches to startup screening are

¹Code, prompts, de-identified result grids, and a synthetic corpus that runs the pipeline end to end: <https://github.com/joyce-acg54/verify-then-argue>

single-pass predictors that extract features or prompt a model once for a decision (Sadia and Cheng, 2026; Maarouf et al., 2025). A smaller line of work coordinates multiple agents, as investor personas, due-diligence committees, or analyst roles, or through graph retrieval and founder self-play (Pei et al., 2025; Liu et al., 2025; Meixin et al., 2025; Wang et al., 2025; Xiong and Ihlamur, 2023). Of these, only DIALECTIC structures the agents as an adversarial debate, with pro and con agents arguing under a critic before a decision (Bae et al., 2026). Multi-agent debate is therefore the exception rather than the norm in startup screening, and how it behaves under unreliable evidence, the focus of this work, remains untested.

Information verification Grounding generation in verified evidence improves factual accuracy and reduces hallucination, through retrieval-augmented generation (Lewis et al., 2020), self-generated verification questions (Dhuliawala et al., 2024), and atomic-fact checking (Wei et al., 2024), building on a long line of automated fact-checking (Guo et al., 2022). Recent systems run claim verification first, extracting each claim, retrieving evidence, and attaching a label before the model reasons, as in automated peer review (Yue et al., 2026) and retrieval-augmented argument graphs (Zhu et al., 2025).

This pattern extends to multi-agent debate, which improves factual accuracy (Du et al., 2023) but multiplies inference cost (Ma et al., 2025), making it especially worthwhile to verify inputs before debating. ED2D retrieves external facts within a five-stage debate to support or challenge arguments

(Han et al., 2026), DelphiAgent mines and filters evidence before a panel reaches consensus (Xiong et al., 2025), and GK MAD fuses retrieved evidence with debate insights for the final decision (Ma et al., 2025).

3 Experimental Design

We test whether verifying a deck’s claims changes how an agent reasons and what it decides in a debate. The agent $\mathcal{A}$ reads a pitch deck x together with an *evidence block* $e(x)$, a short report we attach to the deck that says, for each claim, what checking it against the web turned up. The agent then argues both sides and makes a forced choice between invest and pass (Figure 1). We keep $\mathcal{A}$ and x the same on every run and change only $e(x)$, so any change in the agent’s behavior must come from the evidence. We track two outcomes. *Propagation* is whether a false claim ends up as a *load-bearing premise*, or one of the reasons the final decision actually rests on rather than a point the argument only mentions in passing. The *decision* is the agent’s invest probability $P(\text{invest})$, a number between 0 and 1 for how strongly it leans toward investing. To detect propagation we need to know which claims are false, but real decks do not tell us. We therefore plant known falsehoods, called *canaries*, into real decks and trace whether the agent carries them into its reasoning.

3.1 Verify-then-argue.

Our pipeline, *verify-then-argue*, adds an atomic-claim verification stage before the agent debate, the input-side grounding motivated in Section 1. Both stages are built for this work from standard components, and neither is novel in itself. The verify stage is the instrument that produces the evidence blocks, and the debate is the fixed agent whose behavior we measure. The verify stage turns a deck into checked claims in a few steps. First, we extract *atomic claims*, single verifiable statements. Each extracted claim is tied to a verbatim quote from the slide. A programmatic check voids any quote not found in the page text, and the audit drops claims it judges unfaithful, constraining the extractor from inventing claims the deck never made. Next, we route the claims by verifiability, so that only verifiable ones are searched. Then each verifiable claim is verified with an LLM web-search tool and labeled supported, contradicted, conflicting, or no-evidence, with two auxiliary signals, source re-

liability and cross-search agreement. For example, the claim “Company A has deployed six paid pilots” is routed verifiable and, when search returns nothing, labeled no-evidence, while “Company A’s approach is uniquely defensible” is routed normative and never searched. Table 4 in the appendix shows a larger sample, and Figure 1 walks one claim through the conditions.

Finally, the claims, packaged for the condition under test, are passed to the argue stage, a multi-agent debate we reimplement from DIALECTIC (Bae et al., 2026), with its pro and con argument generation, a devil’s advocate critic, and iterative scoring and refinement. We build on DIALECTIC because it is, to our knowledge, the only startup-screening debate validated against real investor decisions. Step-by-step mechanics are in Appendix B.1, and how much of a deck the extractor and verifier recover is a result we report in Section 5 (RQ3).

3.2 Conditions.

We build a ladder of evidence blocks over a shared set of decks, some carrying the planted canaries and the rest left unmodified for replication. Let $K(x)$ be the claims, v their labels, π a random reshuffle, and $\rho(k)$ and $h(k)$ the C3 reliability and consistency signals. The five conditions range from raw text to fully labeled claims:

- **C0:** the raw deck text x .
- **C1:** the bare claims $K(x)$, with no labels.
- **C2:** each claim with its label, $e_{C2}(x) = \{(k, v(k)) : k \in K(x)\}$.
- **C2-shuf:** C2 with labels shuffled onto the wrong claims, $\{(k, v(\pi(k)))\}$.
- **C3:** C2 plus the two signals, $\{(k, v(k), \rho(k), h(k))\}$.

C2-shuf is the key control. It matches C2 in form but breaks the link between each claim and its label, isolating whether the agent responds to the content of the labels or only to their presence.

3.3 Experiments.

We operationalize the three questions of this paper using two experiments and a probe:

- **RQ1 (impact on reasoning and decision).** Does verifying claims change how agents reason and invest? Experiment **E1** runs the canary decks through the full ladder, and the

contrast between the label-free conditions (C0, C1) and the labeled ones (C2, C3) is verification’s effect on propagation and $P(\text{invest})$. **E2** repeats the propagation measurement on the unmodified decks with real unsupported claims, a canary-free check that the pattern is not an artifact.

- **RQ2 (content versus presence).** When the decision changes, is it driven by the labels’ content or only their presence? Within **E1**, the C2-versus-C2-shuf contrast holds presence fixed while breaking content, isolating which one moves the decision.
- **RQ3 (verifiability of claims).** How much of a deck can be checked on the web at all, and when a claim is not confirmed, is the fact genuinely undocumented or does retrieval miss a documented one? A **coverage probe** measures the verification stage on its own, reporting how many claims are verifiable and splitting the failures into undocumented facts and retrieval misses on facts known to be public.

4 Method

4.1 Models and tooling.

The debate uses gpt-4o-mini for its many turns, the stronger gpt-4o to check claims, and Llama-3.3-70B-Instruct as judge. Two constraints drive these choices. First, to rule out look-ahead bias, every model in the decision path has a 2023 training cutoff, before our post-January-2024 deals, so none can have seen how a startup turned out. Second, to rule out self-preference bias, the judge comes from a different model family than the debaters and so cannot favor its own family’s arguments. Post-debate claim tracing uses Claude Haiku 4.5, which postdates the cutoff, so we keep it out of the decision path and use it only to measure propagation. For web search, Perplexity Sonar is chosen over alternatives like Exa for its reliable date-limited retrieval, and each claim is searched with a context block about the company so retrieval targets the right entity. Remaining settings follow the reference DIALECTIC configuration. To test whether debater capability drives the result, we also run the full ladder with a stronger debater, gpt-4o in place of gpt-4o-mini (Appendix B.5), and the headline findings are unchanged.

4.2 Sampling.

We drew two samples from a pool of startup pitch records from January 2024 onward: a 15-deal pilot sample followed by a stratified random sample of 150 further deals. The pool’s total size is undisclosed under the firm’s confidentiality policy. Of these, 115 yielded machine-readable decks (10 from the pilot and 105 from the stratified sample). Three empty decks were dropped. The 112 retained decks hold 7,360 atomic claims.

4.3 Canary injection.

We plant three canaries into each of the 40 modified decks, then remove 5 to protect personal details, leaving 115 in total. The other 72 decks are left unmodified for E2. Each canary is style-matched and consistently replaces the original facts. Appendix B.6 walks one canary through the full pipeline.

4.4 Debate procedure.

The debate runs two critique-and-refine iterations. Each round generates pro and con arguments, a devil’s advocate critiques them, and a judge scores them on 14 criteria. The criteria span local argument quality, rhetoric, and global quality, and are listed in full in Appendix B.2, together with the decoding configuration, which is held fixed across all conditions. The strongest arguments are kept and refined, and a final judge pass precedes the forced-choice decision.

4.5 Validity controls.

Across conditions the agent’s input is byte-identical apart from $e(x)$, so only the evidence varies. As a further guard against look-ahead, we limit web search to each deck’s date, and any residual leakage would only help flag a false claim, working against our finding rather than creating it. As a stronger check, we run the verifier (gpt-4o) closed-book, with no web access, to test whether it could recover the suppressed facts from memory rather than from retrieval (Appendix B.3).

4.6 Evaluation.

We compute two outcomes from each debate. For propagation, an aligner (Claude Haiku 4.5) maps each source claim to the argument’s load-bearing premises, and propagation is the fraction of ca-

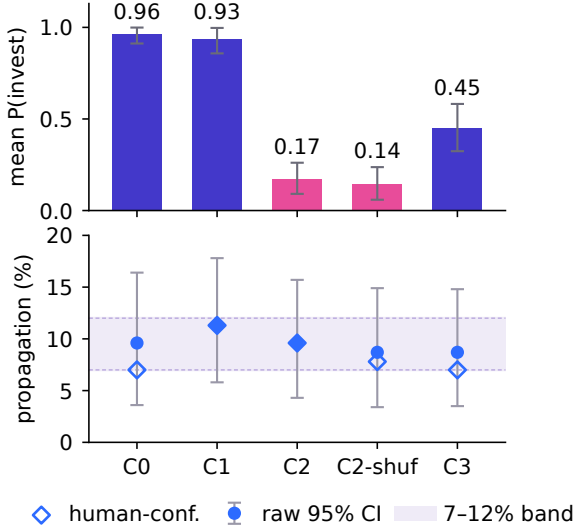


Figure 2: **Labels flip the decision but not propagation (E1).** Top: mean $P(\text{invest})$ by condition; real (C2) and shuffled (C2-shuf, pink) labels suppress it comparably. Bottom: canary propagation into load-bearing premises across the five conditions.

naries whose false content reaches such a premise:

$$\widehat{\text{prop}}(c) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\exists p \in \mathcal{P}_{ic} : p \text{ asserts } i], \quad (1)$$

where $N = 115$ and $\mathcal{P}_{ic}$ is the set of load-bearing premises for the deck containing canary i in condition c . A blinded human audit validates the alignment, and we report both raw and human-confirmed propagation rates. For the decision, we take the mean $P(\text{invest})$ from the agent’s forced-choice first-token log-probabilities, renormalized over the INVEST and PASS tokens. Canaries within a deck are not independent, so we analyze both outcomes at the deck level. We resample whole decks rather than individual canaries in a paired bootstrap, which yields the 95% confidence intervals used throughout. We report raw p-values for a small, pre-specified set of comparisons (Table 2).

5 Results

5.1 Impact on reasoning and decision (RQ1)

Figure 2 shows both endpoints, the decision ($P(\text{invest})$) and propagation, with the per-condition numbers in Table 1. Mean $P(\text{invest})$ is high without verification labels (C0, C1) and roughly $5.5\times$ lower once they appear (C2, top panel). The decision is near-binary per deck, so this collapse is a population flip from mostly invest

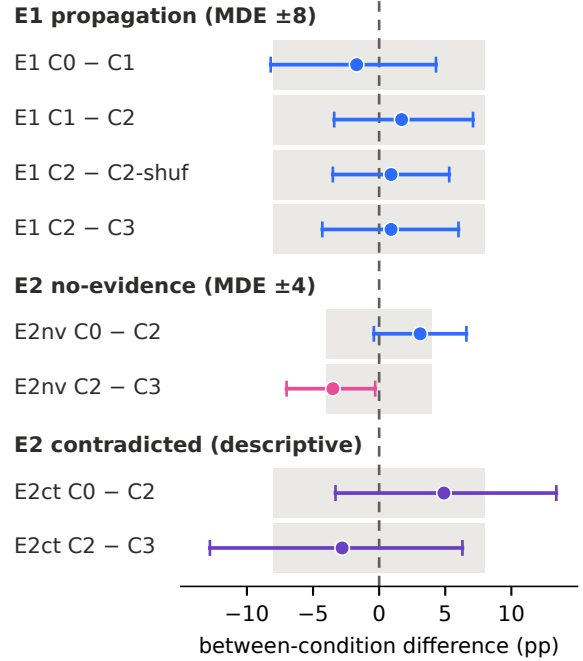


Figure 3: **No contrast beats its detectability floor. Only E2 no-evidence C2–C3 excludes zero, toward more propagation.** Each row is a pre-specified contrast, the between-condition difference in propagation (pp). Dashed line: zero. Gray band: $\pm$ MDE. **Blue:** no detectable effect. **Pink:** excludes zero. **Violet:** E2 contradicted, descriptive only.

to mostly pass (Figure 4, appendix). Canary propagation into load-bearing premises stays between 7% and 12% in every condition (bottom panel), and all four condition comparisons include zero under the paired bootstrap (Figure 3). Even the most favorable comparison rules out a propagation drop larger than about 8 percentage points.

Condition	Mean $P(\text{invest})$	Propagation raw [95% CI]	Human-confirmed
C0	0.964	9.6 [3.6, 16.4]	7.0
C1	0.934	11.3 [5.8, 17.8]	11.3
C2	0.170	9.6 [4.3, 15.7]	9.6
C2-shuf	0.142	8.7 [3.4, 14.9]	7.8
C3	0.452	8.7 [3.5, 14.8]	7.0

Table 1: E1 across the five-condition ladder (C0 raw deck, C1 claims, C2 labels, C2-shuf shuffled labels, C3 labels plus signals), seed 0, 115 canaries per condition. Propagation is the percent of canaries reaching a load-bearing premise, with paired-bootstrap 95% CIs, and human-confirmed keeps only audited links. Summarized in Figure 2.

5.2 Content versus presence (RQ2)

In Figure 2 (top panel), real labels (C2, mean $P(\text{invest})$ 0.170) and shuffled labels (C2-shuf, 0.142) suppress the decision comparably. Their paired difference is 0.028 (95% CI $[-0.10, +0.15]$, not statistically significant, paired permutation $p=0.66$), against a 0.76 $[0.66, 0.86]$ drop from the claims-only condition (Table 2 reports every decision-endpoint contrast and test). Shuffling is a real change, not a near-copy of the labels, since it reassigns the verdict label to a different value for 36% of claims, so the C2 and C2-shuf equivalence is genuine content-blindness and not an artifact of labels that barely differ. Propagation is also comparable across C2 and C2-shuf, and under both the agent invests in only 15% of decks (6 of 40). Adding reliability and consistency signals (C3) raises $P(\text{invest})$ relative to C2 without lowering propagation, though under a stronger gpt-4o debater this restoration largely vanishes (Appendix B.5). On average about 97% of labels do not support their claim, so no deck is dominated by supporting labels.

Contrast	$\Delta P(\text{invest})$	95% CI	p
C0 – C1	+0.03	$[-0.05, +0.12]$	0.44
C1 – C2	+0.76	$[+0.66, +0.86]$	<0.001
C2 – C2-shuf	+0.03	$[-0.10, +0.15]$	0.66
C2 – C3	-0.28	$[-0.40, -0.17]$	<0.001

Table 2: Paired contrasts on the decision (mean $P(\text{invest})$, seed 0, 40 decks). Δ is the mean per-deck difference, with deck-level paired-bootstrap 95% CIs and p from a paired permutation test, not corrected for multiple comparisons.

5.3 Verifiability of claims (RQ3)

About one in four claims is verifiable on the web. A 150-claim human audit puts the verifiable share at $26.5\% \pm 7.2\%$, near the 24.9% measured automatically. Among verifiable claims, most receive a no-evidence label (about 65%) and the rest are supported, contradicted, or conflicting. To separate retrieval misses from facts that are simply undocumented, we use the 82 of 115 canaries whose true value is publicly documented, the other 33 replacing a fact private to the company. We phrase each canary’s *true* value as a claim and send all 82 through the verification pipeline. Recall on these known-public facts is about 11% with results limited to each deck’s date and about 16% without that limit. A true fact is wrongly contradicted in

3.7% of cases under each deck’s date limit, rising to 18.3% without it. The verifier rarely contradicts the planted falsehoods either. Of the 115 planted, 69 survive claim extraction, 10 are routed verifiable and searched, and 1 receives a contradicted label, that one an off-target mis-route. None of the 82 web-verifiable canaries is contradicted, and none is ever labeled supported.

5.4 Replication and robustness

On the 72 decks with no planted canaries, claims labeled no-evidence reach load-bearing premises at about the same rate whether or not labels are shown (C0 versus C2), with no detectable difference between the two. Figure 3 plots all of our planned comparisons against zero, with exact values in Table 5 in the appendix. The gray band marks the minimum detectable effect (MDE), the smallest difference the test could resolve. A confidence interval inside that band is a bounded null rather than a true zero. Almost all contrasts are consistent with no effect, and the one exception, E2 no-evidence C2-versus-C3, points the wrong way, toward more propagation rather than less, and as a single significant result among eight comparisons it is likely a chance finding. In E2, the forest’s bottom group covers claims the verifier labels contradicted. These contrasts are descriptive only, since the verifier wrongly contradicts a true fact in 3.7–18.3% of cases, so a contradicted label need not mark a falsehood.

To check stability across random seeds, we reran all five conditions on all 40 canary decks under three seeds. Condition means barely move (across-seed standard deviation of the mean between 0.006 and 0.029) and propagation varies with a standard deviation of at most 2.3 percentage points. The findings we build on hold at every seed: the collapse once labels appear ranges from -0.76 to -0.84 per seed, and the C2-versus-C2-shuf differences straddle zero (Appendix B.4). Re-running the full ladder with a stronger debater (gpt-4o, Appendix B.5) reproduces both endpoints. The decision still collapses on label presence and is content-blind (C2 vs. C2-shuf -0.01 $[-0.02, 0.00]$), and propagation stays flat across conditions, if anything higher in absolute terms, so the dissociation is not an artifact of debater capability.

6 Discussion

The agent’s decision and its reasoning come apart. Verification moves the decision while leaving the reasoning untouched. Retrieval recovers only a small fraction of the facts already known to be true, the verifier rarely labels planted falsehoods as false, and reattaching each verdict to the wrong claim shifts the decision as much as a correct verdict does (RQ1, RQ2). The agent thus reacts to verdicts without registering which claim they attach to. Most claims about a private company cannot be checked on the open web, and the resulting no-evidence labels are read as a reason for caution rather than as evidence (RQ3, Section 5). This explains why verification helps elsewhere but not here. Where a trusted source documents the facts, checking against it is effective (Yue et al., 2026; Sakhno et al., 2026), but where none exists, labeling a claim false accomplishes little, and the debate weakens whatever evidence it is given (Kraidia et al., 2026). This pattern likely extends to other verify-then-argue systems with little to check against, and within our study it persists under a stronger debater. The constraint is coverage rather than the choice of retriever, so the prudent first step is to measure how much of a deck can be checked, here about a quarter of its claims (26.5%) and one in nine known-true facts (11%), before trusting what verification reports.

For practice, the implication is to keep verification and use it to settle what the open web can settle, and to redirect diligence effort onto the roughly three quarters of claims it cannot reach. That is what founder and reference calls are for, and knowing which claims check out makes those calls sharper. The canary methodology itself is portable. A fund can run it on its own deal flow to measure its local verifiability ceiling.

7 Conclusion

We tested what happens when an LLM agent checks the claims in a startup’s pitch deck before a panel of agents debates whether to invest. Checking made the agent more cautious but no better at telling true claims from false ones, because most claims about a private startup cannot be confirmed on the open web. A verification step can therefore resemble a safeguard while offering little real protection, and the limiting factor is coverage, which neither a better retriever nor a stronger agent closes. Extending agents beyond the deck, to the registries,

records, and networks that diligence actually draws on, is the open problem this ceiling makes precise, and a promising direction for future work.

Limitations

Our main generalization threat is the number of experimental axes we hold fixed. All 112 decks come from one European early-stage fund, so the verifiable fraction and retrieval ceiling describe that distribution, not pitch decks in general, and because the companies are predominantly European, where English-language coverage of private firms is sparser than for US/UK peers, the measured ceiling may be deflated. Recall relies mainly on Perplexity Sonar, with Exa covering only the uncapped arm because its date filter leaked an undated post-cutoff page in a smoke test. Each human audit was performed by a single co-author, and propagation may differ under other model families. Our conclusions describe this configuration, and we do not claim they transfer beyond it.

Our measurement components are validated rather than formally guaranteed. The verdict adjudicator is heuristic and offers no coverage guarantee, so it may reach a verdict without weighing every retrieved item, and we do not quantify how often decisive evidence is missed. The verifiability router is imperfect, with verifiable-class precision 50.0% (30/60), population-weighted recall 47.1%, and 4-class agreement 52.7% (Cohen’s $\kappa = 0.32$, stratified sample), which is why we do not rely on its raw output and instead calibrate the headline verifiable fraction against 150 stratified human labels, adjusting the raw 24.9% to $26.5\% \pm 7.2\%$ (95% CI). Propagation likewise rests on a model-based aligner, but its canary-link precision 87% (86/99) and asserts-falsified precision 91% (74/81) are confirmed by human audit, so the headline propagation numbers are human-verified. The lower-precision claim-link instrument, with precision 65.6% [55.3, 74.6] and miss rate 4.5% [1.8, 11.0] on a 180-pair blinded sample from a single annotator, is therefore confined to secondary metrics.

Two features of the design affect how the null should be read. Propagation is reported over all 115 canaries for comparability, but the agent only sees the 69 that survive claim extraction (60%) in the claim-based conditions C1–C3, while C0 sees the raw deck, where all but about 10 canaries fall within its truncation window. The labeled-condition rates are therefore conservative, and part

of the flatness across conditions reflects extraction dropping 40% of canaries before verification even runs. How much an input-side verification stage can help is thus capped by how many claims extraction and routing keep, not only by evidence coverage. Statistical power is also limited. With only 115 canaries the confidence intervals extend 5 to 8 points to either side of zero and we can detect only effects larger than about 8 points, so a smaller real reduction would not show up, and our null is an upper bound rather than a true zero. The same is true for the decision, where the C2-versus-C2-shuf interval $[-0.10, +0.15]$ narrows to $[-0.02, 0.00]$ in the gpt-4o re-run (Appendix B.5).

The decks and the system-of-record data used to cross-check them are the fund’s confidential internal data and cannot be released, so a full-corpus reproduction is not possible. We instead release the pipeline code, all prompts, the analysis and figure scripts, and the de-identified per-run result grids behind every experiment (transcripts removed, company identifiers hashed), including the three-seed replication, together with a small synthetic deck corpus that exercises the pipeline end to end from a clean checkout (repository linked in Section 1). The original decks, verbatim quote caches, and firm-side ingestion tooling are withheld.

Ethics Statement

The partner firm consented to this study, to publication of aggregate results, and to the release of de-identified per-run outputs with company identifiers hashed and all text removed. One author is affiliated with the firm and accessed its deal flow under that affiliation, the basis for both the consent and a disclosed conflict of interest. The decks come from third-party founders rather than the firm, so we report only aggregate results and name no startup. We quote no deck content and give no detail that could identify an individual company. All example claims are invented. Figure 1 illustrates with a well-known public company and fictitious figures, and the example tables and the appendix walk-through use the released synthetic corpus rather than deck content. The system was evaluated only on the firm’s past deals and never ran in a live workflow, influenced a real investment decision, or produced output seen by deal teams.

Of 120 drafted canary falsehoods we dropped 5 in quality control, removing any that could read as factual assertions about identifiable individuals.

The 115 injected falsehoods exist only in research copies of the decks and were never re-published, returned to the firm’s systems, or able to enter a live workflow. All human annotation was done by a co-author rather than crowdworkers, so no crowdwork-compensation or exposure concerns arise, and the single-annotator dependence is discussed in Limitations.

References

- Jae Yoon Bae, Simon Malberg, Joyce Ann Clarize Galang, Andre Retterath, and Georg Groh. 2026. DIALECTIC: a multi-agent system for startup evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 711–727.
- Xiao-Ping Chen, Xin Yao, and Suresh Kotha. 2009. Entrepreneur passion and preparedness in business plan presentations: a persuasion analysis of venture capitalists’ funding decisions. *Academy of Management journal*, 52(1):199–214.
- Colin Clark. 2008. The impact of entrepreneurs’ oral ‘pitch’ presentation skills on business angels’ initial screening investment decisions. *Venture capital*, 10(3):257–279.
- John H Cochrane. 2005. The risk and return of venture capital. *Journal of financial economics*, 75(1):3–52.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.
- Joyce Ann Clarize Galang, Edoardo Mosca, Simon Malberg, and Georg Groh. 2027. LLMs for venture capital investment: Approaches and open problems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 167–181. Springer.
- Paul A Gompers, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2020. How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.

- Claudia Anahí Guzmán-Solano, Bogdan Kaleb García Rivera, Ignacio Arroyo-Fernández, Mario Alberto Moreno Rocha, Cesar Felipe Martínez Cisneros, and Jesús Ulises Quiroz Bautista. 2025. User-centered approach for AI-powered pitch deck creation. In *Mexican International Conference on Artificial Intelligence*, pages 230–244. Springer.
- Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang. 2026. Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38542–38550.
- Laura Huang and Jone L Pearce. 2015. Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative science quarterly*, 60(4):634–670.
- Steven N Kaplan and Josh Lerner. 2016. Venture capital data: Opportunities and challenges. *Measuring entrepreneurial businesses: Current knowledge and challenges*, pages 413–431.
- Insaf Kraidia, Iyas Qaddara, Alhanof Almutairi, Nada Alzaben, and Samir Brahim Belhouari. 2026. When collaboration fails: persuasion driven adversarial influence in multi agent large language model debate. *Scientific Reports*, 16(1):11640.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.
- Zhongyang Liu, Haoyu Pei, Xiangyi Xiao, Xiaocong Du, Yihui Li, Suting Hong, Kunpeng Zhang, and Haipeng Zhang. 2025. Beyond isolated investor: Predicting startup success via roleplay-based collective agents. *arXiv preprint arXiv:2512.22608*.
- Xiaochen Ma, Guozheng Rao, Lina Xu, Xin Wang, Zaiming Fan, and Zhe Zhang. 2025. Guided and knowledgeable multi-agent debate for fact verification. *Expert Systems with Applications*, page 130103.
- Abdurahman Maarouf, Stefan Feuerriegel, and Nicolas Pröllochs. 2025. A fused large language model for predicting startup success. *European Journal of Operational Research*, 322(1):198–214.
- Cheng Meixin, Zeqi Zhou, Yijia Xiao, Weihao Xuan, Xiangru Tang, Xu Huang, Tinson Xu, Zerui Xu, Jure Leskovec, Yejin Choi, and 1 others. 2025. VCAF: a multi-agent framework for venture capital decision-making using synthetic startup data. In *NeurIPS 2025 Workshop: Generative AI in Finance*.
- Haoyu Pei, Zhongyang Liu, Xiangyi Xiao, Xiaocong Du, Suting Hong, Kunpeng Zhang, and Haipeng Zhang. 2025. The gaining paths to investment success: Information-driven llm graph reasoning for venture capital prediction. *arXiv preprint arXiv:2512.23489*.
- Andre Retterath and Reiner Braun. 2020. Benchmarking venture capital databases. *Available at SSRN 3706108*.
- Rabeya Tus Sadia and Qiang Cheng. 2026. Crunch-LLM: Multitask LLMs for structured business reasoning and outcome prediction. *Neurocomputing*, page 133754.
- Artem Sakhno, Daniil Tomilov, Yuliana Shakhvalieva, Inessa Fedorova, Daria Ruzanova, Omar Zoloev, Andrey Savchenko, and Maksim Makarenko. 2026. Financial transaction retrieval and contextual evidence for knowledge-grounded reasoning. *arXiv preprint arXiv:2603.15459*.
- U.S. Department of the Army. 2006. [Human intelligence collector operations](#). Technical Report Field Manual FM 2-22.3, Headquarters, Department of the Army. Appendix B: source-reliability (A–F) and information-credibility (1–6) ratings.
- Xisen Wang, Yigit Ihlamur, and Fuat Alican. 2025. Startup success forecasting framework (SSFF): a multi-agent framework for startup success prediction. In *NeurIPS 2025 Workshop: Generative AI in Finance*.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.
- Sean Williams, Clay Spinuzzi, and Curtis Newbold. 2020. Toward a heuristic for teaching the visual rhetoric of pitch decks: A pedagogical approach in entrepreneurship communication. *Communication Design Quarterly Review*, 7(4):4–16.
- Cheng Xiong, Gengfeng Zheng, Xiao Ma, Chunlin Li, and Jiangfeng Zeng. 2025. DelphiAgent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing & Management*, 62(6):104241.
- Sichao Xiong and Yigit Ihlamur. 2023. Founder-GPT: Self-play to evaluate the founder-idea fit. *arXiv preprint arXiv:2312.12037*.
- Ling Yue, Chaoqian Ouyang, Hang Xu, Ruijun Huang, Yuchen Liu, Libin Zheng, Wei Liu, Shaowu Pan,

Shimin Di, and Min-Ling Zhang. 2026. Fac-treview: Evidence-grounded peer review with execution-based claim verification. *arXiv preprint arXiv:2604.04074*.

Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation. *arXiv preprint arXiv:2508.20131*.

A Adaptations

Both parts of verify-then-argue are built for this work. The verification pipeline (extraction, routing, search, verdicts, and the reliability and consistency signals) is ours, assembled from established ideas rather than new inventions: breaking text into atomic claims and checking each (Wei et al., 2024), grounding answers in retrieved evidence and self-checking them (Lewis et al., 2020; Dhuliawala et al., 2024), automated fact-checking (Guo et al., 2022), and rating sources by reliability (U.S. Department of the Army, 2006). The multi-agent debate is our reimplementation of the DIALECTIC design (Bae et al., 2026), keeping its pro and con roles and a devil’s advocate. We build on DIALECTIC because it is the only startup-screening debate validated against real investment decisions. What is new is the implementation and the use of verification as a step before the debate. Seven implementation choices affect measurement, and we disclose them here:

- **Shared token cap.** One token budget shared by the page gate, the extractor, and the audit truncated responses on dense pages and silently dropped claims. We split it into per-stage budgets.
- **Date cutoffs.** Different stages had conflicting hard-coded cutoff dates. We replaced them with one cutoff used by both the prompts and the search filter.
- **Few-shot leakage.** Few-shot prompt examples that named companies from the evaluation set were replaced with neutral, fictional ones.
- **Source matching.** The source-reliability check matched domain text loosely, so any domain containing “gov” counted as government. We made the match exact.

- **Verdict model.** An early version of our client pointed at the wrong provider and model. We pinned the intended model at every stage.
- **Judge family.** The judge is from a different model family than the debaters, to avoid self-preference.
- **Question-tree.** DIALECTIC’s prompts reference numbered question-and-answer pairs. Because C0 feeds raw deck text with no indexable units, we replaced the Q&A wording with the evidence block and dropped the index tracking, so prompts stay identical across conditions.

B Method details

This appendix gives the details left out of Section 3 and Section 4. It repeats the measurement-relevant corrections from Appendix A so it stands on its own.

B.1 Claim verification and extraction

Figure 6 shows the full verify-then-argue pipeline. The verify stage runs as a pipeline. (1) **Page gate.** We keep only deck pages that contain verifiable statements. (2) **Extraction.** gpt-4o-mini breaks each page into atomic claims and ties each to a quote from the slide, and a programmatic check voids quotes not found in the page text, which constrains invented claims. Token budgets are 400/4,000/4,000 for the gate, extractor, and audit, after an earlier shared 1,000-token cap was found to drop claims on dense pages. (3) **Deduplication.** We merge claims that recur across slides. (4) **Routing.** We sort claims into verifiable and not, setting aside opinions, projections, and value judgments so that only verifiable claims are searched. Their share is the verifiable fraction in Section 5, and Table 6 breaks all 7,360 claims down by routing category. (5) **Search.** We search each verifiable claim five times through Perplexity Sonar with reworded queries. (6) **Verdict.** gpt-4o reads the retrieved evidence and judges each search. (7) **Source reliability.** We score each source by how trustworthy its domain is, following Admiralty-style grading (U.S. Department of the Army, 2006), and combine these into the C3 reliability signal $\rho(k) \in [0, 1]$. (8) **Consistency.** We measure how much the five searches agree, giving the C3 signal $h(k) \in [0, 1]$ that is high when they agree and low when they conflict. Both signals exist only for searched claims. (9) **Label.** We combine the verdicts into one label per

claim, supported, contradicted, conflicting, or no-evidence. The rule is heuristic, and the resulting labels are not calibrated, so we attach no guaranteed error rate to them.

Claim-extraction quality. Extraction follows the decompose-then-verify approach (Wei et al., 2024), and we check it three ways. *Volume*: 7,929 raw claims, of which 7.2% are near-duplicates, leaving 7,360 (median 58 per deck). *Faithfulness*: 95.7% pass the extractor’s own audit, 4.1% are flagged as not found in the text, and 35.6% of supported claims carry an exact quote, with the rest paraphrasing the slide. This is a model audit plus the quote check, not a human benchmark. *Recall*: of the 115 planted canaries, 69 (60%) survive extraction, and 93% of all claims are factual or numeric. Claims are atomized but not fully self-contained: each is searched with a short company description so retrieval finds the right startup, but a claim whose truth depends on dropped context can still be mis-checked. A human faithfulness audit on a sample is future work.

B.2 Debate implementation

We cap the pro and con sides at 100 words each. The earlier version capped them unequally, which could have let argument length sway the judge. The debate also produces an older score-based decision alongside the invest probability we report, and the two disagreed in early testing.

The judge scores arguments on DIALECTIC’s 14 argument-quality criteria. Four cover local quality (local acceptability, local relevance, local sufficiency, cogency), six cover rhetoric (credibility, emotional appeal, clarity, appropriateness, arrangement, effectiveness), and four cover global quality (global acceptability, global relevance, global sufficiency, reasonableness).

Table 7 lists the decoding configuration. Decoding is held fixed across all conditions, so only the evidence block varies between runs. The only calls left at provider defaults are the search provider’s retrieval call and the Haiku 4.5 aligner (Table 7).

B.3 Leakage-control specifics

Training cutoffs, checked against provider documentation, are October 2023 for gpt-4o-mini (extraction, routing, debate) and gpt-4o (verdicts), and December 2023 for Llama-3.3-70B-Instruct (judge). Only Sonar limits results by date reliably, since the Exa date filter leaks (below). Every deck

postdates these cutoffs, so the models cannot have seen how a startup turned out. As a direct check, the closed-book probe with no web access recovers the true value for only 7 of the 115 suppressed facts (6 of the 82 web-verifiable, 1 of the 33 private), so only those 7 could be flagged from memory rather than from retrieval. In the web-enabled pipeline itself none of the 82 web-verifiable canaries is contradicted (Section 5), and the single contradicted canary is an off-target private one. A second provider does not lift the retrieval ceiling either. In a smoke test, Exa’s date filter leaked, returning results from after each deck’s cutoff date, so we could not rely on it and ran Exa uncapped. Even uncapped, Exa confirms only 25.6% of the 82 web-verifiable true values, and the union of Sonar and Exa reaches only 28.0%, so the low recall is not a single-provider artifact. Older models cost some capability, which we accept as the price of this control.

B.4 Statistical-protocol details

Seed robustness is assessed on the full grid, with all five conditions re-run on the 40 canary decks at three seeds (Table 8). Condition means are stable, with an across-seed standard deviation of the mean between 0.006 and 0.029, and both load-bearing contrasts replicate at every seed. The collapse $C2-C1$ is -0.764 , -0.835 , and -0.797 at seeds 0, 1, and 2, and the equivalence $C2-C2\text{-shuf}$ is $+0.028$, -0.055 , and $+0.049$, with a deck-level paired-bootstrap 95% CI that includes zero at every seed. The headline values in §5 are seed-0 estimates, and the seed-averaged means match them to within 0.02.

Per-deck estimates are far noisier, because decks near the decision boundary swing between invest and pass across seeds (Table 9). The across-seed range of $P(\text{invest})$ for a single deck reaches 1.00, and the conditions differ mainly in how many decks sit near that boundary. In $C0$ and $C1$, 33 of 40 decks move by less than 0.05 across seeds while 4 swing by more than 0.5. In $C2$ the split is 15 stable and 12 swinging, in $C2\text{-shuf}$ 19 and 8, and in $C3$ 8 and 16. The aggregate condition means and the paired contrasts we report are therefore stable while individual deck values are not, which is why we interpret no per-deck value on its own. Propagation shows the same pattern, flat at every seed with an across-seed standard deviation of at most 2.3 percentage points (Table 10).

Verification costs about \$0.70 per deck and a debate about \$0.03 per run. Input-token use scales

with the evidence block, while debate duration does not vary by condition (Figure 5).

B.5 Stronger-debater model (gpt-4o)

To test whether our results depend on the limited gpt-4o-mini debaters, we re-ran the full ladder on the 40 canary decks (seed 0) with gpt-4o as both debater and decision model, keeping the Llama-3.3-70B judge, the labels, and all prompts fixed. A newer model would break the cutoff control, so gpt-4o is the strongest choice that stays leakage-safe. The findings hold and in fact strengthen (Table 11). The decision still collapses on label presence (mean $P(\text{invest})$ about 0.95 at C0 and C1 versus 0.003 at C2) and still ignores label content, with real and shuffled labels now differing by at most 0.02. Propagation stays flat across conditions and roughly doubles in level (about 20% versus 9%), so a stronger debater builds longer arguments that pull in more false content, not less. Only the partial C3 restoration disappears: the stronger reader does not regain confidence when the reliability and consistency signals are added (C2–C3 shrinks from -0.28 to -0.05). The headline gpt-4o-mini setup is therefore conservative for the decision and an underestimate for propagation.

B.6 A canary walk-through

This section traces one planted canary through the pipeline end to end. The deck belongs to Aquilon Robotics, an invented company from the released synthetic corpus, so no fund data appears. The debate numbers come from running the released quickstart on this single deck at seed 0. The mechanics are identical for the real corpus.

Injection. The injected twin replaces one traction line of the original deck with the false claim “Aquilon was named Intralogistics Robot of the Year 2025 by the European Warehouse Automation Council.” The award and the awarding body are invented. The claim has the canonical canary shape: if it were true, it would leave a public trace.

Extraction and the claim stream. The corpus claims file for this deck holds 18 atomic claims, with the canary among them verbatim, so the planted falsehood enters the claim stream that conditions C1 to C3 read. Extraction is the lossiest step for canaries in general, with 60% surviving it in the real corpus (Appendix B.1), and this walk-through follows the path of a canary that survives, the case the labeled conditions actually see.

Condition	$P(\text{invest})$	Decision	Canary echoed
C0	1.00	invest	yes
C1	1.00	invest	yes
C2	0.32	pass	no
C2-shuf	0.10	pass	yes
C3	0.99	invest	no

Table 3: The traced canary deck under the five conditions (seed 0). “Canary echoed” marks whether the planted award appears in the debate’s final arguments.

Routing. In the claims file the canary is routed verifiable, since a named award from a named body is externally checkable, and only verifiable claims are searched.

Search and label. Verifiable claims are searched five times with reworded queries. An invented award leaves no trace to find, and the canary’s entry carries the no-evidence label with source reliability 0.0 and consistency 1.0, the pattern the verifier produces when the searches return nothing. A live rerun of the released verification stage on this invented company behaves the same way at the deck level: none of its verifiable claims comes back supported, and no-evidence is the most common verdict, matching the real corpus, where about 65% of verifiable claims receive no-evidence (Section 5).

Debate. Table 3 shows the same deck under the five conditions. Without labels (C0, C1) the agent invests at 1.00 and the false award is echoed as fact in the final arguments. The con side even concedes it: “Despite Aquilon’s accolade as the Intralogistics Robot of the Year 2025, the company faces substantial challenges...”. Once labels appear (C2, C2-shuf) the decision collapses. Under shuffled labels the award line is still repeated while the argument cites a general lack of verifiable evidence, showing that the caution is not anchored to the canary’s own label. Under C3 the reliability and consistency signals restore the invest decision on this deck, a stronger version of the partial restoration in Section 5. One deck at one seed illustrates mechanics rather than evidence, and per-deck values are noisier than the condition means in Section 5.

C Deck-level claim features and the decision

Decks are written almost entirely in the register of fact: 92.8% of the 7,929 raw extracted claims are factual or numeric (4,718 and 2,642 respectively), and the register is near-uniform across decks, with

Extracted claim	Routing	Verdict
Company A was incorporated in 2023.	verifiable	supported
Operators lose 12% of picking hours to manual handling.	verifiable	conflicting
Company A's CTO led perception at a listed logistics firm.	verifiable	contradicted
Company A has deployed six paid pilots.	verifiable	no-evidence
Company A's gripper achieves a 99.2% pick success rate.	unverifiable	not searched
Company A will reach EUR 4M ARR by 2027.	inference	not searched
Company A's approach is uniquely defensible.	normative	not searched

Table 4: Example extracted claims with routing and verdict, one per category. Synthetic illustrations of the label taxonomy, not deck content (Ethics Statement).

Group	Contrast	Δ	95% CI
E1 prop. (MDE ± 8)	C0 - C1	-1.7	[-8.2, +4.3]
	C1 - C2	+1.7	[-3.4, +7.1]
	C2 - C2-shuf	+0.9	[-3.5, +5.3]
	C2 - C3	+0.9	[-4.3, +6.0]
E2 no-ev. (MDE ± 4)	C0 - C2	+3.1	[-0.4, +6.6]
	C2 - C3 [†]	-3.5	[-7.0, -0.3]
E2 contr. (descr.)	C0 - C2	+4.9	[-3.3, +13.4]
	C2 - C3	-2.8	[-12.8, +6.3]

Table 5: Every pre-specified contrast plotted in Figure 3. Δ is the difference in canary propagation (E1) or no-evidence propagation (E2) in percentage points, with deck-level paired-bootstrap 95% CIs, and MDE is the smallest detectable effect. [†] marks the only contrast that excludes zero. The E2 contradicted contrasts are descriptive only.

Routing category	Claims	Share
Unverifiable	4,738	64.4%
Verifiable (searched)	1,820	24.7%
Inference	442	6.0%
Normative	360	4.9%

Table 6: Routing categories over all 7,360 claims from the 112 decks, canary twins included. The 24.9% in Section 5 is the same measure on the 72 unmodified decks.

Component	Model	Temp.
Debaters (generate, critique, refine)	gpt-4o-mini	0.5
Judge	Llama-3.3-70B-Instruct	0.0
Decision readout (log-probs)	gpt-4o-mini	0.0
Verification (extract, route, verdict)	gpt-4o-mini / gpt-4o	0.0
Retrieval (evidence search)	Sonar	default

Table 7: Decoding configuration, held fixed across all conditions. The only calls at provider defaults are the retrieval call and the Haiku 4.5 aligner, which is measurement-only and off the decision path.

Mean $P(\text{invest})$	Seed 0	Seed 1	Seed 2	Mean
C0	0.964	0.975	0.967	0.969
C1	0.934	0.953	0.962	0.950
C2	0.170	0.118	0.166	0.151
C2-shuf	0.142	0.173	0.116	0.144
C3	0.452	0.436	0.420	0.436
C2 - C1	-0.764	-0.835	-0.797	-0.798
C2 - C2-shuf	+0.028	-0.055	+0.049	+0.008

Table 8: Across-seed stability of the decision endpoint (mean $P(\text{invest})$, 40 canary decks, three seeds). The collapse (C2 - C1) and the equivalence (C2 - C2-shuf) replicate at every seed, and the equivalence CI includes zero at each.

Condition	Med. range	Max range	<0.05	>0.5
C0	0.000	0.991	33/40	4/40
C1	0.000	0.975	33/40	4/40
C2	0.148	0.987	15/40	12/40
C2-shuf	0.059	1.000	19/40	8/40
C3	0.354	0.997	8/40	16/40

Table 9: Per-deck across-seed variability of $P(\text{invest})$ (range over three seeds per deck). The last two columns count decks whose range stays under 0.05 and decks whose range exceeds 0.5. Most decks are stable or swing fully, so deck-level values are noisy near the decision boundary while the 40-deck aggregates are not.

Propagation %	Seed 0	Seed 1	Seed 2	Mean
C0	9.6	7.8	7.8	8.4
C1	11.3	11.3	11.3	11.3
C2	9.6	9.6	8.7	9.3
C2-shuf	8.7	9.6	13.0	10.4
C3	8.7	12.2	10.4	10.4

Table 10: Across-seed stability of the propagation endpoint. Propagation stays flat at every seed, and every pre-specified paired contrast includes zero.

Endpoint	Contrast	gpt-4o-mini	gpt-4o
$\Delta P(\text{invest})$	C1–C2	+0.76 [.66, .86]	+0.93 [.86, .99]
	C2–C2-shuf	+0.03 [−.10, .15]	−0.01 [−.02, .00]
	C2–C3	−0.28 [−.40, −.17]	−0.05 [−.10, −.02]
prop. (pp)	C1–C2	+1.7 [−3.4, 7.1]	+7.0 [−0.9, 15.0]
	C2–C2-shuf	+0.9 [−3.5, 5.3]	−0.9 [−9.3, 7.1]
	C2–C3	+0.9 [−4.3, 6.0]	+0.9 [−4.4, 6.9]

Table 11: **Stronger-debater robustness.** Pre-specified contrasts under the gpt-4o-mini setup versus a gpt-4o re-run (seed 0, 40 canary decks, same labels, judge, and prompts), with deck-level paired-bootstrap 95% CIs.

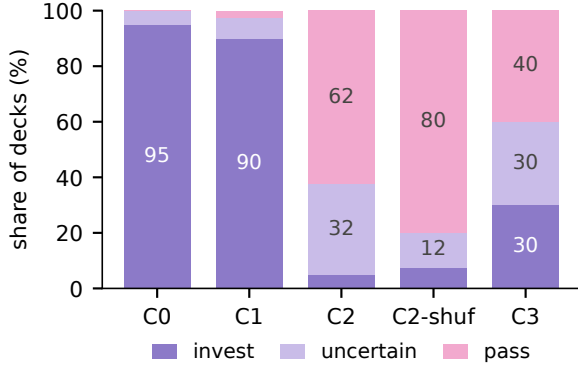


Figure 4: **Adding labels drives a near-binary population flip**, from mostly invest (C0, C1) to mostly pass (C2, C2-shuf). Bins: invest ($P(\text{invest}) > 0.9$), uncertain, and pass ($P(\text{invest}) < 0.1$), seed 0 and 40 decks. Mean $P(\text{invest})$ in Figure 2 averages over this flip.

a per-deck share of mean 0.92 (SD 0.07), ranging from 0.63 to 1.00 over the 112 decks that yield claims. We test whether that register, or any other deck-level claim feature, carries the decision. It does not. Across the 40 canary decks, no feature correlates with $P(\text{invest})$ on the raw deck (Spearman ρ between -0.14 and $+0.09$, all $p \geq 0.40$) or with the per-deck collapse under verification labels, $P(\text{invest})$ at C0 minus C2 (ρ between -0.17 and $+0.17$, all $p \geq 0.29$; Table 12). Two range restrictions bound what this null can show: C0 $P(\text{invest})$ sits near ceiling (mean 0.964), and the register itself barely varies across decks. Within those limits, the agents invest in nearly every deck regardless of how factual it sounds, and how far a deck falls under verification labels is unrelated to its claim mix: the factual register is a property of the genre, not a measurable lever on the decision.

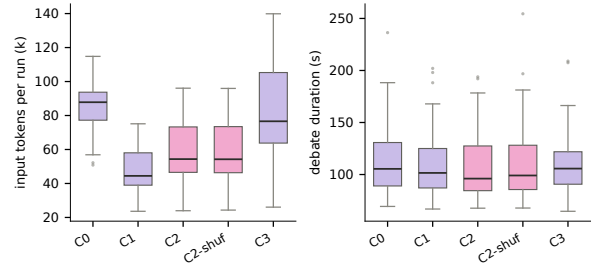


Figure 5: **Input tokens track evidence-block size, and debate duration is flat across conditions.** Left: input tokens per run, lowest for claims (C1) and highest for the raw deck (C0) and C3. Right: duration, medians about 100 s. Pink: C2 and C2-shuf. Output tokens, cost, and agent-pass count are constant and not shown.

Deck feature	$P(\text{invest})$ (C0)		Collapse	
	ρ	p	ρ	p
Factual/numeric register	+0.09	0.59	+0.17	0.30
Verifiable share	+0.06	0.69	−0.17	0.29
Supported share	+0.05	0.75	+0.13	0.44
Contradicted share	+0.09	0.60	−0.07	0.67
No-evidence share	+0.08	0.62	−0.08	0.61
Conflicting share	−0.14	0.40	−0.00	0.98

Table 12: Deck-level claim features against the decision on the 40 canary decks (seed 0). Verifiable share is the fraction of claims routed verifiable, and verdict shares are computed among searched claims. ρ is the Spearman rank correlation of each feature with $P(\text{invest})$ on the raw deck (C0) and with the per-deck collapse under verification labels (C0 – C2); p from a two-sided permutation test ($B=10,000$), uncorrected for multiple comparisons. No correlation is distinguishable from zero.

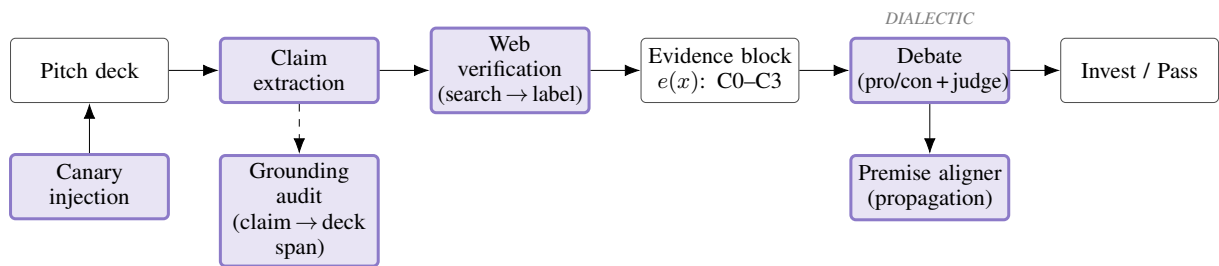


Figure 6: **Verify-then-argue pipeline.** The pipeline runs left to right. Purple marks the parts we build, the debate (gray label) is reimplemented from DIALECTIC, white boxes are inputs and outputs, and only the evidence block $e(x)$ varies across conditions C0 to C3.