Split-Perception: Measuring Divergence in How Large Language Models Assess Company Credibility

Andrew Pomazkov
VeritasLinks, Inc., Irvine, California
andrew@veritaslinks.com

July 2026

Abstract

Large language models are increasingly the first source consulted when an investor, analyst, or counterparty evaluates a company. Yet different models, asked identical questions about the same firm, routinely return materially different assessments — a phenomenon we term split-perception. Drawing on a dataset of AI-generated assessments covering more than 5,000 companies and several million individual model responses collected between 2025 and 2026, we introduce a structured methodology for scoring AI-mediated company perception on a 300–870 scale modeled on consumer credit scoring, decompose the score into visibility, positioning, and risk components, and quantify cross-model divergence. We find that divergence between frontier models on the same company frequently exceeds [XX] points — in one longitudinal case study, 115 points between the highest- and lowest-scoring models on a single measurement date. We argue that split-perception constitutes an unpriced information risk for M&A due diligence, KYB review, and credit assessment workflows that rely, formally or informally, on AI-assisted research, and we outline the conditions under which perception divergence is stable versus noise.

Keywords: large language models, company credibility, perception scoring, due diligence, KYB, information risk, generative engine optimization

JEL Classification: D83, G24, G32, G34, M31, O33

Data underlying this paper were collected via the VeritasLinks measurement platform (veritaslinks.com). Working paper — not peer reviewed. Comments welcome.

1. Introduction

When a venture associate prepares for a first call with a founder, when a KYB reviewer screens a new counterparty, or when a buyer compares vendors, there is a growing probability that the first research step is a prompt to a large language model rather than a visit to the company's website. The answer the model returns is a compression of everything it has absorbed about the firm — its site, review platforms, press mentions, forums, structured data — synthesized into a short verbal judgment. That judgment is delivered at precisely the moment first impressions form, and it is produced without the company's participation or knowledge.

This paper examines a specific and, to our knowledge, unmeasured property of these AI-mediated assessments: they disagree with each other. Two frontier models, asked identical buyer-shaped or diligence-shaped questions about the same company on the same day, can return assessments that differ not merely in tone but in substance — one describing a defensible platform, another a commodity tool; one naming the company among category leaders, another omitting it entirely. We call this split-perception. Where traditional search returned a ranked list a reader could inspect, an AI answer is a single synthesized verdict; when verdicts diverge across models, the "fact of the matter" about a company's standing becomes model-dependent.

Our contributions are threefold. First, we introduce a reproducible methodology for measuring AI-mediated company perception: a structured battery of 1,000–1,500 prompts per company, executed across seven frontier models, aggregated into a composite score on a 300–870 scale deliberately modeled on consumer credit scoring (Section 3). Second, using assessments covering more than 5,000 companies, we document the distribution and structure of cross-model divergence (Section 5). Third, through a 38-run longitudinal case study of a single firm, we distinguish stable perception splits from sampling noise and show that divergence is systematically associated with the citation and authority profile of the company's public footprint (Section 6).

The practical stakes are asymmetric. A company assessed favorably by the model its evaluator happens to use experiences no friction; the same company assessed unfavorably by a different model loses opportunities it never learns existed. For risk and compliance workflows — M&A screening, credit assessment, KYB review — that increasingly incorporate AI-assisted research, split-perception is an unpriced input risk: the conclusion of a screening step depends on an arbitrary tooling choice. Section 7 develops these implications.

2. Background and Related Work

Three literatures intersect here. The first concerns search-mediated firm visibility. A long line of work in information retrieval documents that position in a ranked list powerfully shapes attention and trust: users disproportionately select and believe top-ranked results even when relevance is held constant (Pan et al., 2007). AI answers change the structure of this surface: instead of a ranked list with a long tail, the model names a small set of entities and stops — a winner-take-few regime in which omission is invisible to the omitted party. Aggarwal et al. (2024) formalize this setting as "generative engines" and show experimentally that targeted content interventions can raise a source's visibility in AI-generated answers by up to forty percent — evidence that the surface is both consequential and manipulable. That work, and the practitioner discipline it named, treats model output as a single surface to be optimized; we instead treat the models as a set of potentially disagreeing judges.

The second literature concerns LLM evaluation and inter-model disagreement. Large-scale evaluation efforts document that frontier models diverge substantially and systematically across factual, reasoning, and judgment tasks (Liang et al., 2023), and studies of model-as-judge protocols show that LLM raters exhibit stable, rater-specific biases — position, verbosity, and self-enhancement effects — rather than random error (Zheng et al., 2023). Our setting differs in that the object of judgment is a real-world entity with an observable public footprint, allowing divergence to be related to properties of the underlying evidence rather than to task ambiguity alone.

The third is the credit-scoring tradition. Composite scores on a bounded scale, decomposed into named components with monotonic contribution, have proven durable in consumer finance precisely because they translate a high-dimensional evidence base into a decision-ready quantity with an auditable structure (Hand & Henley, 1997; Thomas, Crook & Edelman, 2017). The methodology in Section 3 imports this design — including the 300–870 range and an explicit risk-deduction component — into the perception domain. The author's prior work implementing the scoring model of a national credit bureau (UBKI, Ukraine, 2014) informs this design choice directly.

Finally, the question of which sources a model treats as authoritative connects to the older literature on credibility assessment of online information (Metzger, 2007) and to its operationalization in search quality rating — the experience–expertise–authoritativeness–trust framework used by human quality raters (Google, 2025). Section 5 shows that cross-model divergence concentrates precisely where these authority signals are weak or conflicting.

3. Methodology: The VeritasScore Framework

3.1 Model set and prompt battery. Each company assessment executes a structured battery of 1,000–1,500 prompts through a two-tier architecture. The measured tier consists of five production-deployed commercial judge models — ChatGPT, Claude, Grok, Perplexity, and DeepSeek — whose responses are the object of measurement. A sixth candidate judge, Gemini, was evaluated and excluded: it failed the within-model consistency criterion described in Section 5.4, returning contradictory verdicts to identical prompts within a single batch, so its divergence from other judges is not interpretable as perception. Judge inclusion thus requires a minimal test–retest reliability, assessed before a model enters the set. The orchestration tier is a fine-tuned Mistral 7B model that conducts the assessment but does not judge: it generates the adaptive elicitation probes of Section 3.3, parses judge responses into the signals of Section 3.4, assembles the firm's digital footprint, and computes the composite score and recommendations. The orchestrator was fine-tuned on the archival corpus of Section 4 — a deliberate loop in which the first pipeline generation's data trained the model that operates the second.

The separation of tiers matters methodologically. Judge outputs are the measured object; the orchestrator is measurement apparatus. Because a single fixed orchestrator generates probes, parses, and aggregates for every judge and every company, any bias it introduces is common across models and firms — and between-model divergence, the central quantity of this paper, is a difference in which common instrumentation effects cancel.

The battery comprises three modules. The brand-perception module mirrors real evaluator behavior through four query families: buyer-shaped queries ("best tools for X"), category queries ("leading companies in Y"), head-to-head comparisons ("A versus B for a small team"), and diligence queries ("risks and weaknesses of Z"). The synthetic buyer-panel module (Section 3.7) places judges in an explicitly evaluative role and records how they reason when asked to decide rather than describe. The narrative-review module (Section 3.8) assesses the firm's narrative document separately from its brand, enabling the brand/narrative decomposition used in Section 6. Each query family is executed in independent sessions, so that within-model sampling variance can be estimated separately from between-model divergence.

3.2 Battery size and stochasticity. Battery size is itself a measurement-design decision. Individual judge responses are stochastic: the same prompt, re-issued in a fresh session, can return materially different content, name different competitors, or omit the firm entirely. At small sample sizes this stochasticity dominates the measurement. In operational testing with reduced batteries, run-to-run dispersion of the composite score was large and exhibited no discernible pattern: scores could not be distinguished from sampling noise, and no stable per-model signal emerged. Dispersion contracts as the battery grows, and composite scores stabilize in the 1,000–1,500-prompt range, which we adopt as the operating point.

We characterize this threshold operationally rather than through formal convergence statistics, which we leave to future work. What matters for the present design is the resulting separation of scales: Section 5.4 treats the residual within-model variance at the operating point as the noise floor, and between-model divergence is judged against it — with the qualification, documented there, that the noise floor is itself model-specific.

3.3 Conversational elicitation. Prompting is conversational rather than single-shot. Initial responses frequently omit signals the measurement requires: a model may name competitors without ordering them, render a judgment without identifying the sources it relied on, or describe a category without placing the firm within it. When required signals are missing, the protocol issues follow-up probes within the same session until the signal set is complete or a fixed probe budget is exhausted.

Probes are generated adaptively rather than drawn from a fixed script: an elicitation layer inspects the incomplete response and formulates the follow-up appropriate to what is missing and to how the model has framed its answer. Adaptivity is bounded by a fixed elicitation policy with an explicit non-steering objective. Generated probes are constrained to neutral phrasing; they may not introduce facts about the company that the model has not itself produced; and they may not signal a preferred or expected answer. Full session transcripts, including every generated probe, are logged, so elicitation is auditable after the fact.

Two properties of this design bear noting. First, reproducibility attaches to the elicitation policy rather than to prompt strings. Because model responses are themselves stochastic, string-level reproducibility is unavailable even to single-shot designs; the appropriate invariance is at the level of the procedure and the aggregate, which the fixed policy and the battery size of Section 3.2 provide. Second, single-shot batteries systematically under-extract precisely the positioning signals on which cross-model divergence concentrates (Section 5.3). Conversational elicitation is therefore not an efficiency optimization but a precondition for measuring the quantity of interest: without it, an absent positioning signal is ambiguous between “the model holds no view” and “the model was not asked to complete its view.”

3.4 Component scores. Elicited responses are parsed into measurable signals — mention rate, mention position, sentiment, category assignment, competitor co-mentions, and cited sources — and aggregated into six named components. Brand visibility captures whether and how prominently the firm is named across query families. Market position captures category assignment and leadership attribution. Customer perception aggregates sentiment and panel preference outcomes. Content presence measures the footprint of answer-shaped content available for models to draw on. Growth potential reflects the forward-looking language models attach to the firm. An explicit risk deduction subtracts for objections, negative attributions, and instability surfaced anywhere in the assessment. Component weights are fixed across companies within a scoring vintage. [EXPAND — input signals and weights per component, or a qualitative description if the weight vector is proprietary.]

3.5 Composite scale and reporting bands. Components aggregate to a composite on a 300–870 scale, deliberately isomorphic to consumer credit scoring: bounded, monotonic, decomposable, and coarse enough to resist over-interpretation of small differences. The composite is reported in three bands with fixed thresholds at 500 and 700: scores of 300–499 are labeled Needs Improvement, 500–699 Good, and 700–870 Excellent. Band thresholds are held constant across companies and scoring vintages, so band membership is comparable across firms and over time. Empirically, the distribution is bottom-heavy. Of the 139 companies in the deduplicated, fully scored population (latest completed assessment per company), 87.1% fall in the Needs Improvement band, 12.9% reach the Good band, and none has yet reached the Excellent band: the maximum observed score is 564, and the median is 437. The unoccupied upper region of the scale is headroom by design rather than a calibration artifact — a property this scale shares with the consumer credit scores it is modeled on.

3.6 Divergence measures. For each company we compute per-model composite scores on the shared 300–870 scale and define split-perception as the range and dispersion of these per-model scores on a single measurement date. Split-perception is distinguished from temporal instability — the variation of the same model's score across repeated runs — which Section 6 measures on a fixed company under a fixed methodology vintage. A company can, in principle, exhibit any combination of the two: stable agreement, stable disagreement, or noisy scores that preclude either conclusion; Section 3.2's operating point exists to make the third case identifiable.

3.7 Synthetic evaluation panels. Assessments additionally include a structured panel of synthetic buyer personas — models conditioned on distinct evaluator roles — that interrogate the company's positioning as a purchasing or investment decision rather than a descriptive question. The panel records objections raised, preference choices between the firm and alternatives, and confidence in those choices, summarized in idea- and confidence-level indices. Panel outputs feed the customer-perception component and the risk deduction. We treat panel results as a measurement of how models reason about the company when placed in an evaluative role — an increasingly common role in practice, per Section 1 — and not as a substitute for human research subjects. [EXPAND — brief protocol description: persona construction, question sequence, and how objections are coded.]

3.8 Narrative-review module. The narrative-review module applies the same multi-model apparatus to a narrative document — the account of category, opportunity, and viability a firm presents to a deciding audience. The document may be an investor-facing page or deck, a business plan submitted with a credit application, or materials prepared for a sale process; the module is agnostic to document type. What defines the assessment is not the format but the evaluator role on which the reviewing models are conditioned — an investment reviewer for a pitch narrative, a credit officer for a loan application, an acquirer's analyst for M&A materials — so the module measures reception by the audience the narrative was written for. Its outputs are reported separately from the brand assessment, which is what enables the brand/narrative decomposition exploited in Section 6.

The module's headline outputs are two indices on a 0–1 scale, each tracked with full run history so their evolution is observable across assessment dates. The *Idea Score* summarizes the reviewing models' aggregate assessment of the strength of the underlying opportunity — whether the problem is real, the approach credible, and the economics plausible for the conditioned role. The *Confidence Score* summarizes their conviction in that assessment. The two are deliberately separated because their divergence is itself diagnostic: a high Idea Score paired with a lower Confidence Score indicates polarized reception — reviewers see the opportunity but are not yet persuaded of this firm's claim to it — a configuration that a single blended metric would render invisible. [VERIFY — confirm the formal definitions of both indices as implemented.]

Beyond the two indices, the module codes four further output classes. *Objections* are the specific reservations reviewing models raise against the narrative, coded into recurring categories — in the dataset, the most frequent concern platform dependency, competitive defensibility, and unit economics. *Identified risks* are objections the models judge material; they feed the composite's risk-deduction component directly, which is why a firm's combined score can sit below its brand-scope score even when brand signals are strong. *Competitor preference* records which alternative, if any, a reviewing model selects when asked to choose — the investor-side analogue of the buyer panel's preference outcome. *Affective response mapping* locates the narrative's reception on an affect field — the emotional register (enthusiasm, skepticism, urgency, indifference) in which models frame their evaluation. [VERIFY — confirm the affect-field dimensions as implemented.]

Methodologically, the module inherits the discipline of Sections 3.2–3.3 — battery sizing above the noise floor and policy-bounded conversational elicitation — so its outputs are comparable across models and across time on the same terms as the brand assessment. Substantively, it measures the object most existing visibility tooling ignores: not how models describe a company, but how they evaluate its story when placed in the deciding role.

We note a dual reading of these outputs. For the measured firm, they constitute advance notice: the objections coded here are, in our operational experience, the objections human reviewers subsequently raise. For evaluator-side readers — the diligence, risk, credit, and investment teams to whom Section 7 is addressed — the same measures run prospectively over a target constitute a structured summary of how the AI layer their own analysts consult already perceives that target's narrative: a credit officer receiving a business plan, for instance, can pair the applicant's document with its measured digital footprint and see both the coded objections and the disagreements among models that a single-assistant workflow never surfaces.

4. Data

The data span two generations of the measurement pipeline, and we are explicit about which claims rest on which; Table 1 summarizes the differences. An earlier pipeline generation, operated between September 2025 and May 2026, assessed more than 5,000 companies and produced several million individual model responses. That generation differed from the current one in three respects that preclude quantitative pooling: it executed smaller batteries of 500–750 prompts per assessment — below the stability operating point established in Section 3.2 — it analyzed only a subset of the signals now parsed, and it lacked several of the metrics introduced with the current methodology, so its outputs are not comparable to the composite scale of Section 3. The corpus is retained in archival form and serves two roles in this paper: it informed the development of the current methodology, including the query families of Section 3.1 and the objection categories of Section 3.8, and it supplies qualitative pattern context. No quantitative claim below is computed on it.

Table 1. Dataset summary by pipeline generation.

	Generation 1 (archival)	**Generation 2 (current vintage)**
Period	Sep 2025 – May 2026	May 2026 – Jul 2026
Companies assessed	5,000+	139 (fully scored, deduplicated)
Model responses	>1,000,000	~420,000
Prompts per assessment	500–750	1,000–1,500
Signal coverage	subset of current signals	full (Sections 3.4–3.8)
Composite score (300–870)	not computed — predates scale	yes, per company and per model
Longitudinal series	—	38 runs, operator's domain
Role in this paper	methodology development; qualitative context	all quantitative claims

All quantitative claims are computed on the current-vintage population: 139 companies holding a complete composite assessment under the methodology of Section 3, taking the latest completed run per company, deduplicated by domain, as of July 2026. This population underlies the distributional statements of Section 3.5 and the divergence analysis of Section 5. Companies enter the dataset by self-selection — a founder or operator initiates an analysis — which we discuss as a limitation in Section 8: the sample over-represents early-stage technology firms attentive to their AI presence and under-represents large incumbents.

For the longitudinal case study we use the platform operator's own domain, for which 38 complete assessment runs exist under the current vintage The series begins on January 29, 2026; its earliest runs coincide with the final development period of the current pipeline, so we read the early trajectory directionally rather than point-for-point, and rest all quantitative claims on measurements from the stabilized period. Using the operator's own firm as the case subject is disclosed as a potential conflict; it is offset by a unique advantage: it is the only company for which every intervening remediation action — directory listings, content changes, review acquisition — is fully observable and timestamped, permitting event-style attribution unavailable for third-party subjects.

The methodology is versioned. A third pipeline generation is in development; when deployed, its outputs will be reported as a distinct scoring vintage under the same discipline, not pooled with the population analyzed here.

5. Findings: The Structure of Split-Perception

5.1 Divergence is large and common. Per-model composite scores are persisted for the subset of the population assessed after per-model persistence was introduced: eleven companies, excluding the platform's demonstration entity. Within this subset, the median cross-model score range on a single measurement date is 100 points and the 90th percentile is 134 points — on a scale whose full reporting bands are 200 points wide. Put plainly: for a typical fully instrumented company, the most and least favorable model verdicts differ by roughly half a reporting band, and for one company in ten, by two-thirds of a band or more. The subset is small, and we report these figures as the first instrumented estimate rather than a settled population statistic; per-model persistence now covers all new assessments, so the estimate will tighten with the population.

Mention-level data, which cover more companies than per-model composites, supply a complementary and starker signal. Among 47 companies with category-query coverage from at least two models, 6.4% are simultaneously omitted entirely by at least one model and ranked in the top three by another. For these firms the split is not a matter of degree: the same company is a category leader in one AI's answer and absent from another's.

5.2 An illustrative case. In mid-July 2026, per-model composite scores for the case firm of Section 6 ranged from 424 (sonar, the retrieval-grounded engine behind Perplexity) to 539 (DeepSeek) — a 115-point split, with the remaining models clustered at 501–519. Read on its own, the pattern suggests a stable interpretive gap: the outlier is the model most dependent on live web citations, and the firm's citation-authority profile was its weakest measured component — the citation gap made visible through the model that weights citations most heavily. The longitudinal evidence of Section 5.4 refines this reading in an important way: the outlier's verdict is not only the most divergent but also the least stable, which changes the practical meaning of the split without diminishing it.

5.3 Divergence localizes in authority components. Component-level scores are not currently persisted per model, so we report this as a qualitative pattern rather than an agreement statistic; quantifying it is left to future work, pending per-model component persistence. Across assessments, models agree far more readily on visibility signals — whether a company is mentioned, and with what sentiment — than on positioning signals: category assignment, leadership attribution, defensibility. The case firm illustrates the pattern: models concur that it exists and is positively received (90% mention rate, 90% positive sentiment) while producing the composite splits documented above. The practical reading: models broadly agree on whether a company exists and is liked, and disagree on what it is and where it stands — precisely the dimensions a diligence process cares about.

5.4 Stability versus noise: two regimes. Repeated runs on the case firm — fifteen assessments over roughly three and a half weeks under the stabilized vintage — separate the model set into two regimes. Four of the five models are individually stable: their within-model score ranges across runs span 44–79 points, with standard deviations between 14.7 and 24.0 — materially below the 100-point median between-model range of Section 5.1. For these models, split-perception behaves as a stable property of the model–evidence pair: the same model keeps reaching approximately the same verdict, and the verdicts differ between models.A third pattern — unstructured within-model noise with no discernible attractors — was observed in the excluded candidate judge (Section 3.1) and motivates the reliability criterion for judge inclusion: divergence is measurable only between judges that are individually consistent.

The fifth model — sonar — exhibits a different regime entirely: a within-model range of 232 points and a standard deviation of 76.3 across the same fifteen runs, including runs scored at the scale floor of 300 and, on multiple dates, materially different scores from separate runs within the same day. We read this as retrieval variance imported into judgment. A model that re-fetches live evidence on every run inherits the volatility of whatever it happens to retrieve; a firm with a thin citation profile offers that retrieval little to stabilize on. Under this reading — which we state as a hypothesis, testable against firms with dense citation profiles — the volatility is not a defect of the model but a measurement of the firm: the thinner the independently citable evidence base, the noisier the retrieval-grounded verdict.

The two regimes carry distinct risks for diligence workflows. Stable splits mean the verdict on a company depends on which model is consulted. Retrieval volatility means it also depends on when. An evaluator relying on a single retrieval-grounded query observes one draw from this distribution.

6. Longitudinal Case Study: One Firm, 38 Runs

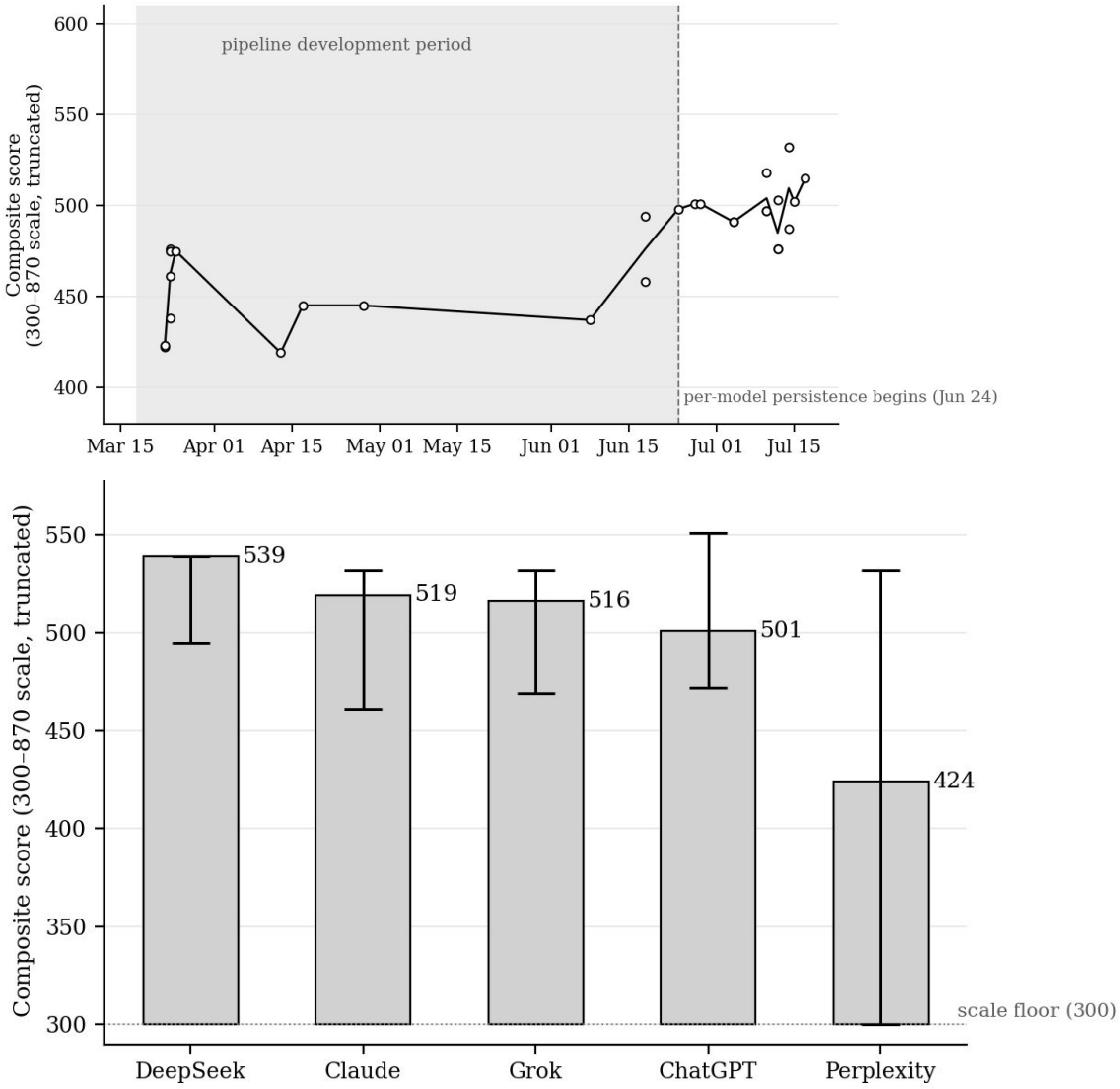
The case subject is the platform operator's own domain, for which 38 assessment runs exist between January 29 and July 2026. The series begins during the final development period of the current pipeline, so we read the early trajectory directionally rather than point-for-point and rest quantitative claims on the stabilized period; per-model scores are persisted for the fifteen runs between June 24 and July 17. Over the series, the firm's brand-scope composite rose from the 410–430 band to approximately 540 — from below the population median of 437 to near the top of the measured distribution, whose maximum observed score is 564 and whose 500-point threshold is reached by 12.9% of companies (Section 3.5). The trajectory was not smooth, and its interior structure is more informative than its endpoints.

Interventions over the series were continuous rather than discrete and were not logged as dated events; Figure 1 accordingly shows the trajectory without intervention markers, and we describe the remediation process qualitatively. It operated as a closed loop. Each assessment's panel outputs (Section 3.7) surfaced the themes on which judges' reservations concentrated; targeted content — documentation and landing pages — was published against those themes; and that content was then introduced into third-party discussion threads on professional and community platforms, where the firm's positioning was actively and sometimes contentiously debated. Operationally, judge outputs shifted within roughly three to seven days of a publication.

Two channel-level observations from this process are offered as hypotheses rather than findings, since neither is instrumented in the present data. First, content that generated third-party discussion moved assessments where equivalent static self-published material — the same claims without an ensuing thread — did not. Second, the panel-level indices of Section 3.8 responded first, with other components following. Both are consistent with the corroboration logic running through Section 5: judges appear to weight discussed, third-party-embedded content over self-description. We flag both for controlled study.

Two features of the series bear on the split-perception thesis. The first is that the composite's rise decomposes cleanly by regime. Within the persisted window, the four individually stable judges moved broadly together — each scored the firm higher at the end of the window than at its start — while the retrieval-grounded judge oscillated across its full observed range (300–532) with no monotone trend. The composite's improvement is therefore carried by the stable judges; the volatile judge contributes variance rather than direction. This is the two-regime structure of Section 5.4 observed longitudinally, and it replaces a simpler reading we considered and rejected against the data — that the retrieval-dependent judge's score was persistently depressed and recovered as authority-building interventions took effect. The run-level data show oscillation, not staged recovery.

The second feature is that assessment scope separates. On the same measurement date, the firm's brand-scope composite (≈540) and its combined score including the narrative document assessed by the module of Section 3.8 (515) differed by 25 points, the gap attributable to risk deductions and panel objections raised against the narrative rather than the brand [VERIFY — confirm the deduction decomposition against the run record]. Models, that is, assess a company and its story separably: the same evidence base supports a stronger verdict on the firm than on the account the firm gives of itself. For the measured population at large this decomposition is testable wherever both assessment scopes exist for the same firm on the same date.



Three qualifications bound what the case can show. First, the subject is operated by the author; the conflict is disclosed in Section 4, and all case figures are reproducible from the platform's public shared reports. Second, it is a single firm: the case demonstrates that the mechanisms of Sections 3 and 5 are observable in one fully instrumented subject, not that their magnitudes generalize. Third, interventions were continuous, overlapping, and not logged as dated events, so attribution is limited to the qualitative loop described above — descriptive, not causal. The case's evidential role is accordingly narrow: it establishes that composite movement, regime structure, and scope separation are all observable within a single firm's measured history.

7. Implications for Risk and Diligence Workflows

The findings of Sections 5 and 6 bear on three workflows in which AI-assisted research now routinely enters decisions about companies. We take each in turn, describing where model-mediated perception enters the workflow, what the measured structure of divergence implies for it, and what procedural mitigations follow. The mitigations are deliberately manual: they require nothing beyond access to several models and a record of what was asked.

7.1 M&A and investment screening. Model-mediated perception enters this workflow at two points. At sourcing, category-shaped queries — "leading companies in X," "best platforms for Y" — produce the short lists from which pipelines are built, and these lists are winner-take-few: an omitted company is not ranked low, it is absent, and its exclusion generates no signal to anyone. In our mention-level data, 6.4% of multi-model-covered companies are simultaneously omitted entirely by one model and ranked top-three by another (Section 5.1); for these firms, pipeline membership is decided by which assistant the sourcing analyst happens to use. At screening, diligence-shaped queries — "risks and weaknesses of Z" — function as a draft of the pre-call memo, produced before anyone at the firm has spoken to the target.

The measured structure of divergence quantifies what this costs. The median 100-point cross-model split of Section 5.1 spans half a reporting band: the same target reads as a different company depending on tooling choice, and the memo that results records conclusions, not the tooling that shaped them — the variation is invisible ex post. The two-regime structure of Section 5.4 adds a temporal dimension: against a thinly cited target, single queries to a retrieval-grounded assistant were observed to vary by more than 200 points across runs, so the outcome depends not only on which model was consulted but on when.

The case study adds a third entry point specific to evaluation: scope. Models assess a company and the account the company gives of itself separably (Section 6) — in the case firm, brand-scope and narrative-inclusive assessments of the same date differed by 25 points, the gap carried by objections against the narrative. Which of the two objects an associate's research actually consulted is determined by query shape, and is likewise unrecorded.

Procedural mitigations follow directly. Query several models and record the spread; treat a wide split not as an inconvenience but as a finding — it is diagnostic of a thin or contested evidence base, which is itself due-diligence information. Where the assistant is retrieval-grounded, repeat the query across several days before treating its answer as a reading. And record model and version in the deal file, so that the tooling that shaped attention is at least reconstructible.

7.2 KYB / KYC and counterparty review. In counterparty workflows, AI assistants most commonly enter as compression: onboarding checks, periodic reviews, and adverse-media screening all require a reviewer to reduce a large open-source footprint to a risk-relevant summary, and an assistant does in minutes what previously took hours. In most institutions this use is an informal accelerant — individually adopted, procedurally invisible, and absent from the documentation that describes how a review was performed.

Split-perception makes the informality a consistency problem. Two reviewers following identical written procedures with different assistants can reach different risk readings of the same entity — the median cross-model split in our data is 100 points — and, where the assistant is retrieval-grounded, the same reviewer can reach different readings of the same entity on different days. For regulated workflows, consistency of procedure is not cosmetic: it is what examinations test. A process whose outcome varies with an undocumented tooling choice fails that standard silently. The omission case is the sharpest instance: an entity absent from one model's answer and prominent in another's (6.4% of multi-model-covered companies in Section 5.1) is precisely the profile that assistant-compressed adverse-media screening can miss without leaving a trace.

The remedy is one institutions already know how to apply. Banks inventory, document, version, and validate the models that touch decisions; AI-assisted research currently functions as a model input while sitting outside that inventory. Bringing it inside means nothing exotic: queries documented, model and version recorded, and the research step named in the procedure it accelerates. The asymmetry between effort and exposure runs in the institution's favor — the documentation burden is minutes per review; the inconsistency it forecloses is currently unbounded and unmeasured.

Divergence itself, once measured, becomes an auditable control rather than a hidden defect. Because cross-model splits are quantifiable, they are documentable, and a simple escalation rule — flag the review when dispersion across consulted models exceeds a set threshold — converts an invisible inconsistency into a recorded event with a defined response. The reviewer's judgment stays where it is; what changes is that the AI layer beneath it acquires the same paper trail as every other input a regulated decision rests on.

7.3 Credit assessment. Perception scores are not credit scores and measure a different object: the state of a firm's AI-mediated reputation, not its capacity to service obligations. Nothing in this section proposes perception measures as inputs to capacity estimation. The relevant intersection is narrower — the soft-information channel, where a lender's qualitative read of a borrower's market standing, management credibility, and competitive position enters the file alongside the financials.

That channel is now often AI-assisted, and the separability result of Section 6 describes its structure. A commercial application typically arrives as a narrative — a business plan — attached to an entity with an ambient public footprint, and models assess the two separably: the plan and the footprint are distinct measured objects that need not agree. A coherent plan resting on a thin or contested footprint is a detectable configuration; so is the reverse, a strong footprint undersold by a weak narrative. An officer's AI-assisted read returns whichever object the query shape elicited, without marking which one it was.

The volatility hypothesis of Section 5.4 locates the channel's weakest point. Retrieval-grounded assessments were least stable exactly where the citable evidence base was thinnest — and thin public footprints are characteristic of the young and small firms for which soft information carries the most underwriting weight. If the hypothesis holds, the soft-information channel is noisiest precisely for the borrowers who depend on it most: an inversion lenders have no current instrument to see, because the model-dependence of an AI-assisted qualitative read is not yet named as a data-quality dimension in credit files.

The mitigations mirror the preceding subsections and stop at the boundary drawn above. Name the input where it is used; record model, version, and date alongside the qualitative read; where the read matters to the decision, take it from more than one model and more than one day, and treat wide dispersion as a signal about the evidence base rather than about the borrower. Perception measures inform the qualitative file; capacity is estimated as it always was.

For the measured companies themselves, the direction of remediation follows from where divergence lives. Section 5.3 localizes it in authority components, and the hypotheses of Sections 5.4 and 6 — that thin citation profiles yield volatile retrieval-grounded verdicts, and that third-party-discussed content moves assessments where static self-publication does not — point the same way: toward independently citable evidence such as third-party reviews, comparison content, and analyst or academic mentions. We state this as the direction most consistent with the present data rather than an established effect; the controlled intervention study flagged in Section 6 is its natural test.

7. Implications for Risk and Diligence Workflows

The findings of Sections 5 and 6 bear on three workflows in which AI-assisted research now routinely enters decisions about companies. We take each in turn, describing where model-mediated perception enters the workflow, what the measured structure of divergence implies for it, and what procedural mitigations follow. The mitigations are deliberately manual: they require nothing beyond access to several models and a record of what was asked.

7.1 M&A and investment screening. Model-mediated perception enters this workflow at two points. At sourcing, category-shaped queries — "leading companies in X," "best platforms for Y" — produce the short lists from which pipelines are built, and these lists are winner-take-few: an omitted company is not ranked low, it is absent, and its exclusion generates no signal to anyone. In our mention-level data, 6.4% of multi-model-covered companies are simultaneously omitted entirely by one model and ranked top-three by another (Section 5.1); for these firms, pipeline membership is decided by which assistant the sourcing analyst happens to use. At screening, diligence-shaped queries — "risks and weaknesses of Z" — function as a draft of the pre-call memo, produced before anyone at the firm has spoken to the target.

The measured structure of divergence quantifies what this costs. The median 100-point cross-model split of Section 5.1 spans half a reporting band: the same target reads as a different company depending on tooling choice, and the memo that results records conclusions, not the tooling that shaped them — the variation is invisible ex post. The two-regime structure of Section 5.4 adds a temporal dimension: against a thinly cited target, single queries to a retrieval-grounded assistant were observed to vary by more than 200 points across runs, so the outcome depends not only on which model was consulted but on when.

The case study adds a third entry point specific to evaluation: scope. Models assess a company and the account the company gives of itself separably (Section 6) — in the case firm, brand-scope and narrative-inclusive assessments of the same date differed by 25 points, the gap carried by objections against the narrative. Which of the two objects an associate's research actually consulted is determined by query shape, and is likewise unrecorded.

Procedural mitigations follow directly. Query several models and record the spread; treat a wide split not as an inconvenience but as a finding — it is diagnostic of a thin or contested evidence base, which is itself due-diligence information. Where the assistant is retrieval-grounded, repeat the query across several days before treating its answer as a reading. And record model and version in the deal file, so that the tooling that shaped attention is at least reconstructible.

7.2 KYB / KYC and counterparty review. In counterparty workflows, AI assistants most commonly enter as compression: onboarding checks, periodic reviews, and adverse-media screening all require a reviewer to reduce a large open-source footprint to a risk-relevant summary, and an assistant does in minutes what previously took hours. In most institutions this use is an informal accelerant — individually adopted, procedurally invisible, and absent from the documentation that describes how a review was performed.

Split-perception makes the informality a consistency problem. Two reviewers following identical written procedures with different assistants can reach different risk readings of the same entity — the median cross-model split in our data is 100 points — and, where the assistant is retrieval-grounded, the same reviewer can reach different readings of the same entity on different days. For regulated workflows, consistency of procedure is not cosmetic: it is what examinations test. A process whose outcome varies with an undocumented tooling choice fails that standard silently. The omission case is the sharpest instance: an entity absent from one model's answer and prominent in another's (6.4% of multi-model-covered companies in Section 5.1) is precisely the profile that assistant-compressed adverse-media screening can miss without leaving a trace.

The remedy is one institutions already know how to apply. Banks inventory, document, version, and validate the models that touch decisions; AI-assisted research currently functions as a model input while sitting outside that inventory. Bringing it inside means nothing exotic: queries documented, model and version recorded, and the research step named in the procedure it accelerates. The asymmetry between effort and exposure runs in the institution's favor — the documentation burden is minutes per review; the inconsistency it forecloses is currently unbounded and unmeasured.

Divergence itself, once measured, becomes an auditable control rather than a hidden defect. Because cross-model splits are quantifiable, they are documentable, and a simple escalation rule — flag the review when dispersion across consulted models exceeds a set threshold — converts an invisible inconsistency into a recorded event with a defined response. The reviewer's judgment stays where it is; what changes is that the AI layer beneath it acquires the same paper trail as every other input a regulated decision rests on.

7.3 Credit assessment. Perception scores are not credit scores and measure a different object: the state of a firm's AI-mediated reputation, not its capacity to service obligations. Nothing in this section proposes perception measures as inputs to capacity estimation. The relevant intersection is narrower — the soft-information channel, where a lender's qualitative read of a borrower's market standing, management credibility, and competitive position enters the file alongside the financials.

That channel is now often AI-assisted, and the separability result of Section 6 describes its structure. A commercial application typically arrives as a narrative — a business plan — attached to an entity with an ambient public footprint, and models assess the two separably: the plan and the footprint are distinct measured objects that need not agree. A coherent plan resting on a thin or contested footprint is a detectable configuration; so is the reverse, a strong footprint undersold by a weak narrative. An officer's AI-assisted read returns whichever object the query shape elicited, without marking which one it was.

The volatility hypothesis of Section 5.4 locates the channel's weakest point. Retrieval-grounded assessments were least stable exactly where the citable evidence base was thinnest — and thin public footprints are characteristic of the young and small firms for which soft information carries the most underwriting weight. If the hypothesis holds, the soft-information channel is noisiest precisely for the borrowers who depend on it most: an inversion lenders have no current instrument to see, because the model-dependence of an AI-assisted qualitative read is not yet named as a data-quality dimension in credit files.

The mitigations mirror the preceding subsections and stop at the boundary drawn above. Name the input where it is used; record model, version, and date alongside the qualitative read; where the read matters to the decision, take it from more than one model and more than one day, and treat wide dispersion as a signal about the evidence base rather than about the borrower. Perception measures inform the qualitative file; capacity is estimated as it always was.

For the measured companies themselves, the direction of remediation follows from where divergence lives. Section 5.3 localizes it in authority components, and the hypotheses of Sections 5.4 and 6 — that thin citation profiles yield volatile retrieval-grounded verdicts, and that third-party-discussed content moves assessments where static self-publication does not — point the same way: toward independently citable evidence such as third-party reviews, comparison content, and analyst or academic mentions. We state this as the direction most consistent with the present data rather than an established effect; the controlled intervention study flagged in Section 6 is its natural test.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), 5–16. https://doi.org/10.1145/3637528.3671900

Google. (2025). Search Quality Evaluator Guidelines. Google LLC. https://guidelines.raterhub.com/searchqualityevaluatorguidelines.pdf

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58(13), 2078–2091.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. Journal of Computer-Mediated Communication, 12(3), 801–823.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2017). Credit Scoring and Its Applications (2nd ed.). Society for Industrial and Applied Mathematics.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track).