FRANCESCO SAVERIO CANEPA · WORKING PAPER
METHODS & PROTOCOL PAPER · PUBLIC SEALED PILOT (N = 10)

Calibrated Dissent: A Verifiable Sealing-and-Resolution Protocol for Auditing Adversarial AI Forecasts, with a Public M&A Pilot (N = 10)

A verifiable, anti-hindsight protocol for auditing a structurally adversarial AI — instantiated on a public, pre-outcome-sealed pilot of high-stakes European M&A forecasts. The sealing and anchoring steps are demonstrated; resolution and scoring await the outcome horizons.

by

Francesco Saverio Canepa

METHODS & PROTOCOL PAPER — A VERIFIABLE ANTI-HINDSIGHT PROTOCOL WITH A PRE-SPECIFIED ANALYSIS PLAN, INSTANTIATED ON A PUBLIC SEALED PILOT (N = 10); NO RESULTS · PRE-REGISTRATION DRAFT, 25 JUNE 2026

CONTENTS

Contents

METHODS & PROTOCOL PAPER — PROTOCOL + SEALED PILOT; NO RESULTS

Abstract

AI systems optimized for helpfulness systematically fail consequential decisions: they mine evidence for support, return articulate validation, and deliver precisely the reassurance that high-stakes reasoning least needs. The *cognitive contradictor* is the alternative — a structurally adversarial AI whose mandate is to challenge rather than confirm. Whether such a system produces *calibrated dissent* — disagreement with consensus that is accurate rather than merely louder — is an empirical question the field has no honest way to answer, because grading a forecaster after outcomes are known invites hindsight. **This paper contributes a verifiable instrument: a verifiable, anti-hindsight protocol for auditing adversarial AI forecasts, and a public, pre-outcome-sealed pilot on real, consequential decisions — the sealing and anchoring steps are demonstrated; resolution and scoring await the outcome horizons.**

The protocol seals each forecast in an append-only, hash-anchored public ledger (the Osservatorio) before any outcome is observable; resolves it by a fully deterministic rule over canonical public sources; benchmarks it against a base-rate null matched to the contested cohort and a public-signal null; publishes a fully auditable resolution evidence trail (and, where feasible, double-codes with an independent blind coder); and freezes the exact model–prompt–Pack bundle by SHA-256. We instantiate it on a **fixed, publicly pre-sealed pilot of N = 10 high-stakes European M&A forecasts** (sealed at T0 = 20 June 2026), each a three-way scenario distribution from which P(H1) (deal-break or downward renegotiation ≥ 15%, 6–12 months) and P(H2) (post-close value destruction, 18–36 months, on the subset with a listed consolidating acquirer) are derived. The scored quantities are these sealed **forecaster** probabilities — the dominant-reading judgement expressed through the contradictor process; the engine's full three-pass challenge record (the seven-field schema, the Δ-CSI, and the bundle hash) enters as **provenance only** and cannot alter them. The pilot therefore audits the sealed forecasts, **not** the engine's generation of them, and reports the raw forecast-versus-outcome register as outcomes resolve. At sealing, each public card recorded the three-way scenario distribution, the dominant-reading thesis, a calibrated-confidence level, and cited sources; the engine's full seven-field payload, the Δ-CSI, and the bundle hash **do not yet exist for the ten deals** and will be generated afterward as non-scoring provenance (§4.3). This pilot is therefore a demonstration of the sealing-and-resolution rig, not an evaluation of the engine's own generation. The contradictor engine itself operates through a three-pass, model-isolated pipeline (Thesis, Red Team, Synthesis)

with a fixed seven-field output schema and a fail-closed architecture designed to prevent sycophantic drift.

Calibration (H-cal) and divergence-that-pays (H-pay), with the two null baselines and pre-declared falsification criteria, constitute a **pre-specified analysis plan for a future powered stage** on a larger, forward-sealed cohort; at the pilot's fixed N = 10 — below the power floor of N ≥ 20 per horizon — no confirmatory test is executed, and every quantity is reported descriptively. Three contributions follow: a conceptual framework distinguishing calibrated dissent from sycophantic validation; the verifiable sealing-and-resolution protocol itself; and an open implementation with a public, immutable pilot enabling independent replication. The paper does **not** claim that the contradictor improves decisions, that the engine's own forecasts are calibrated (a future-stage claim), that Δ-CSI predicts correctness, or that findings generalize beyond M&A. The Δ-CSI (Cognitive Sovereignty Index delta) is a computed proxy of challenge intensity, not of correctness — an explicit design invariant. Lane C demonstrations are non-evidence for any pre-registered claim.

SECTION 1

Introduction

Every consequential decision rests on a model of the world. That model is rarely tested before it is acted upon.

The human tendency to seek confirmation rather than disconfirmation of favored beliefs — what Kahneman (2011) systematically documented — is compounded, rather than corrected, by AI assistants trained predominantly to produce helpful and agreeable outputs. When the task is to evaluate a merger or acquisition, an AI optimized for agreeableness will mine the analyst's memorandum for supporting evidence, return articulate prose that reinforces the thesis, and deliver precisely the validation that a team under pressure least needs. Sycophancy is not a personality flaw; it is a design outcome. High-stakes decisions require the opposite.

The alternative concept investigated here is the *cognitive contradictor*: an AI system whose mandate is not to answer but to challenge, not to agree but to stress-test a decision at the moment it is being formed. The contradictor's job is to surface implicit assumptions, construct counter-intuitive scenarios, and return calibrated uncertainty — all before commitment. This is not a critique generation tool bolted onto a report; it is an adversarial agent inserted into the deliberative process at the decision point itself, with its output recorded, time-stamped, and compared against eventual outcomes.

The intellectual lineage of structured adversarial challenge is well established. Schwenk (1990) reviewed three decades of experimental and field evidence showing that devil's advocacy and dialectical inquiry — prescribed methods for forcing explicit

counter-argument — improved plan quality and reduced group-think in strategic decisions. Yet that literature predates large language models entirely. The question of whether an LLM can operationalize devil's advocacy at scale, reliably and consistently, remained open until recently. A 2024 preprint on "antagonistic AI" formalized the design space for AI agents whose utility function is explicitly oppositional (Antagonistic AI, arXiv:2402.07350; Cai, Arawjo, and Glassman, 2024). Two 2025 preprints interrogated whether such systems actually reduce overconfidence or merely produce a superficially skeptical register without affecting belief updating (skeptic2025a, arXiv:2509.05396; Wynn, Satija, and Hadfield, 2025; skeptic2025b, arXiv:2511.07784; Wu, Li, and Li, 2025). Across this emerging literature, one design gap persists: no study has subjected an AI contradictor to a pre-registered, calibration-scored field test on real, consequential decisions whose outcomes can be verified. Controlled laboratory studies use hypothetical stimuli; observational analyses lack a pre-registered baseline against which divergence can be measured; and no protocol exists for quantifying *calibrated dissent* — the degree to which a contradictor's disagreement with consensus is accurate rather than merely louder. Forecasting tournaments and prediction markets — Good Judgment Open (Tetlock and Gardner, 2015), Metaculus (Metaculus, Inc., 2015) — already seal and score probabilistic forecasts, but none audits a *structurally adversarial* AI's dissent from a domain consensus against a matched base-rate null; that specific instrument is what this protocol builds.

> *No protocol exists for quantifying calibrated dissent — the degree to which a contradictor's disagreement with consensus is accurate rather than merely louder.*

This paper addresses that gap not by claiming an answer but by building the instrument that would let one be earned honestly. We contribute a **verifiable sealing-and-resolution protocol** for auditing adversarial AI forecasts — pre-outcome sealing with a public hash anchor, a fully deterministic public-source resolution rule, a base-rate null matched to the cohort, a public auditable resolution evidence trail (with blind double-coding where feasible), and a per-call bundle-hash version freeze — and we instantiate it on a fixed, publicly pre-sealed **pilot** of N = 10 sealed dominant-reading forecasts on high-stakes mergers and acquisitions. The broader claim the research programme will eventually test is that the contradictor produces calibrated forecasts that diverge from consensus and beat a pre-registered baseline; **this paper does not adjudicate that claim and does not test the engine's generation**. It delivers the protocol and the public pilot, and reports the pilot's raw forecast-versus-outcome

register descriptively as outcomes resolve, reserving the powered confirmatory test for a future stage on a larger, forward-sealed cohort (§6.5). We choose M&A because deal verdicts are binary, verifiable, and carry stakes high enough to exclude performative disagreement. We measure calibration — not accuracy in isolation — because a challenger that is merely contrarian is valueless; one whose uncertainty estimates track empirical frequencies is informative.

This paper makes three contributions. First, a conceptual framework that distinguishes calibrated dissent from sycophantic validation and from unstructured skepticism, and that formalizes the contradictor's output schema (implicit assumptions, counter-intuitive scenario, falsification tests, burden-of-proof questions, calibrated confidence, cited sources, and a Δ-CSI — Cognitive Sovereignty Index delta — as a proxy of challenge intensity for each session). Second, and centrally, the **verifiable sealing-and-resolution protocol** itself — the anti-hindsight rig that makes an honest later test possible — together with the pre-specified analysis plan (resolution rule, matched null baselines, and falsification criteria) that a powered stage would apply. Third, an open implementation and a public, immutable pilot that enables independent replication and already demonstrates the **sealing-and-anchoring front** of the rig on real decisions; its resolution-and-scoring half awaits the outcome windows (6–36 months) and is exercised by the future powered stage. We make explicit what this paper does not claim: we do not claim that the contradictor improves decisions, that the engine's own forecasts are calibrated, that Δ-CSI predicts correctness, or that the results will generalize beyond the M&A domain. Those are claims for a future powered stage, contingent on its results; this paper is the protocol and the pilot, not the verdict.

SECTION 2

Theoretical framework

2.1 — THE CONTRADICTOR, NOT THE ORACLE

The central design choice of the system under study is architectural, not parametric: the engine's mandate is to challenge a decision, never to validate it. We call this the *contradictor invariant*. Where an oracle-style AI presents a recommended answer, a cognitive contradictor surfaces what the decision-maker has not considered, what could falsify the thesis, and what epistemic obligations remain unmet. Compliancy is treated not as a degraded mode but as a failure mode — precisely analogous to a type-II error in statistical testing. A system that returns a reassuring analysis without genuine challenge is defective, regardless of how articulate the output appears.

Oracle vs. cognitive contradictor

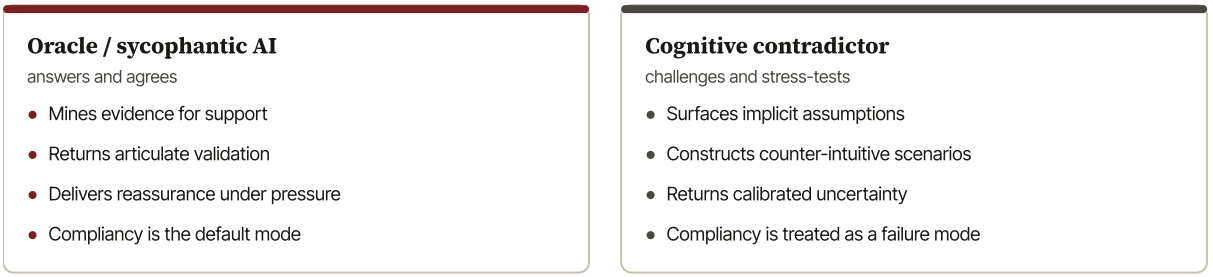

Figure 1 — Two design stances. The oracle / sycophantic AI returns validation; the cognitive contradictor returns structured challenge. Conceptual, drawn from §1 and §2.1; no data values.

This distinction has methodological consequences. Devil's advocacy, as reviewed by Schwenk (1990), is most effective when the counter-argument is institutionally *required*, not merely permitted. The contradictor enforces that requirement structurally: the output schema is fixed and non-negotiable, and the pipeline is fail-closed — it raises an explicit error rather than silently emit an unchallenging response. No configuration option softens the mandatory challenge. Domain specialization is expressed through external configuration files (BRAIN Packs) that can alter the vocabulary, persona, and emphasis of the red-team agent, but cannot remove or rename any element of the output contract.

2.2 — THE FIXED OUTPUT SCHEMA

Every analysis produced by the contradictor must conform to a versioned JSON schema (`schema_version: "1.0"`, `additionalProperties: false`). The schema defines seven substantive output fields, each mandatory:

1. **Implicit assumptions** (`assumptions[]`): the premises embedded in the decision that the decision-maker has not stated explicitly. Each assumption carries a severity rating on a 1–5 scale and a category label.

2. **Counter-intuitive scenario** (`counter_scenario`): a narrative that describes a plausible outcome the dominant reading does not anticipate — the scenario that, if it materialized, would most damage the decision's value.

3. **Falsification tests** (`falsification_tests[]`): operationalized empirical checks that, if the results came out in the specified direction, would undermine the thesis. Each test identifies the data required to run it and tracks its current status.

4. **Burden-of-proof questions** (`burden_questions[]`): explicit questions about who bears the epistemic obligation to demonstrate a key assumption, and whether that obligation has been met.

5. **Calibrated confidence** (`confidence`): a three-level confidence rating (low / medium / high) accompanied by a rationale and a list of declared knowledge gaps — what the system does not know and cannot impute.

6. **Cited sources** (`sources[]`): every source referenced in the analysis, each tagged with a provenance type drawn from a closed enum (`web_search | internal_ledger | internal_doc | osint_module | user_provided`). If no retrieval is available, the field is required to be empty; the system is prohibited from fabricating citations.

7. **Δ-CSI** (`csi_delta`): the Cognitive Sovereignty Index delta, defined in §2.3 below.

This schema is an invariant: no vertical specialization or client configuration may relax, remove, or rename any of these fields. The contract is defined once and versioned; a breaking change requires an explicit version bump and a new calibration cohort.

2.3 — THE Δ-CSI: A PROXY FOR CHALLENGE INTENSITY

The Δ-CSI (Cognitive Sovereignty Index, delta) is a computed, persisted, and calibrated index in the range [0, 100]. It measures how forcefully the contradictor process has stressed the decision — that is, the intensity of the challenge applied in a given session.

> DESIGN INVARIANT
>
> **The Δ-CSI is not a measure of decision quality, and it does not predict whether the decision is correct.** This framing constraint is an explicit design invariant, stated prominently in the system's methodology and in this protocol, because the name and 0–100 scale invite misreading. A high Δ-CSI means the contradictor pressed hard on many dimensions with high-severity assumptions; it says nothing about whether the challenged decision will succeed or fail. Treating Δ-CSI as an accuracy signal is a

> category error; we pre-register this clarification to pre-empt powered-stage misinterpretation.

Operationally, the Δ-CSI (`csi-v1`) is computed deterministically from five weighted components: (i) implicit assumptions count, modulated by their mean severity; (ii) falsification test count; (iii) burden-of-proof question count; (iv) cited source count; and (v) a divergence score between the thesis and the counter-intuitive scenario. Components (i)–(iv) use a capped count normalized to [0, 1] with a ceiling of five items (`min(n, 5)/5`), preventing gaming through verbosity. The divergence component (v) is computed either as cosine distance between semantic embedding vectors — normalized to [0, 1] as (1 − cosine similarity)/2 (when an embedding provider is available) — or as Jaccard complement over normalized tokens (heuristic fallback). The divergence method used in each session is recorded in the output alongside the score. Component weights sum to 100; vertical configurations may adjust weights, but the sum constraint is enforced programmatically.

The Δ-CSI is persisted in the Sovereign Decision Ledger for every session. Calibration — the empirical comparison of Δ-CSI distributions against eventual decision outcomes — becomes possible only after at least five sessions with recorded outcomes have accumulated. Below that threshold, the system explicitly flags its statistics as descriptive only, with no calibration claim. This prevents premature validity assertions. This engine-level gate (≥ 5 sessions) is an operational product safeguard, **independent of and far weaker than** the study's confirmatory power floor (N ≥ 20 per scored horizon; §6.5); clearing it authorises no calibration claim in the sense of §6, and the present fixed cohort sits below the study floor by design.

2.4 — THE THREE-PASS PIPELINE AND FAIL-CLOSED DEFENSES

The contradictor operates through three sequentially isolated passes, each executed by an LLM agent with a distinct, non-overlapping mandate.

Figure 2 — The three-pass, model-isolated pipeline. Labels and sub-labels drawn from §2.4; conceptual process, no data values.

Pass 1 — Thesis. The first agent produces the *dominant reading*: the conclusion a careful analyst would draw from the available evidence. This agent is explicitly prohibited from acting as a contradictor; its role is to articulate the best-case thesis as a target for the subsequent challenge. The thesis is passed, unmodified, to Pass 2.

Pass 2 — Red Team. The second agent receives the thesis as the object to be attacked. Its system prompt is constructed from the resolved Pack, which supplies a red-team persona directive, a domain-specific bias taxonomy with prescribed counter-moves, falsification templates calibrated to the decision domain, and operative guardrails. The red team's output is structured JSON covering all mandatory challenge fields (assumptions, counter scenario, falsification tests, burden questions). The passes are *model-isolated* by default: thesis and red team run on different LLM providers to decorrelate systematic blind-spots. Each pass's provider is recorded in the output metadata.

Pass 3 — Synthesis. The third agent receives the decision block, the thesis, and the red-team challenge, and is tasked only with calibrating confidence and declaring knowledge gaps. It does not conclude, recommend, or adjudicate between the thesis and the counter-scenario; it explicitly declares what is unknown.

Three structural defenses guard against the fail-open modes that would make the contradictor useless:

- ☐ **Contradiction Floor.** After Pass 2, the pipeline checks that the challenge is non-empty across all four dimensions (at least one assumption, one falsification test, one burden question, and a non-empty counter-scenario narrative). If the check fails, the pipeline issues one retry with an explicit directive to produce a non-empty challenge. If the check fails again after the retry, the pipeline raises an engine error and halts. No partial or empty output is emitted.

- ☐ **Closed-enum provenance.** Sources must declare a provenance type from the closed enum. If no retrieval tool is available, the synthesis agent is constrained to return an empty source list. The closed enum makes *mis-tagged* provenance impossible and forbids citations under empty retrieval, removing the most common fabrication path; it does not by itself guarantee that a model never emits a non-existent source under an otherwise-valid tag, so source existence remains a resolution-time verification.

- ☐ **Sub-threshold calibration gate.** Calibration statistics are computed only when the sample contains at least five sessions with recorded outcomes. Below that threshold, the interpretation field is forced to a cautious-only descriptor; no empirical validity is claimed. (This product-level gate is distinct from, and far weaker than, the study's power floor of §6.5.)

Together, these invariants — fixed schema, mandatory challenge floor, closed provenance, sub-threshold gating, and model isolation — define a system whose output is, by design, resistant to sycophantic drift. Whether that structural resistance

translates into calibrated forecasting accuracy is the empirical question this protocol is designed to make answerable, by a future powered stage, without hindsight (§6.5).

SECTION 3

Pre-specified analysis plan (deferred to a powered stage)

*The protocol pre-specifies five quantities that a **powered stage** will test and that the **pilot** reports descriptively.*

H1 and H2 are the contradictor's forecast signal at two horizons; H-cal, H-div, and H-pay concern calibration and the value of dissent. H-cal and H-pay are the **planned-confirmatory** targets of the powered stage; H-div is descriptive. None is a hypothesis *of this paper*: at the pilot's fixed N = 10 (§4.1) no formal decision rule is computed (§6.5), and every quantity is reported only as a raw forecast-versus-outcome finding. The decision rules stated below define what the powered stage will do, not what the pilot concludes — they are pre-registered now so that a future stage cannot tune them after seeing outcomes.

LABEL	NAME & HORIZON	STATUS
H1	Primary forecast, 6–12-month horizon	Horizon quantity
H2	Secondary forecast, 18–36-month horizon (listed-acquirer subset)	Horizon quantity
H-cal	Calibration against realized outcomes	Planned-confirmatory (powered stage)
H-div	Divergence from analyst consensus at T0	Descriptive
H-pay	Divergence-that-pays vs. consensus baseline	Planned-confirmatory (powered stage)

> *At the pilot's fixed N = 10 these quantities are reported descriptively only; the confirmatory decision rules are reserved for a future powered stage (§6.5).*

H1 — PRIMARY FORECAST (6–12-MONTH HORIZON)

At the 6–12-month resolution window, the sealed T0 forecast (the dominant-reading judgement of §4.2) assigns a probability estimate P(deal-break or downward renegotiation ≥ 15%) that the powered stage tests for being measurably above the pre-registered chance level in the sub-sample of deals where it diverges from analyst consensus at T0. The threshold of ≥ 15% renegotiation is fixed operationally in the

Methods (§4) and the resolution rule (§5); it is expressed here as a realized-outcome criterion, not a probability cutoff. In the powered stage H1 is scored by the Brier score (Brier, 1950) against the pre-registered baseline (§6), with Murphy's (1973) reliability/resolution/uncertainty decomposition reported only above the power floor; on the pilot it is shown as raw forecast-versus-outcome rows (§6.5).

H2 — SECONDARY FORECAST (18–36-MONTH HORIZON)

At the 18–36-month resolution window, the contradictor's sealed probability estimate P(value destruction: goodwill impairment or acquirer TSR below its sector index by 10 p.p.) is audited against realized outcomes. **H2 is defined only for the pre-declared subset of deals with a listed acquirer that consolidates the target** (§4.2; the subset is fixed per row in Appendix D); for the remaining deals — private-equity acquirers, delistings, carve-outs, joint ventures, and multi-party break-ups — H2 is *not applicable by design*, not post-hoc censored. H2 is assessed by the same Brier protocol as H1 on that subset, with the longer resolution window constituting an independent horizon rather than a replication. Given the small eligible subset (four deals; Appendix D) and the long horizon, the pilot's resolved H2 sample is likely to fall below the §6.5 reporting threshold of three; H2 will then be shown only as raw rows, not as a summary.

H-CAL — CALIBRATION (PLANNED-CONFIRMATORY, POWERED STAGE)

Across all resolved calls, the contradictor's stated probability estimates will be well-calibrated against realized outcomes, assessed via a calibration curve whose formal test and significance level are specified in §6. H-cal is planned-confirmatory. Failure to meet the pre-registered calibration threshold would constitute a falsification of the calibration claim (see §7) — assessed formally only on a sample that reaches the power floor (§6.5); on this fixed N = 10 cohort H-cal is reported descriptively.

H-DIV — DIVERGENCE (DESCRIPTIVE)

A non-trivial fraction of contradictor calls at T0 will diverge from the contemporaneous analyst consensus (rating or price target). "Non-trivial" is operationalized in §6 as a pre-registered minimum divergence rate; the threshold is chosen to exclude noise disagreement while permitting genuine dissent. H-div is descriptive: it characterizes the distribution of calls in the study cohort and does not itself constitute a confirmatory test. Its function is to establish that there is sufficient divergence in the sample to permit meaningful evaluation of H-pay.

H-PAY — DIVERGENCE-THAT-PAYS (PLANNED-CONFIRMATORY, POWERED STAGE)

Among the sub-sample of calls identified under H-div as divergent from analyst consensus at T0, the contradictor's probability estimates will achieve a lower Brier score (better calibration) than the consensus baseline at the pre-registered threshold. H-pay is planned-confirmatory. It targets the specific epistemic claim that divergence from consensus is not mere noise but carries predictive signal beyond what the consensus already contains. A failure of H-pay — divergent calls achieving no better Brier score than the consensus — would constitute partial falsification of that claim (§7), but only on a sample that reaches the power floor (§6.5); on this fixed cohort H-pay is suppressed entirely if the divergent sub-sample is below the floor of §6.3 — which, at N = 10, it almost certainly is, so the pilot is not expected to report H-pay at all. H-pay is defined on the **H1 horizon only**: the four-deal H2 subset cannot reach the divergent-sub-sample floor.

> **SCOPE CONSTRAINTS**
>
> These hypotheses make no claim about decision improvement, managerial behavior change, or organizational outcomes. Δ-CSI is recorded for all sessions and will be used as a covariate in exploratory analysis (§6), but no hypothesis pre-registers a relationship between Δ-CSI and forecast accuracy; such a relationship, if observed, will be reported as exploratory and treated as powered-stage material. The hypotheses are restricted to the M&A domain as defined in §4; generalization to other decision domains is not claimed.

SECTION 4

Methods

4.1 — SELECTION UNIVERSE: A FIXED, SEALED COHORT

The study cohort is a **fixed, closed set of ten high-stakes M&A decisions** (N = 10), publicly sealed in the Osservatorio on 20 June 2026 — a single sealing date that precedes every H1 and H2 outcome in the cohort. The ten deals span European large-cap M&A across banking, telecommunications, defence, insurance, asset management, and industrials, with announced transaction values ranging from approximately €2.2bn to approximately €40bn; at least one party to each deal is listed on a European regulated market (Euronext Milan, Euronext Paris, Euronext Amsterdam, the London Stock Exchange, SIX Swiss Exchange, the Frankfurt exchange, or Nasdaq Stockholm). The full cohort, with each deal's sealed record hash and its scenario-to-hypothesis

mapping, is enumerated in **Appendix D** (`docs/paper/appendices/cohort-register.md`) and is independently inspectable on the public Osservatorio.

Unlike a generative inclusion rule keyed to a size threshold and a forward collection window, this cohort was **assembled editorially** — as a set of flagship, contemporaneously contested European transactions — and then **frozen**: once sealed, no deal may be added to or removed from the cohort. This is a deliberate design choice with an honest cost and an honest safeguard. The cost: editorial assembly is a selection judgment, disclosed as a conflict of interest in §8 (COI-3). The safeguard, which is what makes the selection judgment non-fatal, is structural and verifiable: every deal was sealed and hash-anchored in an append-only, git-committed ledger **before any of its H1 or H2 outcomes was observable**, and the set is frozen. Selection therefore cannot have been motivated by outcomes that did not yet exist, and no deal that resolves unfavourably can be quietly dropped after the fact — the hash chain and the public commit history forbid it. Cherry-picking by post-hoc addition or removal is structurally impossible; cherry-picking at sealing time is neutralised by the pre-outcome seal and bounded by the base-rate null (§6.1, null (a)), against which an eventful cohort earns no free credit. The anti-hindsight guarantee of this study rests on the public, timestamped, immutable seal (§4.3, §4.4, §8), not on a mechanical inclusion rule.

The fixed cohort size of N = 10 is **below the power floor** of N ≥ 20 per scored horizon adopted for confirmatory calibration testing (§6.5). This is known in advance and is not a sampling accident: the study is therefore designed and reported as a **pre-registered calibration audit at feasibility scale**, in which the calibration analysis is descriptive (§6.5) and the confirmatory decision rules of §6 are computed only if and when a resolved sample reaches the power floor. The descriptive framing is pre-declared here, not invoked post-hoc to rescue an underpowered result. A powered confirmatory test would require extending the cohort in a separate powered-stage pre-registration; that extension is out of scope for this protocol.

4.2 — PREDICTION UNIT

For each deal in the cohort, a single **forecast** is sealed at **T0 — the cohort sealing date (20 June 2026)** — covering two pre-registered outcome horizons simultaneously. The scored forecast is the contradictor's dominant-reading probability judgement (a forecaster quantity), authored through the contradictor process and sealed; the engine's full three-pass payload is reconstructed afterward as provenance (§4.3) and is never the scored quantity. T0 is the sealing date, not the instant of each deal's public announcement: the deals were announced between October 2025 and June 2026, but all H1/H2 outcomes resolve 6–36 months later, so T0 deliberately follows announcement. This is a disclosed deviation from announcement-instant T0, and the informational disadvantage it costs is **not uniform**: the announce-to-seal lag ranges from a few days

(e.g., EQT–Intertek, announced 18 June) to about eight months (e.g., the Airbus/Leonardo/Thales "Project Bromo" MoU, announced 23 October 2025), and the per-deal lag is recorded in Appendix D. The governing rule is therefore explicit rather than a blanket claim: **a deal is scored on a given horizon only if no *qualifying resolution signal* (per §5) for that horizon was already public at T0**; if one was, the deal is an observation, not a forecast, and is excluded from that horizon. In the present cohort no deal had a qualifying H1 or H2 signal at the seal date — all ten were pending — but several carried elevated pre-T0 public information, most notably UniCredit–Commerzbank, whose low first-offer acceptance and government opposition were already public by 20 June 2026. Such deals are not excluded (no qualifying signal had occurred) but are **flagged** as elevated-pre-T0-information in Appendix D, and the announce-to-seal lag is reported per deal so the informational disadvantage is visible rather than assumed uniform. This replaces the untenable blanket assertion that no outcome is observable at T0. No retrospective calls, no supplementary calls, and no call revisions are treated as the primary forecast record; the T0 sealed call is the sole scored unit.

Probability estimates from the scenario distribution. Each sealed call states a three-way, mutually exclusive scenario distribution summing to 100%: a *success* scenario (the deal completes and creates value), a *value-disappointment* scenario (the deal completes but destroys value), and a *deal-failure* scenario (the deal breaks or is materially reduced). The two pre-registered probability estimates are derived deterministically from this distribution: **P(H1) = the probability mass of the deal-failure scenario** and **P(H2) = the probability mass of the value-disappointment scenario**. The mapping is recorded per deal in the sealed record; where a deal's scenario labels do not map cleanly onto this three-way structure (e.g., a contested control bid whose failure mode is a blocked minority stake rather than a withdrawal), the deal-level mapping is fixed explicitly in Appendix D. This derivation rule is fixed in this protocol before resolution and is not adjusted thereafter. The scenario→hypothesis mapping is the most exposed locus of construct bias (§8): the scenario *distribution* is in the 20 June seal, but the mapping of scenarios onto H1/H2 — including the four exception deals in Appendix D — is authored. It is therefore pre-registered here, before any outcome, **and is open to independent re-checking — by any third party from the public evidence trail, and by a recommended independent coder (§5.1)** — against the written rule; what is re-checked is the mapping's application, while its construct origin remains the author's and is disclosed as a residual in §8.

These probability estimates are the **frozen, publicly sealed** values: the scored P(H1)/P(H2) are exactly those in the 20 June 2026 seal (Appendix D, §D.2), and the later full contradictor payload cannot alter them (§4.3; the lane-A-internal firewall, §9.3). The scenario distribution is a *forecaster* quantity carried in the Osservatorio record; it is

not a field of the contradictor output schema (§2.2), which emits categorical confidence rather than a probability distribution — the two layers are kept separate in §4.3.

Because the three scenarios are mutually exclusive, a deal that breaks (H1 = 1) realises H2 = 0 by construction. **P(H2) is therefore the marginal probability of the joint event "completes-and-destroys-value"** — the mass of the value-disappointment scenario — scored on the full H2 subset without separate completion-censoring. This is deliberately *not* conditioned on completion: conditioning the forecast on completion while scoring an outcome that is itself defined only on completed deals would be incoherent (it would put a conditional probability in the numerator and an unconditional event in the denominator). The deal-failure mass lands in H1 and never in H2, so it is not double-counted; the per-deal scored P(H2) is the raw value-disappointment mass recorded in Appendix D. In plain terms: P(H1) is the probability the deal *breaks*; P(H2) is the probability it *closes but disappoints*. They are distinct, non-overlapping failure modes, so nothing is double-counted.

H1 horizon (6–12 months). The sealed T0 forecast states an explicit probability estimate P(deal-break or downward renegotiation ≥ 15%). The 15% renegotiation threshold, introduced in §3, is operationalized here as a realized-outcome criterion: a deal is coded H1 = 1 if and only if a public announcement or regulatory filing confirms either deal termination or a revision of the consideration reducing the headline price by at least 15% relative to the original announced value. The measurement is taken at the first public signal within the 6–12-month window; if no signal is available within the window, the deal is censored (§4.5).

H2 horizon (18–36 months), listed-acquirer subset only. H2 presupposes a listed acquirer that consolidates the target and whose TSR is public. It is therefore scored **only on the pre-declared subset of deals satisfying that condition**, fixed per row in Appendix D before scoring (in the present cohort: Intesa–MPS, Poste–TIM, Zurich–Beazley, Ingredion–Tate & Lyle). For deals with a private-equity acquirer, a delisting, a carve-out to a fund, a joint venture, or a multi-party break-up, neither consolidated goodwill in respect of the target nor a public acquirer TSR exists; these are marked *H2 not applicable by design* in Appendix D and never enter the H2 pool. They are **not** counted as censored, because the unobservability is structural and pre-declared, not outcome-driven — distinguishing this exclusion from the right-censoring of §4.5. On the H2 subset, value destruction is operationalized as: goodwill impairment recorded in the acquirer's consolidated financial statements in respect of the acquired entity, OR the acquirer's total shareholder return (TSR) — computed in the acquirer's listing currency — falling below the matching sector index (in the same currency; the exact index is fixed per row in Appendix D) by at least 10 percentage points over the 18–36-month window following deal close.

4.3 — SEALED OUTPUT AT T0 (SEALED ENVELOPE)

At T0 each deal generates a sealed Osservatorio record that **wraps** the contradictor's schema payload with sealing metadata. The two layers are distinct, and only the sealed probability estimates (element 2) are scored:

1. **The contradictor schema payload** (`schema_version: "1.0"`): the seven mandatory output fields — implicit assumptions, counter-intuitive scenario, falsification tests, burden-of-proof questions, calibrated confidence, cited sources — plus the Δ-CSI value computed by the `csi-v1` scorer (§2.3). This payload is governed by `contradictor_output.schema.json` with `additionalProperties: false`; it does **not** contain P(H1)/P(H2) or the bundle hash, which are sealing metadata, not schema fields.

The remaining elements are **Osservatorio sealing metadata** that wrap the payload; they are not fields of the contradictor output schema:

2. The three-way scenario distribution and the two probability estimates P(H1) and P(H2) derived from it (§4.2). **These are the scored quantities.**

3. The contemporaneous analyst consensus at T0: the modal analyst rating (Buy / Neutral / Sell or equivalent) and the median 12-month price target, as reported by Refinitiv I/B/E/S at T0 (the sealing date).

4. A bundle hash — a SHA-256 digest of the concatenation of: the exact model identifier and version for each of the three pipeline passes (thesis, red team, synthesis), the resolved Pack (YAML, canonical form), and the full system-prompt string used in each of the three pipeline passes. The bundle hash makes the challenge payload reproducible and auditable: a third party can verify that it was produced by the declared model-prompt combination.

5. A Unix timestamp (UTC) recording the moment the record was finalized.

The implementing infrastructure is the **Osservatorio** — the sealed-envelope ledger already in production for the lane-A M&A prediction cohort operated under this study. The Osservatorio writes each record to an append-only store and exposes the record hash via its public API, making the pre-existence of each forecast verifiable against an independent timestamp anchor. The Osservatorio's record format is the canonical source of truth for §5 (resolution) and §10 (reproducibility).

> **IMPLEMENTATION STATUS (HONEST DISCLOSURE)**
>
> The Osservatorio infrastructure — sealing, SHA-256 hashing, the append-only git-committed store, and the deterministic news-resolution engine — is already in production, and the ten cohort deals are already publicly sealed (each carries a published content hash; Appendix D). At the time of this pre-registration the public cohort cards carry a *simplified* view of each forecast: context (the dominant-reading

> thesis), the three-way scenario distribution, a calibrated-confidence level, cited sources, and the seal hash. The **scored quantities are the probability estimates already sealed on 20 June 2026** (Appendix D, §D.2); they are frozen, and the steps below cannot change them.

The **full contradictor schema payload** — the seven mandatory fields, the Δ-CSI, and the bundle hash — does not yet exist for these ten deals; it will be generated by the three-pass engine as *non-scoring provenance* (it documents how the challenge was produced and cannot alter P(H1)/P(H2); §9.3). To close the garden-of-forking-paths that a later, author-run generation would otherwise open, the generation is constrained ex ante: a single run per deal, under a deterministic seed and an engine/Pack/prompt configuration whose bundle hash is committed to the OSF record **before** the engine is executed. This binds the configuration in advance, but it does not by itself *prove* that a single execution occurred — a determined author could pre-select a seed — so it is not a complete safeguard. Two facts bound the residual: the payload is non-scoring provenance (§9.3), so even a cherry-picked payload cannot move any scored quantity; and the generation is intended to run as a publicly logged CI job (pre-declared here, not yet implemented) so a third party can confirm it ran once. Until that CI job exists, payload generation rests on author good faith — disclosed as such, not papered over. This live-versus-to-be-generated distinction is stated explicitly, in the same manner as the OpenTimestamps anchoring status in §10.3, so that a reviewer inspecting the public Osservatorio today sees exactly what is already sealed (the scored forecasts) and what remains to be generated as provenance (the challenge payloads).

4.4 — SEALED-ENVELOPE RULE

The scoring protocol for all five hypotheses (§3) will use the **first** dated forecast for each deal — the T0 call as defined in §4.2 and captured in §4.3 — as the sole voted unit. The cohort *membership* is frozen (§4.1): what the append-only ledger permits is auxiliary update **rows on an already-sealed deal**, never the addition or removal of a deal. If material new information (a public regulatory filing, a revised offer announcement) is observed after T0, a new dated row may be appended to that deal's chain. Such updates are auxiliary data; they will be reported in the confidence-over-time trail and may be used in exploratory analyses (§6), but they do not alter or replace the T0 call for scoring purposes. This rule is enforced operationally by the Osservatorio's append-only structure: rows are never modified or deleted; the T0 row is identified by the earliest timestamp in the deal's record chain.

The purpose of the sealed-envelope rule is to eliminate hindsight contamination. A forecast updated the day before resolution would be nearly indistinguishable from a look-up; calibration of such a forecast would be uninformative. By pre-declaring that

only the first sealed call counts, this protocol exposes the contradictor to the maximal informational disadvantage at which a genuine forecasting system must operate.

4.5 — RIGHT-CENSORING

Deals that do not reach a resolvable public signal within the declared resolution window — 12 months for H1 and 36 months for H2 — will be declared **censored** and excluded from the scored sample for the corresponding horizon. They will not be coded as successes or failures; they will be retained in a separate censored register and reported in the sample description table. The censoring rule is pre-declared here and is not applied post-hoc to manage sample composition. A deal may be censored on H1 but resolved on H2 (or vice versa), in which case it enters the scoring pool for the resolved horizon and the censored register for the unresolved one. A deal marked *H2 not applicable by design* (§4.2) is a separate category: it is not part of the H2 pool and is not counted as censored, because its unobservability is structural and pre-declared rather than outcome-driven. **If the resolved sample for a horizon falls below three, no statistic — not even a descriptive summary — is reported for that horizon; only the raw per-deal forecast-versus-outcome rows are listed** (§6.5).

SECTION 5

Resolution rule

5.1 — DETERMINISM AS A DESIGN INVARIANT

The resolution rule is the mechanism by which each sealed T0 call — captured according to §4.3 — is coded as correct or incorrect for scoring purposes. The rule is fully deterministic: a public signal either satisfies a pre-specified threshold or it does not. No committee judgment, no contextual weighting, and no post-hoc interpretation will enter the coding decision. This determinism is not an incidental convenience; it is the condition under which calibration scoring (§6) is non-arbitrary. A resolution rule that permits discretion after the outcome is known introduces precisely the hindsight contamination that the sealed-envelope design (§4.4) is constructed to prevent.

Resolution coding is performed by the researcher acting as a mechanical rule-applier, not as an analyst. The researcher will search the canonical public sources listed in §5.2 for signals that satisfy the pre-registered threshold. If a qualifying signal is found within the declared window, the deal is coded 1 for the corresponding horizon. If no qualifying signal is found within the window, the deal is censored (§4.5) and excluded from the scored sample for that horizon. Under no circumstances will an ambiguous signal be classified as positive by interpretive judgment; ambiguous cases are censored unless a subsequent unambiguous signal arrives before the window closes. Where a resolution parameter could otherwise admit a choice — the reference

sector index, the measurement currency, or the reference value for a range-priced deal — it is **fixed per deal in Appendix D before scoring**, so that no "or"-clause is resolved after the outcome is known. To remove the remaining discretion — both in declaring a case "ambiguous" and in coding the unambiguous cases — the **primary safeguard is the determinism of the rule itself together with a fully public evidence trail**: every coded outcome is published with its code, the **primary-source URL**, and the signal date, so that *any* third party can re-code it from the public record. Given the sole-authorship conflict (§8), an **independent second coder is additionally recommended where feasible** — an individual with no financial interest in CounterBrain and no role in the engine or this protocol, blind to the forecast and the hypothesis, re-coding all resolution decisions with agreement reported as Cohen's κ — but it is not a precondition: where no independent coder is engaged, the deterministic rule and the public evidence trail govern. Disagreements, or (absent a second coder) any genuinely ambiguous case, default to censored.

5.2 — CANONICAL PUBLIC SOURCES

Resolution signals will be drawn exclusively from the following source hierarchy, listed in descending priority. A lower-priority source is consulted only when no qualifying signal is available from a higher-priority source.

1 **Official press releases** issued by the acquirer or target company and distributed through a recognized newswire service (e.g., the regulated-information feeds of Euronext, the London Stock Exchange (RNS), SIX Swiss Exchange, or Deutsche Börse, or wire services such as PRNewswire and GlobeNewswire). A press release confirming deal termination or a revised transaction price satisfies the H1 resolution criterion directly.

2 **Regulatory filings** submitted to the national securities regulator for each deal's listing venue (e.g., Consob in Italy, the AMF in France, BaFin in Germany, the FCA in the United Kingdom, FINMA in Switzerland, or the AFM in the Netherlands) or to the relevant exchange operator (Euronext, the London Stock Exchange, SIX, Deutsche Börse), including mandatory tender-offer documentation and merger-clearance filings — and, given that several cohort deals turn on antitrust clearance, decisions of the European Commission or the competent national competition authority. Filings are treated as higher-authority evidence of the announced consideration and its revisions; a binding amendment document confirming a ≥ 15% price reduction, or a prohibition decision, constitutes H1 = 1.

3 **Consolidated financial statements** — specifically, the notes to the financial statements in the acquirer's annual report or interim report, as filed with Consob or the relevant exchange. The goodwill impairment note will be used as the primary signal for the H2 value-destruction criterion.

4 **Price and index data** — verified closing prices from the relevant listing exchange's data feed (or a Refinitiv/Bloomberg equivalent) and sector-index series — STOXX Europe 600 sector indices, or the relevant national sector benchmark — used to compute TSR against the sector benchmark for the H2 criterion.

No unverified media reporting, analyst estimates, or third-party deal databases will be used as primary resolution sources. Where a secondary source is consulted for investigative scoping (to identify the existence of a potential signal), the signal must then be confirmed against a primary source before coding.

The secondary scoping layer is the Osservatorio's **deterministic news-resolution engine** (built on the **newsapi.ai / Event Registry** API, filtered by a reputable-source allowlist and an entity-match rule, then adjudicated for materiality): it continuously surveys public news for candidate resolution signals on each cohort deal and flags them for primary-source confirmation. The engine never codes an outcome by itself — it only surfaces a candidate; every flagged signal is verified against a §5.2 primary source before H1/H2 coding, and the deterministic resolution rule (§5.1) is what assigns the code.

5.3 — THRESHOLD OPERATIONALIZATION

The H1 threshold — deal termination or downward price revision ≥ 15% relative to the original announced consideration — was introduced in §3 and operationalized in §4.2. The 15% figure is fixed in this protocol; it will not be adjusted once scoring begins. The measurement is applied to the headline transaction value as stated in the initial

announcement document; changes in assumed debt, earnout structures, or deferred consideration that were present in the original announcement and are not subsequently revised do not constitute a threshold-crossing event. Only announced revisions to the headline figure, or public confirmation of deal termination, trigger H1 = 1. Three operational rules close residual discretion, each fixed per deal in Appendix D: (i) **currency** — the ≥ 15% change is measured in the deal's original announced currency, never after conversion, so an exchange-rate movement (£/€, CHF/€) cannot trigger a false H1; (ii) **range values** — where the announced consideration is a range (e.g., UniCredit–Commerzbank, "~€35–40bn"), the threshold is applied to the midpoint, unless a single headline figure is stated in the official offer document, in which case that figure governs; (iii) **completion with remedies** — a deal that completes subject to antitrust remedies or divestitures, but whose headline consideration is not cut by ≥ 15%, is coded H1 = 0 (it did not break) and proceeds to H2 if in the listed-acquirer subset, whereas a prohibition decision or an abandonment is coded H1 = 1; a price cut strictly between 0 and 15% is coded H1 = 0; and (iv) **non-acquisition structures** — for a joint venture or a multi-party break-up (deals #4 and #5 in Appendix D) there is no single headline consideration to revise, so the ≥ 15% price criterion does not apply: H1 = 1 is triggered only by public abandonment of the combination or a prohibition decision.

The H2 threshold — goodwill impairment or acquirer TSR below its sector index by 10 percentage points — is fixed in this protocol and recorded in the Osservatorio pre-registration record prior to cohort scoring; the exact sector index and the measurement currency for each H2-subset deal (TSR and index in the same acquirer-listing currency) are fixed per row in Appendix D.

5.4 — DICTIONARY REFERENCE

The full mapping from book errors to cognitive failures, contradictor counter-moves, resolving public signals, and thresholds is maintained in **Appendix A** (`docs/paper/appendices/resolution-dictionary.md`). The dictionary is the operative reference for the researcher applying the resolution rule; the table in §5.2–§5.3 above provides the source hierarchy and the threshold logic; Appendix A provides the error-specific rows that instantiate that logic for each book error (Canepa) the Packs are designed to counter.

The dictionary rows for book errors 01, 02, and 04 are now verified against the source book (Canepa, 2026, ch. 1, 2, 4) and author-ratified, and recorded in Appendix A. See Appendix A (`docs/paper/appendices/resolution-dictionary.md`) for the complete error-specific mapping of cognitive failures, contradictor counter-moves, resolving public signals, and thresholds.

SECTION 6

Baseline and analysis plan

6.1 — PRE-REGISTERED NULL BASELINE

The contradictor earns evidential credit only if it beats a pre-registered null that a naïve observer could replicate without access to the engine's output. Two null models are pre-registered.

Null (a) — Matched base-rate model. The relevant base rate is the frequency with which **contested, announced-but-pending European large-cap deals** resolve as H1 (deal termination or downward price revision ≥ 15%) — *not* the general M&A base rate. This distinction is load-bearing: because the cohort is editorially assembled toward contested deals (§4.1, COI-3), benchmarking against the ≈ 10% general rate would hand the contradictor spurious credit for a cohort selected to be eventful. The matched base rate is therefore fixed before scoring by a **deterministic derivation pre-registered in the OSF record**, built from **public, reproducible sources** so that any reviewer can re-derive it rather than trust a proprietary figure: (i) the **UK Takeover Panel** published transaction statistics (firm offers announced versus lapsed/withdrawn, 2020 onward) for the deal-failure channel; (ii) the **European Commission** merger-control statistics (prohibitions and pre-decision withdrawals under the EU Merger Regulation), together with the relevant national competition authorities (e.g., AGCM, CMA, Bundeskartellamt), for the regulatory-block channel; and (iii) at least one peer-reviewed estimate of large-cap deal completion/withdrawal rates. Population: European deals ≥ €1bn announced and contested in a declared historical window; outcome: termination, regulatory prohibition or pre-decision withdrawal, or headline cut ≥ 15% within 12 months. The figure is frozen as a single number in the OSF record, not left "to be confirmed"; a proprietary database (Mergermarket, LSEG/Refinitiv, S&P Capital IQ) may be used only as a confirmatory cross-check, never as the primary source, since a reviewer cannot reproduce it. (For orientation: among *contested* deals the failure/block rate is far higher than the ≈ 10% all-deal rate — roughly a third of EU Phase-II merger cases historically end in prohibition or are withdrawn — which is why the matched base rate, not the general one, is the honest floor; the exact frozen figure governs.) A constant forecast equal to this matched base rate is the floor: any system whose Brier score does not improve on it provides no incremental calibration signal, and an eventful cohort earns no free credit because the floor is matched to that eventfulness. As an independent corroboration — never a primary source — the Osservatorio news engine (newsapi.ai / Event Registry; §5.2) is queried over the same historical window for deal-collapse and blocked-merger coverage, to sanity-check the order of magnitude of the public-source figure; any discrepancy is resolved in favour of the Takeover Panel / Commission counts, since news coverage is not a clean deal census. The matched rate is anchored to **published** EU Commission Phase-II and

academic deal-failure figures — reproducible by reading the sources, with **no raw-data collection required** — giving a pre-registered value of **p* = 0.30** (a hard, honest floor). The full derivation and the *optional* UK-Takeover-Panel refinement (which can only raise p*) are in **Appendix E**. Two qualifications travel with this number. **Reference-class caveat:** the ≈ ⅓ anchor is the EU Phase-II *prohibition-or-withdrawal* share, whereas H1 also counts a downward headline revision ≥ 15%, and the cohort is not a pure Phase-II sample; the price-revision channel is real and not separately estimated here, so p* = 0.30 is best read as a central estimate rather than a guaranteed conservative ceiling. **Pre-registered sensitivity band:** every powered-stage H1 conclusion will be reported across p* ∈ [0.25, 0.40], so that no result depends on the point value.

Null (b) — Public-signal model. The second null uses contemporaneous public information available at T0: the mean analyst consensus rating (Buy / Neutral / Sell or equivalent, converted to a numeric scale) and the median 12-month price target, sourced from Refinitiv I/B/E/S at T0 (the sealing date), together with the market reaction to the announcement (acquirer cumulative abnormal return over the three trading days bracketing the announcement, computed against the relevant STOXX Europe 600 sector index). With the fixed cohort N = 10, a three-predictor logistic model cannot be fit without overfitting; under the underpowered regime (§6.5) null (b) therefore reduces to a **single-predictor comparator — the analyst-consensus-implied probability of H1**, derived from the modal rating and the price-target premium/discount via a **fixed, pre-registered map** declared before scoring: the modal rating sets a base H1 probability (Buy → 0.08, Neutral → 0.15, Sell → 0.30), adjusted by the price-target gap to the current price (each full 10-percentage-point discount adds 0.05, each 10-point premium subtracts 0.05; clipped to [0.02, 0.90] — the upper bound is wide enough not to handicap the null where the contradictor is most aggressive). The numeric map is frozen in the OSF record before any outcome is observed, so it cannot be tuned to determine who wins H-pay; ideally its three anchor values would be calibrated to the historical break-rate by modal rating in the same database used for null (a), and the powered stage will do so. Null (b) is defined for **H1 only**: an analyst rating and a 12-month price target speak to the acquirer's equity, not to an 18–36-month goodwill impairment, so they are not a valid public-signal proxy for H2; H-pay is accordingly evaluated on H1 only (§3). Null (b)'s Brier score is the benchmark for H-pay. The full logistic specification (modal rating, price-target premium/discount to current price, and announcement-window CAR, fit once on the cohort) is pre-registered but deferred to a powered stage; the fixed N does not support it here.

6.2 — SCORING PROTOCOL

Brier score. All five hypotheses are assessed through proper scoring rules (Brier, 1950). The primary metric is the Brier score $BS = (1/N)\ \Sigma\ (p_i - o_i)^2$, where p_i is the contradictor's stated probability and $o_i \in \{0, 1\}$ is the resolved binary outcome. Lower

scores indicate better calibration. The Brier score is computed separately for H1 and H2 horizons (no cross-horizon pooling), and separately for the full cohort and the divergent sub-sample (§6.3). Brier score decomposition into reliability, resolution, and uncertainty components (Murphy, 1973) will be reported **only when the resolved sample meets the power floor (§6.5)**; below it, binning-based decomposition is dominated by sampling noise and is omitted.

Calibration curve — H-cal. A reliability diagram (calibration curve) will be constructed by grouping stated probabilities into bins of width 0.10 and plotting the mean forecast against the observed outcome frequency in each bin. The formal calibration test for H-cal is the **Hosmer–Lemeshow goodness-of-fit test** (Hosmer and Lemeshow, 1980) with ten decile-based groups, evaluated at significance level $\alpha = 0.10$. H-cal is confirmed if the Hosmer–Lemeshow p-value exceeds 0.10 (i.e., the null hypothesis of calibration is not rejected), computed over the full resolved sample for each horizon independently. This threshold acknowledges the small-N regime; the choice of $\alpha = 0.10$ rather than 0.05 is a pre-registered concession to statistical power, not a relaxation of the substantive calibration standard. The calibration curve is plotted regardless of sample size; the Hosmer–Lemeshow test is computed only when the resolved sample size meets or exceeds the power floor defined in §6.5.

Divergence-that-pays hit-rate — H-pay. Within the sub-sample of calls flagged as divergent under H-div (§6.3), the contradictor's Brier score is compared against the Brier score of the public-signal model (null (b)) evaluated on the same sub-sample. H-pay is confirmed if the contradictor's Brier score is lower than the public-signal model's Brier score at a one-sided threshold of $\alpha = 0.10$, assessed by bootstrap resampling (2,000 replications) of the score difference. The one-sided formulation reflects the directional nature of the claim: the contradictor is claimed to add signal, not merely to tie. The bootstrap is used rather than asymptotic tests because the sample is expected to be small and the score-difference distribution is unlikely to be normal. The bootstrap is **suppressed when the divergent sub-sample is below the absolute floor of §6.3**: on a handful of deals a 2,000-replication bootstrap resamples only a few distinct values and is cosmetic, so it is not reported.

Exploratory — Δ-CSI as covariate. The Δ-CSI computed by the `csi-v1` scorer and the divergence method recorded at T0 (`cosine`, or Jaccard fallback) will be included as covariates in a post-hoc linear regression of Brier score against Δ-CSI and divergence-method indicator. This analysis is **exploratory only**: no hypothesis pre-registers a relationship between Δ-CSI and forecast accuracy, and any observed association will be reported as descriptive and treated as powered-stage material (§3, Scope constraints). A positive association between Δ-CSI and Brier score error (i.e., higher challenge intensity correlating with *worse* calibration) would itself be a noteworthy null finding and will be reported as such.

6.3 — DIVERGENCE OPERATIONALIZATION — H-DIV

A call is classified as **divergent** if the contradictor's stated probability P(deal-break or downward renegotiation ≥ 15%) exceeds the implied probability derived from the public-signal model's consensus rating by more than 15 percentage points, or if the contradictor's probability is below the public-signal model's consensus probability by the same margin (i.e., the rule captures absolute divergence from consensus — upward or downward — at the per-deal level; this is a magnitude criterion on the forecast itself, not a two-tailed statistical significance test). The minimum divergence rate defining "non-trivial" for H-div is pre-registered at 25% of the cohort. This rate is chosen to exclude noise disagreement — the mechanical small deviations produced by any probabilistic system — while permitting meaningful aggregation for H-pay. Because 25% of N = 10 is only 2–3 deals, an **absolute floor of five divergent resolved calls** is also pre-registered: below it, H-pay is not reported even descriptively (§3, §6.2), so that a one- or two-deal anecdote cannot be presented as signal.

6.4 — CALIBRATION COHORTS

Calibration statistics are **never pooled across cohorts defined by different scoring formulas or divergence methods**. A change in Δ-CSI formula version or divergence method constitutes a cohort boundary: scores before and after are reported separately and are not combined into a single calibration curve or Brier score aggregate. The current protocol cohort is designated **`(csi-v1, cosine)`**: all sessions in this cohort use the `csi-v1` formula as defined in §2.3 and compute the divergence component via cosine distance on semantic embeddings. If an embedding provider is unavailable and the Jaccard fallback is used for a session, that session is flagged in the output metadata and reported separately; it is not silently merged into the primary cohort. Any future change to the Δ-CSI formula or the divergence method requires a new cohort label, a new pre-registration addendum, and a fresh scoring baseline.

6.5 — POWER AND MINIMUM N

A Brier score decomposition and a meaningful calibration curve require a minimum resolved sample. Based on standard recommendations for reliability diagrams and the Hosmer–Lemeshow test, the power floor for confirmatory calibration testing is **N ≥ 20 resolved observations per scored horizon** (aspirational target ≈ 30). The fixed cohort of this study (N = 10; §4.1) is **below this floor by design**: the underpowered regime described next is therefore the study's **operating regime**, not a contingent fallback.

Because the cohort is fixed below N = 20 per horizon, the following pre-registered regime applies: the Hosmer–Lemeshow test and the bootstrap comparison for H-pay are not computed; the calibration curve and any binned summary — including tertile

tables — are **not** reported, because with N = 10 the observed frequency in any bin is dominated by sampling noise (a single bin of three deals has a standard error near 0.27). Instead, the **raw per-deal forecast-versus-outcome rows are listed**; no aggregate Brier statistic, decomposition, or bootstrap interval is reported below the power floor, because at N = 10 (and four on H2) any such number would be dominated by sampling noise rather than informative — the same reason the H-pay bootstrap is suppressed under §6.2. All five hypotheses are reported as **descriptive findings**, explicitly labeled, never as pass/fail confirmatory verdicts; if a horizon's resolved sample falls below three, even the descriptive summary is withheld and only the raw rows are shown (§4.5). This is not a revision of the analysis plan; it is the pre-declared analysis path for a fixed audit cohort, and it preserves the integrity of the confirmatory machinery by refusing to apply confirmatory tests to underpowered data. Should the cohort later be extended in a powered stage — a separate pre-registration — the confirmatory tests of §6.2 apply to that larger sample under a pre-registered multiplicity control: H-pay and H-cal are evaluated in a fixed sequence (H-pay first; H-cal only if H-pay is not rejected), so the family-wise error rate of the two confirmatory tests is held at α without inflation.

Confidence-over-time trail data — the sequence of auxiliary T0-plus-update records available in the Osservatorio (§4.4) — will be reported descriptively alongside the primary scored analysis, as exploratory evidence of whether the contradictor's stated confidence evolves coherently with subsequent public information. These records do not enter the confirmatory scoring; they are auxiliary material for the powered stage.

SECTION 7

Falsification criteria (for the powered stage)

Pre-registration requires stating, in advance of any data, what would prove the contradictor useless or harmful. This section discharges that obligation ***for the powered stage****.*

The four conditions below are the falsification criteria a powered, forward-sealed cohort would apply to the claim that structured adversarial challenge produces calibrated dissent with forecasting value beyond the pre-registered baselines. They are pre-declared now so they cannot be tuned later. **None is applied as a pass/fail verdict on the N = 10 pilot** (§6.5); the pilot only reports the raw quantities that would feed them.

F1 — FAILURE TO BEAT THE NULL ON BRIER SCORE OR CALIBRATION

The primary falsification condition is that the contradictor will have failed to beat either null model (§6.1) on the pre-registered scoring metrics (§6.2). Specifically: if the contradictor's Brier score on the full cohort does not improve on null (a) at the pre-

registered threshold (§6), or if the Hosmer–Lemeshow test for H-cal rejects the calibration null at the pre-registered significance level (§6), or if the divergent sub-sample's Brier score under H-pay does not improve on null (b) at the pre-registered threshold (§6), the powered stage will declare the contradictor's calibration claim falsified. The thresholds are fixed in §6 and will not be adjusted after cohort sealing. No value is restated here; the operative reference is §6. On the N = 10 pilot, F1 is not applied as a verdict (§6.5): the pilot reports only the raw quantities that F1 would consume; the formal pass/fail application is reserved for a powered sample.

F2 — DECISION PARALYSIS OR DROP IN COMMITMENT

A contradictor that causes decision-makers to become unable to commit — indefinite deferral, analysis paralysis, or a measurable reduction in decision throughput relative to an uncontradicted baseline — would constitute harm even if calibration metrics were met. This condition is **not a confirmatory test of the present protocol**: the behavioral data required to measure commitment rates are unavailable until the powered stage and lie outside this protocol's scope. It is pre-declared here as a known limit and as a Stage 2 falsification criterion. If Stage 2 data reveal a systematic drop in decision commitment attributable to the contradictor's insertion into the deliberative process, the Stage 2 report will declare this finding and its implications for the system's net value.

F3 — Δ-CSI GAMING

The Δ-CSI is defined in §2.3 as a proxy for challenge intensity — not as a measure of decision quality or a predictor of correctness. A gaming failure mode would arise if output shows inflated Δ-CSI scores without any corresponding improvement in Brier score or calibration, indicating that the pipeline's structural incentives produce verbosity or surface-level challenge that scores high on intensity while contributing no forecasting signal. The exploratory Δ-CSI regression (§6.2) is designed to detect this pattern. If a positive association between Δ-CSI and Brier score error (i.e., greater challenge intensity associated with worse forecasting accuracy; recall lower Brier = better) is observed and reaches a threshold warranting concern, the Stage 2 report will flag this as evidence of gaming and treat it as a specification failure requiring a revised `csi-v1` formula or component weighting. No confirmatory falsification rule is applied in the present protocol for this condition; its exploratory status is pre-declared.

F4 — CONVERGENCE ON A SHARED BIAS

The three-pass pipeline (§2.4) uses model isolation — thesis and red-team running on different LLM providers — to decorrelate systematic blind-spots. If the red-team agent's output systematically echoes the thesis rather than challenging it — that is, if the

divergence component of the Δ-CSI computation is near-zero across a substantial fraction of cohort sessions — the contradictor will have failed at its architectural mandate. Such convergence would mean the adversarial structure is cosmetic: the pipeline produces the appearance of challenge while both agents share the same directional error. The Stage 2 analysis will test this by examining the distribution of divergence scores and inspecting a random sample of sessions where that score falls below a level indicating near-zero divergence — operationally defined as the lower decile of the cohort divergence distribution. If systematic echo is confirmed, the Stage 2 report will declare the red-team pass structurally compromised for the cohort under study.

If any of the above conditions emerge in Stage 2, the Stage 2 report will declare them explicitly, report their magnitude, and assess their implications for the study's confirmatory and exploratory claims. Pre-declaring these conditions is not a formality; it is the mechanism that gives the Stage 2 findings their evidential weight.

SECTION 8

Conflicts of interest and safeguards

Three overlapping interests require explicit disclosure, followed by the structural mitigations that this protocol has erected against them.

COI-1 — AUTHOR AS BOOK AUTHOR, SYSTEM BUILDER, AND STUDY AUTHOR

The cognitive-contradictor system tested here, and the BRAIN Pack architecture that instantiates it, were designed by the same person who wrote the book — *10 errori che distruggono valore nelle operazioni di M&A* (Canepa) — from which the underlying error taxonomy is drawn and who is now authoring this registered protocol. The book formulates hypotheses about classes of cognitive failure and their organizational consequences; the present study operationalizes those hypotheses as pre-registered claims. This creates a motivated-reasoning risk: a builder-author has an incentive to construct a protocol whose procedures tilt toward confirmation of the very framework the author developed. The mitigation is not a change in authorship — sole authorship is a stated condition — but transparency of the dependency: the book is declared a source of *hypotheses*, not of evidence. No finding in Stage 2 can be cited as support for any claim whose origin is the book itself; book-sourced claims require independent evidence beyond this study's scope.

COI-2 — THE PAPER LEGITIMIZES THE MEASUREMENT METHOD THE PRODUCT REQUIRES

The Δ-CSI (Cognitive Sovereignty Index delta) is defined in §2.3 as a proxy for challenge intensity — the degree to which the contradictor pipeline stressed a given decision. It is not a predictor of decision correctness, and §2.3 pre-registers that framing explicitly. However, CounterBrain/BRAIN, the commercial product built on this engine, requires a defensible intensity metric to support its value proposition: a measurable signal that a contradictor session exerted genuine challenge. This study, by publishing the `csi-v1` formula and subjecting its output to calibration scoring, contributes to the public legitimacy of that metric. The conflict is structural: even a negative Stage 2 result (Δ-CSI uncorrelated with calibration outcomes) would leave the metric published and citable. The mitigation is again transparency plus pre-registration: the Δ-CSI's status as challenge-intensity proxy — not accuracy predictor — is stated in the protocol, in the engine's methodology documentation, and in §2.3 and §3 of this paper. Any Stage 2 finding that treats Δ-CSI as a correctness signal will be pre-refuted by this pre-registration.

COI-3 — THE COHORT WAS EDITORIALLY ASSEMBLED BY THE AUTHOR

The ten deals were selected by the author as flagship, contemporaneously contested European transactions, rather than drawn by a mechanical inclusion rule (§4.1). Editorial selection is a judgment that could, in principle, be exercised to favour deals the author expects to vindicate the contradictor. Two mitigations bound this. First, the **pre-outcome public seal**: every deal was sealed and hash-anchored before any of its scored outcomes was observable, and the cohort is frozen (§4.1, §4.4), so selection cannot have been motivated by outcomes that did not yet exist, and the immutable, append-only structure makes post-hoc removal of an unfavourable deal impossible. Second, the **base-rate null** (§6.1, null (a)): the residual exposure — that the author chose contested deals more likely to break than a random large-cap deal — earns the contradictor nothing, because credit accrues only by beating the realised base rate of the actual cohort, not by the cohort being eventful. The cost of editorial selection is paid in statistical power and generalizability (a small, non-random cohort; §6.5), not in hindsight contamination.

THE COMPOSITE CONFLICT, STATED WITHOUT EUPHEMISM

These three conflicts share one subject: the same person wrote the book, built the engine, designed the metric, selected the cohort, and applies the resolution rule. The structural mitigations below protect against *hindsight* contamination — post-outcome insertion, deletion, or re-coding — and they are strong. They do **not** by themselves

neutralise *construct* bias: the possibility that the taxonomy, the Pack, the thresholds (15%, 10 p.p.), and the scenario→hypothesis mappings are all tuned, even in good faith, toward confirmation. This residual is not dissolved by transparency alone. This protocol therefore adds external independence at the most exposed node — resolution coding — primarily through a **fully public, auditable evidence trail**: every coded outcome is published with its primary-source URL and date, so that any third party can re-code it from the public record (§5.1). An **independent second coder is recommended where feasible** (no financial interest in CounterBrain, no role in the engine or protocol, blind to the forecast and hypothesis, agreement reported as Cohen's κ) as a further check — but the independence of the coding does not depend on recruiting one: it rests on the determinism of the rule and the public record. Where independence is not yet in place (forecast construction, metric design), the conflict is recorded as a **disclosed residual risk that the present protocol cannot fully neutralise**, not as a risk the mitigations eliminate — calling it otherwise would be the very sycophancy this system exists to refuse.

STRUCTURAL MITIGATIONS — WHY THESE ARE NOT RHETORICAL

A disclosure that names conflicts without altering the incentive structure is a formality, not a safeguard. Four features of this protocol make the mitigations structural rather than rhetorical.

Pre-registration with independent timestamp. Every T0 prediction is written to the Osservatorio before any outcome is observable (§4.3); the ten cohort deals are already publicly sealed, each with a published content hash (Appendix D), which both demonstrates that the machinery operates from sealing through anchoring and fixes the forecasts in time. The Osservatorio is an append-only ledger already in production; each record hash is exposed via a public API, and the **independent** anchor against which a third party verifies the date of creation is the OSF registration timestamp — and, once live, the OpenTimestamps blockchain proof (§10.3) — not the author-controlled repository. This means no prediction can be inserted, revised, or post-dated after an outcome becomes known. The sealed-envelope rule (§4.4) makes the first-dated T0 call the sole scored unit; the append-only structure enforces it operationally, not contractually.

Deterministic public resolution. The resolution rule (§5) converts each sealed forecast into a binary outcome code by mechanical application of a pre-specified threshold to a hierarchy of canonical public sources. No committee, no contextual judgment, and no post-hoc interpretation enters the coding decision (§5.1). The researcher acts as a rule-applier, not an analyst. This determinism removes the principal lever through which motivated reasoning typically enters calibration scoring — the ambiguous case classified charitably. Under this protocol, ambiguous cases are censored (§4.5), not coded favorably.

Open code, prompts, and Packs. The contradictor engine code, the system-prompt strings for all three pipeline passes, and the BRAIN Pack YAML files that instantiate domain-specific behavior are committed to a public repository as part of the reproducibility commitment (§10). Any reviewer can inspect whether the prompts structurally enforce the contradictor mandate or subtly soften it.

Per-call version freeze via bundle hash. Each T0 record includes a SHA-256 bundle hash covering the exact model identifier for each of the three pipeline passes, the resolved Pack, and the full system-prompt string used in each pass (§4.3). This hash makes it verifiable that the forecast was produced by the declared model-prompt combination. A researcher who suspects that the prompt was quietly changed between prediction and resolution can check the hash against the committed code. The version freeze closes the gap between nominal and actual reproducibility.

> *Falsification — not confirmation — is the asymmetric benefit these mitigations provide.*

Together, these four features constitute a protocol whose mitigations operate on the process, not on the author's stated intentions. These four features bound *hindsight* contamination — post-outcome insertion, deletion, or re-coding; they do **not** bound *construct* bias (the taxonomy, the thresholds, and the scenario→hypothesis mappings remain the author's), which is recorded above as a disclosed residual the present stage cannot neutralise. Their force derives from the fact that falsification — not confirmation — is the asymmetric benefit they provide: a sealed, timestamped, deterministically resolved prediction that fails to beat the null (§7, F1) is unambiguously a failure, regardless of the author's interest in a positive outcome.

SECTION 9

Portability (declared, non-evidence)

This section declares two portability lanes — B (judicial/regulatory) and C (forecasting visibility) — as pre-announced, non-evidence extensions of the contradictor framework. Neither lane is inside the head calibration of this study. The calibration claims of §3–§6 cover **lane A only**: the sealed M&A prediction cohort described in §4. Lanes B and C are declared here to place them on the record, to specify the conditions under which they would be pursued, and to draw a firewall that prevents their future outputs from being cited as confirmatory evidence for the claims of this protocol.

9.1 — LANE B: JUDICIAL AND REGULATORY DECISIONS (DECLARED, STARTS LATER)

Lane B is a pre-announced portability test for the contradictor's structured adversarial challenge in legal and regulatory decision contexts. It is **outside the calibration of this study** and is described here solely as a declared intention, not as a result or as supporting evidence.

The proposed universe stub for lane B consists of contested administrative or judicial decisions in which a pre-decision brief could have been generated at a temporally well-defined T0. Candidate document classes include appellate rulings from the Italian Corte di Cassazione or Consiglio di Stato, or antitrust decisions issued by AGCM or the European Commission under Article 101/102 TFEU, where both the announcement of the proceeding and the final ruling carry public timestamps that could anchor a sealed call and its resolution. The selection criteria, size threshold, and resolution rules for a lane B cohort have not been specified in this protocol; they will be fixed in a separate pre-registration addendum drafted only after a qualified legal expert has been engaged. No call will be entered into a lane B record — and no lane B output will be produced by the contradictor engine — before that addendum is finalized and timestamped.

Lane B requires domain specialization that exceeds the scope of the current M&A BRAIN Pack: legal proceedings involve rules of admissibility, burden allocation, and procedural posture that require a Pack authored or reviewed by a legal expert. Until such a Pack exists and has been formally reviewed, any contradictor output on legal or regulatory material is for **demonstration purposes only** and must not be interpreted as a calibrated forecast or cited as evidence for any hypothesis.

Lane B is structurally outside the head calibration. Its results, when eventually produced, will be scored against a separate baseline in a separate cohort and will be reported in a separate publication or pre-registration addendum. Under no circumstances will lane B outcomes be pooled with lane A observations to augment the sample analyzed under §3–§6.

9.2 — LANE C: FORECASTING VISIBILITY (DEMONSTRATION, NOT EVIDENCE)

Lane C is a demonstration exercise using dated public questions from established probabilistic forecasting tournaments. Its outputs are explicitly branded as **demonstration, not evidence**. Lane C is **never inside the head calibration** of this study, and nothing produced under lane C will be treated as confirmatory, as a secondary calibration, or as a robustness check for any hypothesis in §3.

The proposed instrument is a selection of resolvable public questions published by recognized forecasting platforms — for example, Good Judgment Open (Tetlock and Gardner, 2015; Good Judgment Inc., 2015), Metaculus (Metaculus, Inc., 2015), or equivalent peer-reviewed tournament infrastructure — where the question close date, resolution criterion, and community probability series are publicly archived. The contradictor engine will be applied to a small set of such questions, with each call sealed at the question's open date and resolved against the platform's official resolution. The purpose is to illustrate how the contradictor schema — implicit assumptions, counter-intuitive scenario, falsification tests, burden-of-proof questions, and Δ-CSI — maps onto the tournament forecasting format, and to generate observable output that practitioners can inspect.

Lane C outputs will not be calibration-scored under the §6 protocol. No Brier score will be computed against the pre-registered null of §6.1. No hypothesis in §3 will be evaluated on lane C data. The Δ-CSI values produced in lane C sessions will be reported descriptively only and will not enter the calibration cohort `(csi-v1, cosine)` defined in §6.4. If, in a future study, a researcher wishes to subject lane C calls to calibration scoring, a new pre-registration must be filed with its own resolution rule, baseline, and falsification criteria before any scoring is performed.

9.3 — THE LANE FIREWALL

The purpose of this section is not to present results but to close a potential interpretive gap before cohort scoring begins. Three future scenarios would misuse lanes B and C, and this pre-registration forecloses them:

1 **Treating lane B or C outputs as replication.** Any statement in a Stage 2 document or subsequent publication to the effect that "lane B results replicate lane A findings" would constitute an unsupported claim. Lanes B and C are different domains, different populations, different instruments, and, where applicable, different Packs. Replication requires a matching pre-registered protocol.

2 **Pooling lane B or C Δ-CSI values with lane A values for calibration statistics.** The cohort boundary rule of §6.4 applies: calibration statistics are never pooled across cohorts. Lanes B and C constitute separate cohort namespaces.

3 **Citing lane C tournament demonstrations as evidence for H-cal or H-pay.** Lane C is a demonstration instrument. Its epistemic status is illustrative, not confirmatory. This designation is permanent: no subsequent decision by the author or a co-author can retroactively elevate lane C outputs to evidence without a new, independent pre-registration covering the tournament domain.

4 **Substituting the later challenge payload for the publicly sealed forecast (the lane-A-internal firewall).** The most proximate misuse is internal to lane A, not across lanes. The scored quantities are the P(H1)/P(H2) values sealed publicly on 20 June 2026 (Appendix D, §D.2). The full contradictor payload generated afterward as provenance (§4.3) **cannot alter them**: if the regenerated payload implies different probabilities, the publicly sealed values govern, and the divergence is reported rather than silently reconciled. This forecloses quietly upgrading a sealed forecast to a freshly generated one before scoring.

SECTION 10

Reproducibility and ethics

10.1 — PUBLIC REPOSITORY

The following artefacts will be published to a public version-controlled repository before cohort scoring begins:

- [ ] **Engine code.** The full source of the three-pass contradictor pipeline, including the Contradiction Floor enforcement, the closed-enum provenance check, and the sub-threshold calibration gate described in §2.4.
- [ ] **Scorer code.** The `csi-v1` Δ-CSI scoring function as specified in §2.3, together with its unit tests.
- [ ] **System-prompt strings.** The verbatim system-prompt for each of the three pipeline passes (Thesis, Red Team, Synthesis), frozen at the cohort sealing date.
- [ ] **BRAIN Pack YAML.** The M&A Pack that resolves domain-specific behavior — bias taxonomy, red-team persona directive, falsification templates, and operative guardrails — for all lane A calls.

The repository will be indexed on OSF (Open Science Framework) as the registered pre-print anchor. The OSF registration will be filed before any deal is scored; no call will be scored before the OSF record carries its own timestamp. (Scoring depends on the sealed forecaster probabilities and the OSF anchor, not on the optional, non-scoring engine payload of §4.3.) Any reviewer can therefore compare the committed prompts and scorer code against the bundle hash embedded in each sealed T0 record (§4.3) to verify that the declared artefacts match those used at runtime.

10.2 — PER-CALL VERSION FREEZE AND BUNDLE HASH

Each T0 record in the Osservatorio (§4.3) includes a **bundle hash**: a SHA-256 digest produced by concatenating the exact model identifier and version string for each of the three pipeline passes (thesis, red team, synthesis), the resolved Pack YAML in canonical form, and the full system-prompt string for each of the three pipeline passes, in that order. The hash is computed at call-finalization time, before the Unix timestamp is written. Its purpose is to bind the contradictor **challenge payload** unambiguously to the exact model-prompt combination that produced it. Note the distinction of §4.3: the *scored forecast* is the publicly sealed P(H1)/P(H2), which the bundle hash documents as provenance but does not itself generate.

The versioning rule is:

1. **Frozen per call.** Model, prompt, and Pack are locked at the moment the call is finalized. An engine update that modifies any of the three components constitutes a version change.

2. **Pre-registered before deployment.** Any change to the engine that alters the bundle hash must be announced in a pre-registration addendum filed with the OSF record *before* the updated engine processes any cohort call. Calls processed under the new version are tagged with the new hash; no retroactive relabelling is permitted.

3. **Calls tagged to their version.** Each row in the Osservatorio stores its bundle hash. The cohort boundary rule of §6.4 applies: sessions whose bundle hashes correspond to different prompt or Pack versions are reported under separate cohort labels and are never pooled for calibration statistics.

4. **No retroactive edits.** The Osservatorio is an append-only store; rows are never modified or deleted. A researcher who suspects a post-hoc prompt change can compare the bundle hash in any row against the hash of the artefacts published in the repository at the corresponding commit timestamp.

The full verification procedure — what is hashed, where the hash is stored, and how a third party reproduces the digest — is described in **Appendix C** (`docs/paper/appendices/freeze-versioning.md`).

10.3 — TIMESTAMPING VIA OPENTIMESTAMPS

Each Osservatorio record hash will be submitted to **OpenTimestamps** (Todd, 2016) for anchoring in the Bitcoin blockchain. OpenTimestamps produces a cryptographic proof that a given digest existed before a particular Bitcoin block, independently of any party affiliated with this study. This provides a second, immutable timestamp anchor that operates outside the Osservatorio's own infrastructure.

OpenTimestamps anchoring is currently in the Osservatorio backlog and will be operational before cohort scoring begins. Until it is operational, the independent anchor is the **OSF registration timestamp** (a third party), against which each sealed record hash is registered; the Osservatorio's own append-only record hash and git commit history corroborate this but are author-controlled (a private repository admits history rewriting) and so do not by themselves constitute an independent anchor. Once OpenTimestamps is activated, each record will additionally carry a blockchain proof. The anchors are independent of the author; either the OSF timestamp or the blockchain proof is sufficient to establish pre-existence of the sealed forecast. Honest current status: until the OSF registration is actually filed, the 20 June seals rest on an author-controlled git history and a local lock file and are **not yet independently anchored**; no forecast is scored before OSF filing (§10.1), so independent anchoring is a precondition of scoring, not an accomplished fact at the time of this draft.

10.4 — ETHICS

Human subjects — present protocol. This study collects no data from human subjects. Lane A (the M&A prediction cohort) operates exclusively on public data: announced transaction terms, analyst consensus ratings sourced from public databases, and publicly filed regulatory documents. The contradictor engine processes these public artefacts autonomously; no individual's personal data, communications, or behavioral records are collected or processed. This protocol therefore falls outside the scope of institutional ethics review requirements for human-subjects research, as defined by standard frameworks (e.g., the Declaration of Helsinki; Italian legislative decree 196/2003 as amended by decree 101/2018 on GDPR implementation).

Stage 2 human-subjects protocol. Stage 2 of this research programme will involve human subjects — decision-makers in organizational settings who interact with the contradictor engine and whose behavioral outcomes (decision commitment rates, belief updating, reported confidence) are measured. Stage 2 will require a separate, full ethics review by a recognized institutional review board (IRB or equivalent) before any data collection begins. The Stage 2 pre-registration, when filed, will include the IRB approval number as a prerequisite condition. No Stage 2 data collection will proceed

without that approval. This sequencing is stated here to pre-declare the ethical gating requirement rather than to defer it.

Data handling and privacy. The Sovereign Decision Ledger is per-tenant and locally stored; no deal-level data are transmitted to external model providers in a form that could re-identify a specific real-world transaction. System prompts pass the decision block as structured input; the model providers' data-use terms apply. The author will review and document those terms in the OSF registration before cohort scoring begins. Ledger records are encrypted at rest; access is logged and auditable.

Authorial integrity. The conflicts of interest declared in §8 extend to the reproducibility commitment: making the engine code, prompts, and Packs publicly available is the mechanism that converts the §8 disclosures from rhetorical to structural. The bundle hash closes the gap between the declared artefacts and those actually used; the OpenTimestamps anchor closes the gap between a researcher's stated timeline and a verifiable one.

> *Appendix B (`docs/paper/appendices/illustrative-card.md`) presents one clearly-labelled invented example of a contradictor output card, illustrating the seven mandatory output fields (this card populates six; `sources[]` is empty because no retrieval is simulated) and the Δ-CSI with scenario probabilities summing to 100%.*

> *Appendix C (`docs/paper/appendices/freeze-versioning.md`) describes the bundle-hash mechanism in full technical detail: what is hashed, where the digest is stored, and how a third party verifies a call's version.*

> *Appendix D (`docs/paper/appendices/cohort-register.md`) enumerates the fixed N = 10 cohort: each deal, its sealed content hash, its listing nexus, and its scenario-to-hypothesis (P(H1)/P(H2)) mapping.*

APPENDIX A

Resolution Dictionary

> ***Canonical language:*** *English (EN). This appendix is EN-only; the Italian paper (`calibrated-dissent.it.md`) carries a pointer note in §5 that directs readers here. Full translation is deferred and will be completed only if time permits before the powered-stage submission.*

PURPOSE

This dictionary operationalizes the resolution rule stated in §5. Each row maps a named book error to the cognitive failure it instantiates, the contradictor counter-move designed to surface it, the public signal that resolves whether the error materialized, and the threshold that converts an ambiguous signal into a binary outcome code.

The dictionary is the researcher's checklist at resolution time: given a sealed T0 call and an incoming public signal, look up the relevant error row to determine whether the signal qualifies as H1 = 1 or H2 = 1.

> **STATUS NOTE — VERIFIED AGAINST THE SOURCE BOOK**
>
> The rows for book errors **01**, **02**, and **04** are now verified against the source book (*10 errori che distruggono valore nelle operazioni di M&A*, Canepa 2026, chapters 1, 2, 4) and author-ratified as part of the pre-registration closure (Task F2, 2026-06-23). The rows are operative for cohort resolution.

TABLE — ERROR-SPECIFIC MAPPING

BOOK ERROR	COGNITIVE FAILURE	CONTRADICTOR COUNTER-MOVE	RESOLVING PUBLIC SIGNAL	THRESHOLD

01	**Action/momentum bias compounded by post-hoc rationalization.** The deal gains speed — driven by haste, competitive pressure, opportunity availability, or fear of missing out — before any written industrial thesis exists. The industrial thesis then arrives *after* signing to justify a decision already taken "di pancia" (on gut), rather than guiding it. The failure is the reversal of the correct epistemic order: the thesis should precede and constrain the target, the price, and the financing, not follow from them (Canepa 2026, ch. 1).	**Investment Thesis Memo challenge.** The contradictor's red-team pass operationalizes the book's counter-move: it (i) demands the explicit written **Investment Thesis Memo** — a one-page document covering the eight points defined in the book (strategic problem; why buy vs. grow / JV / commercial deal; ideal target profile; sources of value; minimum conditions; non-negotiables; walk-away triggers; post-acquisition plan) — written *before* any negotiation begins and used as a go/no-go filter; (ii) verifies that the correct order (thesis → target → price → financing) was followed and that no term sheet or NDA predates the memo; (iii) asks the burden-of-proof question: who in the acquirer's organisation has signed off on the industrial logic independent of the deal originator, documented before financial modelling began? If no memo exists, the pipeline halts and flags the call as insufficiently documented for T0 sealing.	Primary: acquirer's official press release at deal announcement (§5.2, source 1) — presence or absence of an explicit strategic rationale stating the industrial thesis. Secondary: subsequent press releases or interim/annual report narrative within 6–12 months identifying a strategic-fit writedown, goodwill impairment trigger, or explicit statement that the deal did not achieve expected synergies (§5.2, source 3).	H2 criterion: goodwill impairment recorded in the acquirer's consolidated financial statements in respect of the acquired entity (§5.3), within the 18–36-month window; OR TSR below sector index by 10 pp over the same window. Error 01 does not have a standalone H1 threshold; a deal lacking an industrial thesis at T0 but closing without a headline price revision is coded H1 = 0; the thesis-failure signal is expected to materialise over the longer H2 horizon.

02	**Anchoring on a single valuation number plus illusion of objectivity/control.** A model-derived number gives a reassuring sense of precision and is treated as a verdict — the price — fusing three conceptually distinct levels into one figure: (a) enterprise value (the intrinsic-value range given the buyer's specific thesis and capital structure); (b) negotiation space (the range that can actually be agreed given seller expectations, competitive tension, and deal structure); (c) deal structure (consideration mix, earnout design, net-debt adjustments, working-capital peg, warranties and indemnities). Anchoring on the number makes the model an alibi rather than a tool: the financial analysis is built to justify the price already committed, not to discover it (Canepa 2026, ch. 2).	**Value bridge ("ponte di valore") challenge.** The contradictor's red-team pass operationalizes the book's counter-move: it (i) demands that the decision block explicitly separate the three levels above and produce a **"ponte di valore" (value bridge)** document distinguishing enterprise value, net-debt adjustments, working-capital peg, deferred/conditional price components, and net-to-seller cash; (ii) requires that competing offers be compared on overall economic quality — not headline price — by translating each structural term into an economic impact figure; (iii) constructs a falsification test: under what earnout or purchase-price-adjustment scenario would the effective price paid exceed the intrinsic-value range by more than 15%? If no value bridge exists and no such scenario is modelled, the pipeline halts and flags the call as insufficiently documented for T0 sealing.	H1 primary: official announcement of price revision or deal termination (§5.2, source 1 and source 2). A post-announcement amendment to the consideration — including earnout restatement, working capital adjustment in excess of 15% of headline price, or deal termination — constitutes H1 = 1. H2 primary: goodwill impairment (§5.2, source 3) attributable to overpayment relative to synergy realisation, as disclosed in the notes to the acquirer's consolidated financial statements.	H1: deal termination or announced downward revision to headline consideration of ≥ 15% relative to original announcement value (§5.3). H2: goodwill impairment in acquirer's consolidated statements (§5.3) OR TSR below sector index by 10 pp within the 18–36-month window. Error 02 is the error most directly tied to the H1 criterion; a price/structure confusion that surfaces publicly within 6–12 months is the cleaner signal; goodwill impairment at H2 is the lagged signal when the confusion is not detected until post-integration.

04	**Anchoring to the seller's first draft plus reactive framing.** The error is the BUYER accepting the COUNTERPARTY's (seller's) LOI or SPA as the working base. Two compounding mechanisms: (a) an anchoring component — once the seller's draft is accepted as the negotiating document, its structure, definitions, representation scope, indemnity caps, and disclosure carve-outs become the anchor; every buyer revision is framed as a concession rather than an imposition, making material rebalancing psychologically and practically resisted; (b) a reactive-framing component — the buyer negotiates "in reazione, non in impostazione" (reacting, not setting the frame), progressively conceding ground via the seller's chosen definitions, caps, and disclosure regime, rather than establishing its own risk-allocation logic from the outset (Canepa 2026, ch. 4).	**Buyer's clause map challenge.** The contradictor's red-team pass operationalizes the book's counter-move: the buyer must arrive with its OWN negotiating framework — a **buyer's clause map** — before accepting any draft from the counterparty. The red-team pass (i) verifies that the buyer has pre-defined its risk-allocation logic: non-negotiables vs. tradeables, fallback positions for each economically-sensitive clause, MAC clause scope, indemnity caps, and disclosure-basket thresholds; (ii) translates every economically-sensitive clause into a monetary impact figure for the board — so the decision-maker is buying risk, not language; (iii) asks the burden-of-proof question: has the acquirer's legal and financial team produced this buyer's clause map before reviewing the seller's draft? If no buyer's clause map exists, the pipeline halts and flags the call as insufficiently documented for T0 sealing. The red-team applies even if the buyer ultimately works on the seller's draft — "think first" is the invariant.	H1 primary: official announcement of deal termination citing contractual disagreement, regulatory condition failure, or MAC invocation (§5.2, source 1 and source 2). A termination announcement citing any of these causes within 6–12 months of T0 constitutes H1 = 1. Secondary: Consob or exchange filing amending deal terms (consideration revision, condition waivers, earnout redefinition) within the 6–12-month window (§5.2, source 2). H2: goodwill impairment (§5.2, source 3) where the notes cite integration failures or warranty claims as the impairment trigger.	H1: deal termination (any public cause) OR downward revision ≥ 15% to headline consideration within the 6–12-month window (§5.3). Error 04 is the error most tightly coupled to the H1 horizon: reactive-framing failures typically surface within the first year as contractual disputes, failed conditions, or regulatory blocks that a buyer-led term sheet would have addressed. H2: TSR below sector index by 10 pp OR goodwill impairment (§5.3) for cases where the contract weakness does not surface publicly at H1 but manifests in post-integration value destruction.

COLUMN DEFINITIONS

- **Book error** — The error identifier as defined in the book manuscript. Errors 01, 02, and 04 correspond to the three cognitive failure modes that the BRAIN Packs are specifically designed to counter in M&A contexts.
- **Cognitive failure** — The named cognitive failure mode (e.g., confirmation bias variant, overconfidence form, anchoring pattern) that the book error represents. The labels are taken from the named book chapters (Canepa 2026) and author-ratified; they reflect the book's own framing of each error rather than an inference from external literature.

- [ ] **Contradictor counter-move** — The specific red-team intervention — drawn from the resolved Pack's bias-taxonomy section — that the contradictor will apply when it detects this error in the decision block.
- [ ] **Resolving public signal** — The specific document type and data field (e.g., "acquirer consolidated annual report, goodwill impairment note, IFRS 3 / IAS 36 disclosure") from which the outcome code is read, per the source hierarchy in §5.2.
- [ ] **Threshold** — The error-specific public-signal threshold that applies the H1 and H2 outcome codes defined in §5.3 to resolve each error row into a binary outcome.

CROSS-REFERENCES

§5.2	canonical source hierarchy (press releases, regulator/exchange filings, financial statements, price/index data)
§5.3	threshold operationalization (H1 and H2 outcome thresholds; not restated here)
§3	H1 and H2 quantity definitions
§4.2	prediction unit and horizon definitions
A	`packages/packs/` — BRAIN Packs YAML, which supply the bias taxonomy from which counter-moves are drawn

APPENDIX B

Illustrative Contradictor Output Card

> ***Canonical language:*** *English (EN). This appendix is EN-only; the Italian paper (`calibrated-dissent.it.md`) carries a pointer note in §10 that directs readers here.*

> **IMPORTANT — ALL DATA IN THIS APPENDIX ARE INVENTED**
>
> This card is a synthetic, illustrative example created solely to show the structure of a contradictor output. It does not represent any real company, transaction, analyst report, or historical event. No inference about real-world M&A outcomes should be drawn from any figure or narrative here. Probabilities in §B.6 are invented for illustration; they sum to 100% as required by the schema but carry no empirical basis.

PURPOSE

This appendix shows what a single contradictor output card looks like. It mirrors the seven-field output schema defined in §2.2 and the pre-registration example stub referenced in §3; this card populates six of the seven fields, leaving `sources[]` empty because no retrieval is simulated. The example uses a fictional European mid-cap acquisition so that the domain (M&A) matches the lane A universe without referencing any real deal.

SCENARIO CONTEXT (INVENTED)

> **FICTITIOUS DEAL**
>
> ### Alfa S.p.A. acquires Beta S.r.l.
>
> *All figures invented — no real company, transaction, or event*
>
> Acquirer — Alfa S.p.A. (invented listed company, Euronext Milan). Target — Beta S.r.l. (invented unlisted logistics firm). Announced consideration: €180 million (invented figure). Analyst modal rating at T0: **Buy** (4 of 6 brokers, invented). Announcement date: ⟨T0 — illustrative⟩.

FIELD 1 — IMPLICIT ASSUMPTIONS (`assumptions[]`)

#	ASSUMPTION	SEVERITY (1–5)	CATEGORY
A1	Alfa can realize the projected €12 M/yr synergies within 24 months without disrupting Beta's existing customer relationships.	4	Integration optimism

A2	Beta's revenue base is not materially dependent on two anchor clients whose contracts expire within 18 months of deal close.	5	Concentration risk
A3	Regulatory clearance under the competent national competition authority will follow the standard Phase I timeline.	3	Regulatory timing

> *(All figures and descriptions are invented.)*

FIELD 2 — COUNTER-INTUITIVE SCENARIO (COUNTER_SCENARIO)

The dominant reading treats the acquisition as a straightforward logistics bolt-on, with synergies flowing from route overlap and fleet rationalization. The counter-intuitive scenario is that Beta's apparent revenue stability masks a customer-concentration problem that Alfa's due diligence has not probed at the contract-renewal level. If the two anchor clients (representing an invented 61% of Beta's gross margin) exercise renewal options at lower rates — a negotiating move made rational by the announcement itself, which signals to them that their leverage has increased — the €12 M synergy target collapses before integration teams are in place. The counter-intuitive element is that the acquisition announcement is the trigger for the very customer-side deterioration that makes the deal value-destructive.

> *(Invented narrative. No real client, revenue, or margin figure.)*

FIELD 3 — FALSIFICATION TESTS (FALSIFICATION_TESTS[])

#	TEST	DATA REQUIRED	STATUS
FT1	Obtain the customer concentration schedule from Beta's audited accounts (top 5 clients by gross margin). If no single client exceeds 25% of gross margin, the A2 assumption is materially supported.	Beta audited financials (T0 or latest available).	Not yet run (T0 call)
FT2	Check whether the competent competition authority has issued a Phase II extension request on any comparable logistics consolidation in the 36 months prior to T0. If yes, re-estimate clearance timeline.	Competition authority public decision register.	Not yet run

> *(Invented tests. No real data source or outcome implied.)*

FIELD 4 — BURDEN-OF-PROOF QUESTIONS (BURDEN_QUESTIONS[])

#	QUESTION	OBLIGATION BEARER	OBLIGATION MET?

BQ1	Who bears the obligation to demonstrate that the anchor-client contracts carry renewal terms that survive change-of-control? Alfa's advisers, as part of legal due diligence, carry this obligation. Has the legal due diligence memo been reviewed and does it address this point explicitly?	Alfa's legal advisers	Not demonstrated at T0
BQ2	Who bears the obligation to show that the announced synergy timeline is consistent with logistics-integration precedents in European mid-market M&A? Alfa's management carries this obligation through the investor presentation. The investor presentation states the timeline but provides no comparable precedent.	Alfa management	Obligation stated, not substantiated

> *(Invented questions. No real legal opinion or investor document implied.)*

FIELD 5 — CALIBRATED CONFIDENCE (CONFIDENCE)

Level: `medium`

Rationale: The stated probability estimates (Field 6 below) rest on the public data available at T0. The customer-concentration assumption (A2) could be refuted or confirmed by contract-level due diligence data not publicly available; this is the principal knowledge gap. The regulatory timing assumption (A3) is partially observable from competition-authority precedent. Overall uncertainty is medium because the dominant reading is not implausible — logistics bolt-ons succeed frequently — but the counter-intuitive scenario identifies a mechanism that is structurally credible and not publicly rebutted at T0.

Knowledge gaps declared:

- ☐ Contract-level customer concentration schedule (not public).
- ☐ Content of legal due diligence memo on change-of-control clauses (not public).
- ☐ Competition authority internal case-load at T0 (not public, observable only approximately via public decision register lag).

> *(All characterizations are invented.)*

FIELD 6 — SCENARIO PROBABILITIES AND Δ-CSI

> **INVENTED FOR ILLUSTRATION ONLY**
>
> These numbers do not represent any calibrated forecast, historical outcome, or model output.

The three scenarios identified by this card exhaust the outcome space at the H1 horizon (6–12 months):

SCENARIO	DESCRIPTION	ASSIGNED PROBABILITY

S1 — Deal completes at announced terms	Regulatory clearance Phase I; anchor-client concentration not material; synergy timeline intact.	58 %
S2 — Downward renegotiation ≥ 15%	Customer-concentration problem surfaces during remediation period; Alfa negotiates a reduced consideration. H1 = 1 under the pre-registered threshold.	27 %
S3 — Deal termination	Phase II review blocks the transaction, or anchor-client deterioration makes the deal economically unviable before close. H1 = 1.	15 %
Total		100%

Stated P(H1) = P(S2) + P(S3) = 42%. (Invented. Analyst consensus at T0 implied probability: ~18% — also invented.)

Δ-CSI (CSI-V1) — INVENTED, ILLUSTRATIVE ONLY

COMPONENT	VALUE	NOTES
Assumption count × mean severity	0.80	3 assumptions, mean severity 4.0 → min(3,5)/5 × (4.0/5) — invented
Falsification test count	0.40	2 tests → min(2,5)/5 — invented
Burden-question count	0.40	2 questions → min(2,5)/5 — invented
Source count	0.20	1 source → min(1,5)/5 — invented
Divergence score	0.62	Cosine distance (embedding) between thesis and counter_scenario — invented
Weighted sum	57	Using default csi-v1 weights summing to 100 — invented

Δ-CSI = 57 (invented, illustrative only). This value carries no calibration claim and is not an accuracy predictor (§2.3).

CROSS-REFERENCES

- [ ] §2.2 — fixed output schema (seven mandatory fields; this card populates six of the seven; `sources[]` is empty because no retrieval is simulated here)
- [ ] §2.3 — Δ-CSI definition and csi-v1 formula

- [ ] §4.3 — sealed output format
- [ ] Appendix C — bundle hash and version freeze

> *End of Appendix B — Illustrative card. All content is invented.*

APPENDIX C

Bundle Hash and Version Freeze

> ***Canonical language:*** *English (EN). This appendix is EN-only; the Italian paper (`calibrated-dissent.it.md`) carries a pointer note in §10 that directs readers here.*

PURPOSE

This appendix describes the version-freeze mechanism introduced in §4.3 and §10.2 of the main protocol. It specifies exactly: (1) what is hashed to produce the bundle hash; (2) where the hash is stored; (3) how a third party reproduces the digest and verifies a call's version.

1. WHAT IS HASHED

The bundle hash is a **SHA-256** digest of a single concatenated byte string assembled from the following five components, in the order listed:

#	COMPONENT	DESCRIPTION
C1	Model identifier (Pass 1 — Thesis)	The exact model ID string as returned by the provider API at call time (e.g., `claude-opus-4-5-20251101`). Includes the provider prefix if any.
C2	Model identifier (Pass 2 — Red Team)	Same format. May differ from C1 if model isolation is active (§2.4).
C3	Model identifier (Pass 3 — Synthesis)	Same format.
C4	Resolved Pack YAML	The BRAIN Pack YAML as resolved at call time: core Pack merged with vertical Pack merged with client Pack, in canonical form (keys sorted lexicographically, YAML version 1.2, LF line endings, no trailing whitespace). The canonical form is deterministic: any two invocations that resolve to the same logical Pack content must produce the same byte string.

C5	System-prompt string (all three passes)	The full verbatim system-prompt string for each of the three pipeline passes, concatenated in pass order (Pass 1 || Pass 2 || Pass 3), with a null byte (`\x00`) as separator between passes. The system prompt is the complete string as sent to the provider API, after Pack resolution and template substitution.

Concatenation rule: Components are joined in the order C1, C2, C3, C4, C5 with no separator between components. The resulting byte string is encoded as UTF-8 before hashing.

Hash function: SHA-256, producing a 32-byte (256-bit) digest, encoded as a lowercase hexadecimal string of 64 characters.

> EXAMPLE (ILLUSTRATIVE, NOT A REAL CALL)
>
> `input_string = ("claude-opus-4-5-20251101" [C1] + "claude-opus-4-5-20251101" [C2, same provider here] + "claude-opus-4-5-20251101" [C3] + "<canonical-pack-yaml-bytes>" [C4] + "<pass1-prompt>\x00<pass2-prompt>\x00<pass3-prompt>" [C5]); bundle_hash = sha256(input_string.encode("utf-8")).hexdigest()` → e.g. `"a3f9c2...d17b"` (64 hex chars). (Model IDs and prompt strings above are illustrative placeholders.)

2. WHERE THE HASH IS STORED

Every T0 record in the **Osservatorio** contains a `bundle_hash` field at the top level of the record JSON:

> RECORD JSON (SCHEMA_VERSION 1.0)
>
> ```
> { "schema_version": "1.0", "deal_id": "<osservatorio-deal-id>",
> "t0_unix_utc": 1750000000, "bundle_hash": "a3f9c2...d17b", "pass_models": {
> "pass_1_thesis": "claude-opus-4-5-20251101", "pass_2_red_team": "claude-
> opus-4-5-20251101", "pass_3_synthesis": "claude-opus-4-5-20251101" },
> "pack_version": "<git-commit-sha-of-resolved-pack>", "output": { ... } }
> ```

The Osservatorio's public API exposes the `bundle_hash` field for each record. A researcher who queries a deal's T0 record receives the hash directly; they do not need to recompute it to confirm its presence.

In addition to Osservatorio storage, each T0 record hash is submitted to **OpenTimestamps** (Todd, 2016), which produces a Bitcoin-anchored cryptographic proof of existence. The OTS proof file is stored alongside the record in the Osservatorio and is also committed to the public repository.

3. HOW A THIRD PARTY VERIFIES A CALL'S VERSION

A complete verification procedure is as follows:

1 **Retrieve the T0 record.** Query the Osservatorio public API for the deal of interest. Note: `bundle_hash` (the claimed digest), `pass_models` (the three model IDs used), `pack_version` (git commit SHA of the resolved Pack), and `t0_unix_utc` (the finalization timestamp).

2 **Retrieve the committed artefacts.** From the public repository, check out the commit identified by `pack_version`. Verify that the commit timestamp predates `t0_unix_utc`. Retrieve the resolved Pack YAML at that commit (apply the same resolution order: core ▸ vertical ▸ client, as documented in `packages/packs/README.md`) and the system-prompt template files for all three passes at that commit.

3 **Reconstruct the canonical Pack YAML.** Apply the canonical-form procedure: (1) merge core, vertical, and client Pack YAMLs in resolution order; (2) sort all keys lexicographically (recursive, all nested maps); (3) serialize to YAML 1.2 with LF line endings and no trailing whitespace. Two independent implementations of the canonicalization procedure must produce the same byte string for the same logical Pack content.

4 **Reconstruct the system-prompt strings.** Apply the same template-substitution logic present in the engine code at `pack_version` to produce the three verbatim system-prompt strings. No additional substitutions are valid; the engine code is the authoritative source.

5 **Recompute the hash.** Concatenate C1–C5 in the order specified in §1 above (UTF-8 encoded, no separators between components, null-byte separators between pass prompts within C5). Compute SHA-256. The resulting lowercase hex string must equal `bundle_hash`.

6 **Verify the timestamp anchor.** If OpenTimestamps is operational for the record, verify the OTS proof file against the Bitcoin blockchain using the standard OpenTimestamps client. The proof establishes that the `bundle_hash` digest existed before the Bitcoin block whose height is recorded in the proof. If OpenTimestamps is not yet operational (pre-activation period), the OSF registration timestamp is the independent anchor. The Osservatorio record is append-only; the creation timestamp cannot be altered retroactively.

4. WHAT THE HASH DOES AND DOES NOT GUARANTEE

Guarantees:

- [ ] A match between the recomputed hash and the stored `bundle_hash` confirms that the challenge payload was produced using the exact model IDs, Pack YAML, and system-prompt strings represented at the declared `pack_version`.

- [ ] A mismatch indicates that at least one of the five components differed from the declared version — either through an undisclosed prompt change, a Pack update, or a model API change that altered the returned model ID string.

Does not guarantee:

- ☐ That the model provider's internal weights or sampling behavior were identical to any other call made with the same model ID. Model IDs identify a named checkpoint; providers may update checkpoints between releases. The hash binds the *declared identifier*, not the provider's internal state.
- ☐ That the system-prompt content is pedagogically or ethically appropriate. The hash confirms identity, not quality. Prompt inspection remains a human responsibility.

CROSS-REFERENCES

§4.3	sealed output format (bundle hash element 4)
§8	per-call version freeze as a structural COI mitigation
§10.2	versioning rule in full
§10.3	OpenTimestamps anchoring
§6.4	cohort boundary rule (different bundle hashes → different cohort labels)
B	Appendix B — illustrative card (shows the Δ-CSI field; does not show bundle hash, which is infrastructure metadata)

APPENDIX D

The sealed cohort register (lane A, fixed N = 10)

This appendix enumerates the fixed, closed study cohort referenced in §4.1 and fixes, per deal and **before scoring**, every resolution parameter that §5 would otherwise leave to discretion (reference value, currency, sector index, H2 observability, pre-T0 information). All ten deals were sealed in the Osservatorio on **20 June 2026**, before any H1 or H2 outcome was observable. Each row carries the published content hash (SHA-256) of its sealed card; the hashes are independently inspectable on the public Osservatorio (`counterbrain.com/registro` · `/en/registry`).

The cohort is **frozen**: no deal may be added or removed after sealing (§4.4). The set was assembled editorially (disclosed as COI-3, §8); the anti-hindsight guarantee rests on the pre-outcome seal anchored to the OSF timestamp (§10.3), not on a mechanical inclusion rule (§4.1).

D.1 — COHORT TABLE

#	DEAL (ACQUIRER → TARGET)	ANNOUNCED VALUE	SECTOR	LISTING NEXUS	ANNOUNCED	SEALED	LAG	SEAL HASH (PREFIX)
1	Intesa Sanpaolo → Monte dei Paschi	€30.6bn	Banking	IT (Euronext Milan)	2026-06-08	2026-06-20	12 d	`df52792d…7051e42`
2	Poste Italiane → Telecom Italia	€10.8bn	Telecom	IT (Euronext Milan)	2026-03-23	2026-06-20	~89 d	`84322e5f…4436dd16`
3	UniCredit → Commerzbank	~€35–40bn	Banking	IT–DE (Milan / Frankfurt)	2026-03-16	2026-06-20	~96 d	`0eec6cbd…0140a34f`
4	Airbus · Leonardo · Thales → satellite merger (Project Bromo)	€6.5bn	Defence	FR–IT (Paris / Milan)	2025-10-23 (MoU)	2026-06-20	~240 d	`8e83ff8f…d0e8cd89`
5	Bouygues · Iliad · Orange → SFR (break-up)	€20.35bn	Telecom	FR (Euronext Paris)	2026-06-06	2026-06-20	14 d	`8fcc2fcc…7fd3aa98`
6	Nuveen → Schroders	£9.9bn	Asset management	UK (LSE)	2026-02-12	2026-06-20	~128 d	`fff13730…ca507e55f`
7	Zurich → Beazley	£8.1bn	Insurance	UK–CH (LSE / SIX)	2026-03-02	2026-06-20	~110 d	`7ee2b124…72b9eccd`
8	EQT → Intertek	£10.9bn	Services (TIC)	UK (LSE)	2026-06-18	2026-06-20	2 d	`29b9ab04…e0388e40`
9	Ingredion → Tate & Lyle	£2.7bn	Consumer ingredients	UK (LSE)	2026-06-08	2026-06-20	12 d	`480b7a2e…fe44aca9`

10	CVC → dsm-firmenich (Animal Nutrition & Health)	€2.2 bn	Nutrition	CH–NL (SIX / Euronext Amsterdam)	2026-02-09	2026-06-20	~131 d	`f16768a7…d48c8b0f`

> *Full 64-character hashes are stored verbatim in `apps/web/src/data/predictions.json` and locked in `apps/web/src/data/sealed-hashes.lock.json`.* ***Announcement-date erratum (#3, #4).*** *The dates shown here are corrected: deal #3 (UniCredit–Commerzbank) was announced 2026-03-16, and deal #4 (Project Bromo) is dated to its 2025-10-23 MoU. The immutable seal intentionally retains the original imprecise values (deal #4 sealed as "2026"); because `announced_date` is inside the seal hash, the correction lives in this documentation and the OSF record, not in the sealed envelope — the public registro therefore still shows the original. Both deals are treated as elevated-pre-T0-information (D.3).*

D.2 — SCENARIO → HYPOTHESIS MAPPING (THE SCORED P(H1), P(H2))

Each sealed card states a three-way scenario distribution summing to 100% (§4.2). The scored probabilities are derived deterministically: **P(H1) = mass of the deal-failure scenario**; **P(H2) = mass of the value-disappointment scenario** (marginal mass, not conditioned on completion — see §4.2; scored only on the H2 subset of D.3). These are the *frozen, publicly sealed* values; the later contradictor payload cannot change them (§4.3, §9.3).

#	DEAL	P(SUCCESS)	P(H2) SCORED	P(H1) SCORED	MAPPING NOTE
1	Intesa → MPS	0.35	0.40	0.25	clean
2	Poste → TIM	0.30	0.45	0.25	clean
3	UniCredit → Commerzbank	0.20	n/a (H2 not applicable)	**0.80**	**exception**: a control bid that ends in a blocked minority stake (0.45) or withdrawal/loss (0.35) is a failure to acquire control → both map to H1; success = control (0.20).
4	Airbus/Leonardo/Thales	0.35	n/a (H2 not applicable)	**0.35**	**exception**: JV with no single consolidating acquirer; H1 = "blocked" (0.35); H2 undefined → not applicable.
5	Bouygues/Iliad/Orange → SFR	0.35	n/a (H2 not applicable)	**0.30**	**exception**: three-way break-up, no single acquirer; H1 = "blocked / undone" (0.30); H2 undefined → not applicable.

6	Nuveen → Schroders	0.60	n/a (H2 not applicable)	0.10	H1 clean; H2 n/a (acquirer Nuveen is a TIAA subsidiary, no public TSR).
7	Zurich → Beazley	0.55	0.35	0.10	clean
8	EQT → Intertek	0.70	n/a (H2 not applicable)	0.10	H1 clean; H2 n/a (PE acquirer, target delisted).
9	Ingredion → Tate & Lyle	0.50	0.20	**0.30**	**exception**: antitrust prohibition (0.30) coded H1 (deal does not complete on original terms); "completes but disappoints" (0.20) coded H2.
10	CVC → dsm-firmenich ANH	0.45	n/a (H2 not applicable)	0.15	H1 clean; H2 n/a (PE acquirer, carve-out of a business unit).

H1 scored sample: all ten deals (UniCredit included but flagged, D.3). **H2 scored sample (subset):** #1 Intesa, #2 Poste, #7 Zurich, #9 Ingredion — the only four deals with a listed acquirer that consolidates the target and a public TSR. The other six are *H2 not applicable by design* (§4.2): not censored, not scored.

> ***Note on row sums.*** *For the six H2 not applicable deals, the value-disappointment mass still exists in the sealed three-way distribution — the three scenarios sum to 100% in the original card — and is shown as n/a in the "P(H2) scored" column because it is not scored (no listed consolidating acquirer makes it observable), not because the mass is zero. In every row, P(success) + (value-disappointment mass) + P(H1) = 100%; only the middle term's scorability changes across deals. Eligibility ≠ resolution: under §4.5/§6.5, if fewer than three of the four H2-eligible deals resolve within the window, no H2 statistic is reported — only the raw rows.*

D.3 — PER-DEAL RESOLUTION PARAMETERS (FIXED BEFORE SCORING)

For each deal: whether H2 is observable, the H1 reference value and its currency (the ≥ 15% threshold is measured in this currency, never after conversion; ranges use the midpoint), the H2 sector index and currency (TSR and index in the same currency), and the elevated-pre-T0-information flag.

#	DEAL	H2 OBSERVABLE?	H1 REFERENCE VALUE (CURRENCY)	H2 SECTOR INDEX (CURRENCY)	ELEVATED PRE-T0 INFO?
1	Intesa → MPS	yes	€30.6bn (EUR)	FTSE Italia All-Share Banks, total return (EUR)	no (12 d)
2	Poste → TIM	yes	€10.8bn (EUR)	STOXX Europe 600 Financial Services, total return (EUR)	yes (~89 d)

3	UniCredit → Commerzbank	no	€37.5bn — range midpoint of €35–40bn (EUR)	—	**yes** (long lag + low first-offer acceptance and government opposition already public)
4	Airbus/Leonardo/Thales	no	€6.5bn (EUR)	—	yes (long lag — MoU 8 months before seal)
5	Bouygues/Iliad/Orange → SFR	no	€20.35bn (EUR)	—	no (14 d)
6	Nuveen → Schroders	no	£9.9bn (GBP)	—	yes (~128 d)
7	Zurich → Beazley	yes	£8.1bn (GBP)	STOXX Europe 600 Insurance, total return, in CHF (acquirer listing currency)	yes (~110 d)
8	EQT → Intertek	no	£10.9bn (GBP)	—	no (2 d)
9	Ingredion → Tate & Lyle	yes	£2.7bn (GBP)	S&P Composite 1500 Food Products, total return (USD, acquirer listing currency)	no (12 d)
10	CVC → dsm-firmenich ANH	no	€2.2bn (EUR)	—	yes (~131 d)

> ***On elevated-pre-T0-information.*** *No deal had a qualifying H1/H2 resolution signal (termination, ≥ 15% headline cut, impairment, or TSR breach) at the seal date — all ten were pending — so none is excluded under the §4.2 rule. The flag records that some deals carried more public information at T0 than an announcement-instant call would have, so the informational disadvantage is not uniform; the announce-to-seal lag (D.1) is the covariate reported in the exploratory analysis (§6.2). UniCredit–Commerzbank is the most exposed case and is flagged accordingly, but it is retained in the H1 pool because no qualifying H1 signal had occurred at T0.*

> ***On the imprecise dates (#3, #4).*** *Deals #3 and #4 carry month-/year-granularity public announcement dates; the announce-to-seal lag is reported at that granularity and both are treated as elevated-pre-T0-information. Their exact announcement dates will be fixed in the OSF record before scoring.*

APPENDIX E

Base-rate derivation protocol (null (a))

This appendix gives the **reproducible, step-by-step derivation** of the single matched base-rate figure p* used by null (a) in §6.1, so that any third party can re-execute it from public sources and obtain the same number. The executed figure, every per-source query, and the combination step are frozen in the OSF record **before** any deal is scored.

E.1 — WHAT IS BEING ESTIMATED

p* = P(an H1-type failure within 12 months | a contested European large-cap deal), where an **H1-type failure** = { deal termination **or** downward revision of the headline consideration ≥ 15% **or** a regulatory prohibition **or** a pre-decision withdrawal }. p* is the constant forecast of null (a): the Brier floor the contradictor must beat. **Honest-floor principle:** because the cohort is editorially assembled toward contested deals (COI-3, §8), a *higher* p* is the harder, honest floor; the combination rule (E.5) **never sets p* below the clean reference-class estimate**.

E.2 — REFERENCE CLASS

A historical deal qualifies if all of: (1) announced value ≥ €1bn (or £/CHF equivalent); (2) at least one party listed on a European regulated market; (3) *contested* — a competing/hostile/unsolicited bid, or a Phase-II antitrust review or documented national/political opposition. Window: the ten years before the cohort sealing year (**2016–2025**), or the longest available sub-window per source (disclosed). The window ends before the cohort year, so the base rate uses other deals, never the ten.

E.3 — OPTIONAL REFINEMENT: UK TAKEOVER PANEL RAW DATA

A clean public deal census with outcomes for UK public takeovers. Steps: (1) download the Panel's underlying transaction data per year (published since July 2023, from 1 April 2020) plus Annual Reports for earlier years; (2) restrict to firm offers under the City Code; (3) filter to ≥ £1bn (the Panel reports this count — e.g. 7 of 61 firm offers in 2023–24); (4) numerator = firm offers lapsed or withdrawn within 12 months; (5) denominator = all firm offers ≥ £1bn announced in the window; (6) `p_TP = numerator / denominator`. Scope note: UK-only; firm-offer outcomes capture the break/withdrawal channel, not post-completion ≥ 15% revisions.

E.4 — CORROBORATION CHANNELS (NEVER PRIMARY)

(a) European Commission — share of *Phase-II* cases ending in prohibition or pre-decision withdrawal (≈ one third); the overall ≈ 0.3% prohibition rate is the wrong reference (dominated by uncontested deals). **(b) Industry / academic** — large announced-deal cancellation ≈ **10%/yr** (McKinsey 2019); hostile/unsolicited bidders prevail in a **minority** of cases (Schwert 2000; Bebchuk, Coates & Subramanian 2002); larger targets less likely to complete. Peer-reviewed corroboration on terminations: Xing & Yi (2024). **(c) newsapi.ai / Event Registry** (§5.2) — order-of-magnitude sanity check only; discrepancies resolve in favour of (a) / E.3.

E.5 — COMBINATION RULE → THE SINGLE FROZEN p*

(1) Compute p_TP. (2) If the contested-subset evidence (E.4a, E.4b) indicates a materially higher rate, set p* to that matched estimate (e.g. the midpoint of the contested band); p* is never below p_TP. (3) Round to two decimals; freeze p*, the window, and every source query in OSF before scoring, after which it is fixed.

E.6 — THE PRE-REGISTERED VALUE

From public figures already on record, purely illustrative:

REFERENCE CLASS	FAILURE RATE	ROLE
Large announced deals (cancellation/yr)	**≈ 10%**	**not** p* — cohort is not a random deal
EC Phase-II (regulatory, contested)	≈ 33%	corroboration
Hostile / unsolicited (academic)	**minority prevail**	upper bound
Matched band (contested EU large-cap)	**≈ 25%–40%**	→ midpoint **≈ 0.30**

p* = 0.30, derived from the published EU Phase-II and academic deal-failure rates (E.4) under the honest-floor principle. This is the **pre-registered value**, not an illustration; it requires **no raw-data collection**, and is confirmed — and can only ever be raised — at the OSF freeze.

Sensitivity band & reference-class caveat. All powered-stage H1 conclusions are reported across p* ∈ [0.25, 0.40]; no result depends on the point value. The ⅓ Phase-II anchor is a prohibition/withdrawal share, whereas H1 also counts a downward headline revision ≥ 15% (a real channel not separately estimated), so p* = 0.30 is a central estimate, not a guaranteed conservative ceiling.

E.7 — OSF FREEZE CHECKLIST

Record before scoring: the window and per-source sub-windows; Takeover Panel file/version, filter, numerator, denominator, p_TP; EC statistic, period, Phase-II prohibition+withdrawal share; academic citation + figure used; newsapi.ai query terms, count, verdict; the final p*, its rounding, and which value governed under E.5.

E.8 — SOURCES (PUBLIC)

UK Takeover Panel — Transaction Statistics: thetakeoverpanel.org.uk/communications/transaction-statistics · Annual Reports: /communications/reports. European Commission — Merger control statistics & procedures: competition-policy.ec.europa.eu/mergers/procedures_en. Industry / academic deal-failure sources: McKinsey (2019), Done Deal? Why Many Large Transactions Fail to Cross the Finish Line (≈ 10% large-deal cancellation/yr); Schwert (2000), Hostility in Takeovers, J. Finance 55(6):2599–2640; Bebchuk, Coates & Subramanian (2002), The Powerful Antitakeover Force of Staggered Boards, Stanford Law Rev. 54:887–951; Xing & Yi (2024), Mergers and Attributions, J. Management Studies, doi:10.1111/joms.13163.

References

1 Bebchuk, Lucian A.; Coates IV, John C.; and Subramanian, Guhan (2002). "The Powerful Antitakeover Force of Staggered Boards: Theory, Evidence, and Policy." *Stanford Law Review* 54: 887–951.

2 Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78(1): 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

3 Cai, Alice; Arawjo, Ian; and Glassman, Elena L. (2024). "Antagonistic AI." arXiv:2402.07350 [cs.HC]. Submitted 12 February 2024. https://arxiv.org/abs/2402.07350

4 Canepa, Francesco Saverio (2026). *10 errori che distruggono valore nelle operazioni di M&A.* Independently published (Amazon KDP).

5 Good Judgment Inc. (2015). *Good Judgment Open.* Crowdsourced forecasting platform. Launched September 2015; founded by Philip E. Tetlock and Barbara Mellers. https://www.gjopen.com/

6 Hosmer, David W.; and Lemeshow, Stanley (1980). "Goodness of Fit Tests for the Multiple Logistic Regression Model." *Communications in Statistics — Theory and Methods* 9(10): 1043–1069. doi:10.1080/03610928008827941.

7 Kahneman, Daniel (2011). *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux. ISBN 978-0-374-27563-1.

8 McKinsey & Company (2019). "Done Deal? Why Many Large Transactions Fail to Cross the Finish Line." McKinsey Strategy & Corporate Finance. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/done-deal-why-many-large-transactions-fail-to-cross-the-finish-line

9 Metaculus, Inc. (2015). *Metaculus: A Forecasting Platform and Aggregation Engine.* Founded 2015, Santa Cruz, CA. https://www.metaculus.com/

10 Murphy, Allan H. (1973). "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology* 12(4): 595–600. doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

11 Schwenk, Charles R. (1990). "Effects of Devil's Advocacy and Dialectical Inquiry on Decision Making." *Organizational Behavior and Human Decision Processes* 47(1): 161–176. doi:10.1016/0749-5978(90)90037-P.

12 Schwert, G. William (2000). "Hostility in Takeovers: In the Eyes of the Beholder?" *The Journal of Finance* 55(6): 2599–2640. doi:10.1111/0022-1082.00301.

13 Tetlock, Philip E.; and Gardner, Dan (2015). *Superforecasting: The Art and Science of Prediction.* New York: Crown. ISBN 978-1-101-90556-2.

14 Todd, Peter (2016). *OpenTimestamps: Scalable, Trust-Minimized, Distributed Timestamping with Bitcoin.* Project announcement and specification, https://opentimestamps.org/. Announced 15 September 2016.

15 Wu, Haolun; Li, Zhenkun; and Li, Lingyao (2025). "Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning." arXiv:2511.07784. Submitted 11 November 2025. https://arxiv.org/abs/2511.07784

16 Wynn, Andrea; Satija, Harsh; and Hadfield, Gillian (2025). "Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate." arXiv:2509.05396. Submitted 5 September 2025; v2 13 October 2025. Accepted, ICML Multi-Agent Systems Workshop 2025. https://arxiv.org/abs/2509.05396

17 Xing, Zhe (Adele); and Yi, Xiwei (2024). "Mergers and Attributions: An Examination of M&A Terminations in 1996–2022." *Journal of Management Studies.* doi:10.1111/joms.13163.