# Reliability before Validity: Stability and Sensitivity of LLM Risk-Pattern Selection for AI Deployment Risk Assessment

Gareth McConomy[1], Raymond Bond[1], David H. Glass[1], Jun Liu[1], and Holly Toner[2]

[1] School of Computing, Ulster University, Belfast, Northern Ireland, UK
mcconomy-g@ulster.ac.uk
[2] Kainos plc, Belfast, Northern Ireland, UK

**Abstract.** A growing line of work uses large language models (LLMs) to match a free-text description of an AI deployment against a fixed library of named risk patterns, framing risk identification as a case-based-reasoning (CBR) task within AI-governance pipelines. The approach is promising, but its selections must first be reliable, and no operational oracle is available: domain experts frequently disagree, so there is no stable ground truth to recover. We therefore take a reliability-precedes-validity stance and measure reproducibility, stability, and sensitivity rather than correctness. We compare four arms (unstructured, chain-of-thought, holistic CBR, and a deterministic formal CBR scorer) across four LLMs and ten cases, probing each method with meaning-preserving paraphrases and meaning-altering edits, scored by top-10 set-membership overlap with bootstrap CIs; with ten cases this is a characterisation, not a powered test. We find that (i) instability is distributed across the detection and scoring stages, with scoring often the larger source and the split model-dependent; (ii) the unstructured arm is the most reproducible yet the least sensitive, the configuration a stability-only metric would reward; (iii) only the two CBR arms that use the detection labels are stable *and* sensitive, indicating that structure is what buys sensitivity; and (iv) the holistic and formal scorers are a statistical tie in sensitivity at this scale, the formal one adding determinism and auditability but trading stability through the coarseness of its scores. We make no claim about correctness, which is out of scope by design.

**Keywords:** Large language models · Case-based reasoning · AI risk assessment · Reliability · Reproducibility · Stability · Sensitivity.

## 1 Introduction

A growing line of research proposes using large language models (LLMs) to identify the risks of an AI deployment by matching a free-text description of the deployment against a library of named risk patterns, a case-based-reasoning (CBR) task [1–5]. Each pattern is a named, recurring way an AI deployment can

cause harm, defined by a short list of *enabling conditions*: observable features of a deployment that make that risk more likely. As a route to AI governance [6, 7] the approach is promising but not yet established: before such a tool can be relied upon in practice, its selections must be *reliable*. We treat reliability as comprising three measurable properties: *reproducibility* (repeating the same case yields the same selection), *stability* (rewording the case without changing its meaning leaves the selection unchanged), and *sensitivity* (a material, risk-relevant change to the case does change the selection). Reliability is distinct from *accuracy* (whether the selected risks are the right ones), which we do not address. Whether such selections are reliable is not obvious, because LLM outputs shift under semantically irrelevant input changes [12, 16]; establishing it is the question we take up.

What makes this setting unusual is that reliability cannot be replaced by a correctness check. For AI-risk assessment no operational oracle is available: governance-mapping work from the MIT AI Risk Repository [8, 9] reports that LLM risk-coverage judgements frequently disagree with human consensus, and expert reviewers with one another. Where experts do not converge, validity is contested and costly to adjudicate rather than simply recoverable, so we take a reliability-precedes-validity stance: before asking whether a selection is *right*, we ask whether the process is reproducible, stable, and sensitive, the properties measurable without an oracle.

We study four selection *methods*, which we call *arms* (A1–A4), over a fixed library of 86 AI-risk patterns and ten deployment cases, across four LLMs, probing each with meaning-preserving paraphrases (stability) and meaning-altering edits (sensitivity), read on a stability × sensitivity plane with bootstrap CIs. Because two arms share a condition-detection stage (an LLM step that labels each pattern's enabling conditions PRESENT, PARTIAL, or ABSENT) and differ only in their scorer, the design also localises where instability forms. With $n = 10$ cases this is a characterisation, not a powered test.

*Aim, research questions, and objectives.* Our *aim* is to characterise the *reliability* (not the accuracy) of LLM risk-pattern selection for AI deployment, and to localise where any unreliability arises. Three research questions follow. **RQ1**: how reproducible and stable is each method under repetition and meaning-preserving rewording? **RQ2**: is each method sensitive to genuine, risk-relevant changes in the case, or merely stable because it does not react to content? **RQ3**: when a method is unstable, does the instability arise in the condition-detection stage or in the scoring stage? Our objectives are to (i) define expert-free reliability measures that need no ground truth; (ii) compare four methods across four LLMs and ten cases under both perturbation types; and (iii) attribute instability to a pipeline stage.

## 2 Related Work

*LLM inconsistency and paraphrase sensitivity.* That LLM outputs vary under semantically irrelevant input changes (paraphrase, reordering, formatting) is

well documented, from perturbation-based behavioural testing [13] to enterprise perturbation benchmarks [12]. We claim no contribution on this general phenomenon; we take it as our premise and focus on *where* in a structured pipeline the instability forms.

*Stability and sensitivity.* The most closely related work distinguishes an output's stability under meaning-preserving change from its sensitivity to meaning-altering change. Riasat [14] separates stability, correctness, sensitivity, and validity, arguing internal consistency is an insufficient proxy for correctness; Pirouzkhah et al. [15] study semantics-preserving perturbations for code revision; Kostíc et al. [16] show meaning-preserving perturbations destabilise LLM scores and rankings; RobustFlow [17] is closest but targets workflow *generation* and remedies instability by training. We adopt this framing as background: it differs in object (general reasoning, code, workflow generation) and in assuming or training for ground truth, whereas we evaluate case-based risk-pattern selection where no operational oracle is available, characterising reliability rather than training for it.

*CBR with LLMs.* The combination of case-based reasoning [18, 19] with LLMs is an active area [4, 5], but is typically quality-framed, validated against ground truth or human alignment, and does not, to our knowledge, decompose instability by pipeline stage; our use is reliability-oriented, not quality-claiming.

*The no-oracle premise.* Our no-oracle premise follows governance practice: the MIT AI Risk Repository [8, 9] reports that LLM risk-coverage assessments frequently disagree with human consensus, and expert reviewers with one another, so no stable ground truth exists to recover. Validity is not thereby unattainable, merely contested and costly to adjudicate; a multi-expert layer remains future work. The regime is familiar: metamorphic testing addresses exactly this oracle problem, checking relations between outputs under controlled input transformations [10], as our paired probes do; validating LLM generative-agent systems without ground truth is likewise open [11].

*Gap.* We are not aware of prior work at this intersection: CBR risk-pattern selection for AI-deployment governance, in a no-oracle setting, with stage-localised instability. The constituent ideas, cited above, are not ours.

## 3 Method

Figure 1 summarises the design. We measure the *reliability* (reproducibility, stability, and sensitivity) of LLM risk-pattern selection, not its correctness. There is no expert, no answer key, and no recall metric anywhere in the design; validity is out of scope by construction, because no oracle for case–pattern applicability exists even in principle (Sec. 1). Every quantity we report is a property of how the *selection process* behaves under repetition and input perturbation.

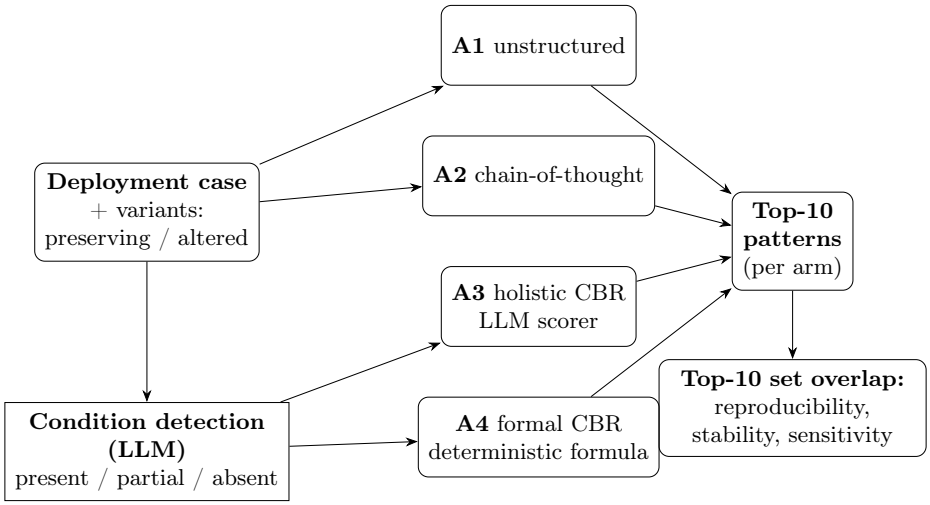



**Fig. 1.** Experimental pipeline. A deployment case, with its meaning-preserving and meaning-altering variants, is processed by four selection arms. A1 and A2 read the full pattern library directly; A3 and A4 instead use a shared condition-detection stage and differ only in their scorer (holistic LLM judgement vs. a deterministic formula), the single-variable contrast in the design. Each arm returns its top-10 patterns; selections are compared by top-10 set-membership overlap to measure reproducibility, stability under rewording, and sensitivity under content change.

### 3.1 Materials: a fixed pattern library and ten fresh cases

All four arms operate over one fixed library of **86** AI-risk patterns (identifiers P2–P87; P1 is absent, which is immaterial). Each pattern carries a short list of *enabling conditions*: between three and five per pattern (mean 3.92), 337 condition slots in total, each a brief textual label such as "Safety-Critical Operational Domain." The enabling conditions are the unit the detection stage classifies and the substrate of the formal scorer. They are graded, compensatory evidence, not individually necessary or jointly sufficient: the formal scorer averages their satisfied weights, so more satisfied conditions raise a pattern's applicability. For example, pattern P3 (Automation-Induced Skill Degradation Under Reduced Practice) carries enabling conditions such as "Routine Tasks Suitable for Automation" and "Skills Are Perishable Without Practice"; the detection stage decides, for a given case, whether each is PRESENT, PARTIAL, or ABSENT. The library is the identical shared input for every arm, model, and perturbation variant; there is no pre-retrieval or candidate shortlist, and each arm ranks over the whole 86-pattern space.

The ten deployment cases are fresh, span three domains (finance ×3, health ×3, public sector ×4), run 809–998 words each, and carry no annotations or answer key. They were authored by a separate LLM (Sec. 3.8) distinct from every LLM under test, so none is evaluated on text it wrote. One finance case, for

**Table 1.** The four arms. Every arm outputs a ranked top-10 of the same 86 patterns.

| Arm | Input | Processing |
|---|---|---|
| A1 unstructured | library + case | one LLM call selects and ranks; no scaffold |
| A2 chain-of-thought | library + case | A1 plus a generic step-by-step instruction |
| A3 informal CBR | labels + case | LLM holistically scores and ranks, weighing labels as evidence; no fixed formula |
| A4 formal CBR | labels only | deterministic similarity function (Sec. 3.4); no further LLM call |

instance, describes a building society extending a self-trained loan-underwriting engine from small-business loans to larger commercial mortgages, with loan officers increasingly reluctant to override it; full input and output traces are in the linked repository.

### 3.2 The shared condition-detection stage

Two of the four arms (A3 and A4) use a shared condition-detection step. For a given case text, the LLM under test (not a separate detector) classifies every enabling condition of every pattern as PRESENT, PARTIAL, or ABSENT, with a short ($\leq 15$ word) quoted evidence span justifying each label. Labels are keyed by the integer pair (`pattern_id`, `condition_index`), which is the only join key; the condition text is copied through for human audit only.

Crucially, detection is **re-run on every repeat and variant**, never cached: its instability is part of what the dependent arms' reliability measures, and fixing it would make the formal arm appear falsely perfect; it is a measured quantity, not noise to remove.

### 3.3 The four arms

Each arm selects and ranks the top $K = 10$ patterns for a case; Table 1 summarises each arm's input and processing. The arms differ in the reasoning scaffold placed between the case and the ranking.

A3 and A4 differ in exactly one variable, the scorer (LLM holistic judgement vs. a deterministic function) applied to identical labels: the only single-variable contrast in the design. A1 and A2 are whole-library baselines; that comparison is not fully controlled, since the structured arms receive an additional detection pass (extra decomposition and computation), so their advantage is the effect of the detection-plus-scoring package, not prompt wording alone (Sec. 6).

### 3.4 The formal similarity function (A4)

For a case and a pattern with enabling conditions of detected statuses $s_1, \ldots, s_N$, the formal score is the normalised mean of status weights

$$\text{sim}(\text{case}, \text{pattern}) \;=\; \frac{1}{N} \sum_{i=1}^{N} w(s_i),$$

where $w(\textsc{present}) = 1.0$, $w(\textsc{partial}) = 0.5$, $w(\textsc{absent}) = 0.0$, and any condition lacking a detected label defaults to ABSENT. Patterns are ranked by descending score, with ties broken by ascending integer `pattern_id`. The function is fully deterministic and emits a per-condition evidence trace, giving A4 a mechanically reconstructible audit trail; it is an ordered weighted averaging (OWA) aggregation [20] at orness 0.5 (neutral), with the OWA generalisation a pluggable seam not exercised here. The coarse status set collapses many scores into ties, so we report tie frequency and lean on the content-sensitivity probe (Sec. 3.5) rather than rank order alone.

### 3.5 Reliability dimensions and the set-overlap metric

All measurement is automated; no human grades any output. We characterise three expert-free dimensions:

1. **Reproducibility** (*same input, same output?*): agreement across three identical repeats of a method on an unchanged case (e.g. ask one model to select for the same case three times and check whether the top-10 patterns stay the same).
2. **Stability** (*reworded input, same output?*): invariance of the selection under meaning-preserving rewordings of the case (same facts, different wording).
3. **Sensitivity** (*materially changed input, changed output?*): the selection *should* move under a genuine, risk-relevant change to the case (e.g. removing human oversight). This guards against *inert stability*: a method that looks consistent only because it barely reacts to anything.

To compare two top-$K$ selections we use a **set-membership overlap** metric: the membership overlap $|A \cap B|/K$ of the two top-10 *sets* and its complement $\Delta = 1 - |A \cap B|/K$. We avoid ordered rank correlation, which would credit A4's arbitrary-but-stable identifier tiebreak as genuine stability; set overlap counts only membership churn.

From this we derive three change quantities per arm and model: *noise* $\Delta$ (run-to-run change on the unchanged original), *preserving* $\Delta$ (paraphrase variant vs. original), and *altered* $\Delta$ (content-altered variant vs. original). We summarise stability as $1 - \text{presv}\Delta$ and sensitivity as $\text{altrd}\Delta - \text{presv}\Delta$ (movement under content change in excess of movement under mere rewording). The desired ordering is $\text{altrd}\Delta \geq \text{presv}\Delta \geq \text{noise}\Delta$. Sensitivity so measured is necessary, not sufficient: it certifies that an arm *responds* to risk-relevant change, not that the response is *appropriate*, which only expert judgement could certify.

### 3.6 Stage decomposition: localising where instability forms

The matched A3/A4 pair localises instability to a stage. A4's scorer is deterministic, so *all* of A4's instability is inherited from detection. Pinning the labels and re-running only the holistic A3 scorer isolates **scoring-stage** instability, so comparing fixed-label against detection-varying A3 partitions A3's instability between the two stages. A **coherence rule** keeps this clean: in repeat $r$ of a (model, case, variant), A3 and A4 use detection repeat $r$; A1 and A2 never touch detection.

### 3.7 Perturbation design

For each case a separate LLM generates six perturbation variants of the case text. **Four are meaning-preserving** (every fact, number, and named entity held fixed, only the wording changed) and drive stability: *stylistic_ rewrite* (different register and word choice), *structural_ reorder* (same content, reordered), *lossless_ summary* ($\approx 25\%$ shorter, every fact kept), and *detail_ elaboration* ($\approx 25\%$ longer, no new facts). **Two are content-altered** and drive the sensitivity probe: *altered_ 1* makes one autonomy change (e.g. removing a human override) and *altered_ 2* one change to domain, affected population, or stakes, so a governance assessment *should* reach a different conclusion; the model states the change explicitly.

### 3.8 Models, parameters, and run protocol

Four co-equal LLMs are evaluated under test: `gpt-4o`, `gemini-2.5-pro`, `grok-4`, and `deepseek-chat` (a non-reasoning model). A fifth, separate LLM (`claude-sonnet-4-6`) authors the cases and generates all perturbations and is never placed under test. The LLM arms run at deployment temperature 0.3; $K = 10$.

We collect three identical repeats of the unchanged original per (model, arm), two repeats on each preserving variant, and three repeats on each altered variant, across all 10 cases and 4 models, every call logged. Confidence intervals come from a **nonparametric bootstrap that resamples the 10 cases** [21] ($B = 2000$, 95% percentile); we report bootstrap CIs rather than null-hypothesis tests.

## 4 Results

All reliability quantities below are top-10 set-membership overlaps over the full run of 2,720 arm units (480 originals plus 2,240 perturbation units; no failed units). Throughout we report *relative*, not absolute, reproducibility: even on unchanged text and the most stable arm, roughly a fifth of the top-10 churns from one run to the next.

**Table 2.** Reproducibility decomposition: mean top-10 set overlap across three identical repeats, by model. "A4 (det-var)" is the deterministic formal arm (all its variance inherited from detection); "A3 vary" lets detection move, "A3 fixed" pins detection to one repeat's labels. The final column is the detection stage's estimated share of A3's total top-10 instability.

| model | A1 | A2 | A4 (det-var) | A3 vary | A3 fixed | detect. share |
|---|---|---|---|---|---|---|
| gpt-4o | .783 | .643 | .750 | .683 | .813 | 41% |
| gemini | .823 | .787 | .707 | .747 | .830 | 33% |
| grok | .800 | .817 | .710 | .740 | .887 | 56% |
| deepseek | .783 | .660 | .810 | .747 | .887 | 55% |

### 4.1 Detection-stage reproducibility

Before any scorer is applied, the shared detection stage is itself only partially reproducible. Across three identical repeats, 80.0–85.9% of (case, pattern, condition) labels are identical all three times (gemini 80.0, gpt-4o 80.8, grok 81.4, deepseek 85.9); most movement is PARTIAL-boundary wobble, with hard PRESENT↔ABSENT flips rarer and concentrated in gemini (7.0% vs. 1.4–3.4% elsewhere). Across models only 59.9% of cells are unanimous (4/4). So the detection stage is itself a real instability source, present before either CBR scorer runs and measured by the stage decomposition rather than discarded as noise.

### 4.2 Where instability forms: reproducibility decomposition

Pinning the detection labels and re-running only the holistic A3 scorer lifts A3's top-10 reproducibility for every LLM (Table 2, "A3 fixed" > "A3 vary"), confirming detection is a genuine but not dominant instability source. It contributes an estimated 33–56% of A3's top-10 instability and the holistic scorer the rest; even with labels frozen, A3 reproducibility reaches only 0.81–0.89. The split is model-dependent: instability is distributed across both stages, not localised to either.

### 4.3 Stability and sensitivity across the four arms

Table 3 reports the three change quantities per arm (model-averaged) and Fig. 2 plots all 16 (arm, model) points with bootstrap CIs. Two arms satisfy the desired ordering altrd$\Delta \geq$ presv$\Delta \geq$ noise$\Delta$: A3 and A4, the label-based arms (those that score patterns from the detection labels). A1 is stable but barely moves under content; A2 inverts the ordering, the signature of degenerate stability.

Sensitivity is a *paired* quantity: the per-case difference altrd$\Delta$ − presv$\Delta$, read from its own bootstrap interval, not inferred from whether the marginal CIs in Table 3 overlap. The four-arm model-averaged paired sensitivity headline (bootstrap 95% CI, resampling the 10 cases) is: A1 +1.2 pts $[+0.2, +2.2]$; A2

**Table 3.** Per-arm change quantities, **averaged across the four LLMs** (top-10 set-membership change $\Delta$ in %, with bootstrap 95% CIs resampling the 10 cases, $B = 2000$): noise$\Delta$ (run-to-run on the original), presv$\Delta$ (paraphrase vs. original), altrd$\Delta$ (content-altered vs. original). Stability margin = presv$\Delta$−noise$\Delta$ (smaller is more stable); sensitivity margin = altrd$\Delta$−presv$\Delta$ (larger is more sensitive). These are model-averaged figures and mask per-model structure: in particular A3 and A4 are null on grok; see Fig. 2 for the per-model resolution.

| arm | noise$\Delta$ [95%CI] | presv$\Delta$ [95%CI] | altrd$\Delta$ [95%CI] | stab. margin | sens. margin |
|---|---|---|---|---|---|
| A1 | 20.2 [18.6,21.8] | 20.0 [18.1,21.9] | 21.2 [19.0,23.2] | −0.2 | +1.2 |
| A2 | 27.3 [23.2,31.1] | 27.9 [25.1,30.7] | 27.6 [24.6,30.5] | +0.6 | −0.3 |
| A3 | 27.1 [23.9,29.9] | 27.1 [24.3,29.6] | 31.9 [28.8,34.9] | +0.0 | +4.8 |
| A4 | 25.6 [22.6,28.6] | 28.6 [26.5,30.7] | 33.2 [30.7,35.4] | +3.0 | +4.6 |

$-0.3$ pts $[-1.6, +1.1]$; A3 $+4.8$ pts $[+3.5, +6.3]$; A4 $+4.6$ pts $[+2.2, +6.7]$. A3 and A4 exclude 0 and are roughly four times A1's; A1 clears 0 only slightly; A2 straddles 0.

Per model, only A3 and A4 have sensitivity CIs that exclude 0, and only on the three non-grok models, where their sensitivity runs roughly +4 to +6 points (Fig. 2); on grok both straddle 0 (+2.3% and +2.6%), so weak sensitivity here is an LLM property, not an arm property. Every A1 and A2 cell straddles 0. These results *indicate* that structure, which uses the condition labels, is what buys sensitivity, while the unstructured and CoT baselines do not reliably move under content change.

### 4.4 The unstructured baseline is stable but minimally sensitive

A1 is the most stable arm (model-averaged preserving overlap ∼80%) yet only minimally sensitive to content. Its model-averaged per-case sensitivity is +1.2 pts, bootstrap 95% CI $[+0.2, +2.2]$: the interval clears 0, so A1 is not strictly inert, but the effect is barely above zero and roughly a quarter of the ∼5 pt sensitivity of the structured arms. A stability-only metric would have ranked A1 best; the sensitivity probe *characterises* it instead as stable but only weakly content-sensitive: the stability trap.

### 4.5 A4's stability gap is genuine score coarseness, not a tiebreak artefact

A4 is the least stable arm under rewording (stability margin +3.0 pts, Table 3). Its coarse three-level scores produce heavy ties: the rank-10 cutoff falls inside a group of patterns sharing one score in 680/680 (100%) of runs, so which tied patterns fill the final slots is settled by the identifier tiebreak. The extra churn is not mere tiebreak shuffling: 75.4% of preserving-variant swaps are genuine own-score movement (the pattern's own score changed under rewording), and the tie-cut share matches rerun noise (24.6% vs. 25.6%); it is paraphrase moving

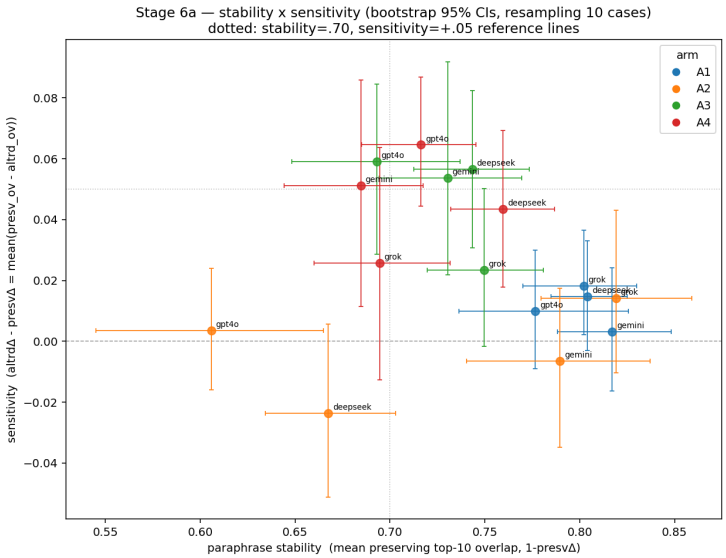



**Fig. 2.** Stability × sensitivity across the 16 (arm, model) points. The $x$-axis is stability (mean overlap under meaning-preserving rewording); the $y$-axis is sensitivity (altrd$\Delta$−presv$\Delta$); bars are bootstrap 95% CIs (resampling the 10 cases). Only the label-based arms A3 and A4 have sensitivity CIs clear of 0, and only on gpt-4o, gemini, and deepseek; grok is the weak-sensitivity model for both A3 and A4.

coarse scores across boundary tiers, coarseness-brittleness of the formal scorer, not a tiebreak artefact. Per LLM, the mean number of the 10 top slots fixed by score rather than the identifier tiebreak is 8.0 on grok and 7.1 on gpt-4o, but only 1.4 on gemini, whose PRESENT-heavy detection saturates the score so that the top-10 is mostly identifier order.

### 4.6 Holistic vs. formal scorer: a statistical tie in sensitivity

The matched A3/A4 pair (identical labels, scorer the only difference) lets us ask whether the holistic or the formal scorer is more content-sensitive. The per-model bootstrap CI on the sensitivity difference (A4−A3) straddles 0 on every model (point estimates −1.3% to +0.6%), so at $n = 10$ the two scorers are statistically indistinguishable in sensitivity. We therefore do not crown a scorer: both are sensitive; the formal one adds determinism and a mechanical audit trail but trades stability via its coarseness (§4.5). A qualitative check (unscored) confirms face validity: across the 20 altered variants, the patterns entering and leaving the top-10 track the stated change.

## 5 Discussion

Stated directly: explicit condition detection makes selection responsive to substantive change; the deterministic scorer buys auditability, not sensitivity, and cedes stability through coarse scores and ties.

*Stability alone is a trap.* A reliability metric built on reproducibility alone would have been actively misleading here: A1, the unstructured baseline, is the most stable arm yet barely moves under genuine content change (§4.4), so a stability-only evaluation would have crowned the arm that engages least with the case. Any reliability assessment therefore needs a sensitivity probe alongside its stability measure; without one, the inert configuration scores best.

*Structure buys sensitivity.* Only the two arms that use the shared condition-detection labels (A3, A4) reach the stable-and-sensitive region, on three of the four LLMs (§4.3); the unstructured and chain-of-thought baselines, which see the same library and case but no detection step, do not reliably move under content change: the condition-detection scaffold, not the extra CoT instruction, forces engagement with case content. We do not claim the structured arms select *better* patterns, only that they respond to content where the unstructured arm is almost inert.

*Where to invest: detection, not the scoring mathematics.* The stage decomposition (§4.2) spreads instability across both stages, the holistic scorer often the larger source; yet replacing it with a deterministic function does *not* improve sensitivity (the A3↔A4 tie, §4.6) and adds tie-brittleness (§4.5). Effort is better spent stabilising the shared condition-detection input than elaborating the scoring mathematics, which buys auditability but neither sensitivity nor stability.

*Model-dependence: no universal best pairing.* No single (arm, model) pairing dominates: the label-based arms are sensitive on gpt-4o, gemini, and deepseek but null on grok (weak sensitivity is a property of the LLM, not the arm), while gemini's detection saturates the formal scorer into tie-degeneracy; model choice is not separable from mechanism.

*Practitioner implications.* In practice: prefer the structured, label-based pipeline over freehand prompting when sensitivity matters; treat the top-10 as a *shortlist for human attention*, not a ranking (membership churns ∼20–33%, and within-tie order is a tiebreak artefact); use the output as a thinking aid, not an oracle; and check sensitivity to content change, not just run-to-run reproducibility.

## 6 Limitations

Several limitations bound these results. With $n = 10$ cases this is a *characterisation, not a powered test*: the bootstrap intervals are wide and per-cell (arm

× model) results are indicative, not conclusive. There is no oracle by design: we measure reliability, not validity (no ground truth, expert panel, or recall metric enters anywhere), so a stable or sensitive arm is not thereby accurate. A single fixed 86-pattern library and one governance framing bound transfer to other libraries, domains, or framings. Reported reproducibility is relative: "stable" means "no worse than the noisy baseline." A single separate LLM authored both the cases and the perturbations, bounding ecological validity; though disjoint from every LLM under test, an independent or human-authored source would strengthen the design, in particular human checks that preserving rewrites preserve and altering edits alter only what is intended. The arm comparison is not fully controlled: the structured arms receive an extra detection pass, so their advantage reflects that whole package, not scaffold wording alone. Finally, sensitivity rests on a sparse two-point gradient per case (one autonomy change, one domain/population/stakes change) against a three-repeat noise floor: enough to characterise whether an arm responds to content, but not to map the response surface densely.

## 7 Conclusion and Future Work

We brought the stability-versus-sensitivity lens to CBR AI-risk-pattern selection where no operational oracle is available, and added a stage-localisation method attributing pipeline instability to detection or scoring. Across four arms, four LLMs, and ten cases: reproducibility alone is misleading, only the label-based CBR arms are both stable and sensitive, instability is distributed across detection and scoring, and the two scorers tie in sensitivity, the formal one buying auditability at the cost of stability. None of this speaks to correctness, which remains out of scope.

Future work has two tracks. First, the deferred *fusion* direction (combining the holistic and formal scorers, or structured and unstructured arms) has a prerequisite: a fuller sensitivity response surface than the present two-point contrast, e.g. by Monte Carlo sampling of the perturbation space; even a dense surface shows that a method *is* sensitive, not that its sensitivity is *appropriate*. Second, validation at scale: more cases, an independent or human-authored case and perturbation source, and, once validity is in scope, a multi-expert study relating reliability to correctness.

**Acknowledgments.** This work was undertaken as part of a PhD funded by Kainos plc.

**Disclosure of Interests.** The authors have no competing interests to declare. For completeness: this work forms part of a PhD funded by Kainos plc, and one author (H. Toner) is employed by Kainos plc.

**Data and code availability.** Pattern library, cases, perturbations, run logs, and analysis code: `https://github.com/ggmconomy/Ai_stability_testing`.

**Declaration on Generative AI.** The authors used Claude (Anthropic) for literature synthesis during knowledge-graph construction, manuscript drafting and language editing, and case-study development; generative AI was excluded from the experiments themselves.

## References


1. McConomy, G., Bond, R., Glass, D., Liu, J., Toner, H.: Augmenting large language models with causal risk patterns for AI deployment risk assessment. In: Proc. IEEE ICAD 2026 (to appear)
2. McConomy, G., Bond, R., Glass, D., Liu, J., Toner, H.: Bridging causal reasoning and fuzzy calibration for trustworthy AI risk assessment. In: Proc. FLINS-ISKE 2026 (to appear)
3. McConomy, G., Bond, R., Glass, D., Liu, J., Toner, H.: Layer-appropriate augmentation for human-AI risk assessment: integrating case-based reasoning with fuzzy calibration. In: Proc. HHAI 2026 (to appear)
4. Gan, W., Dao, M.-S., Zettsu, K.: Case-based reasoning augmented large language model framework for decision making in realistic safety-critical driving scenarios. arXiv:2506.20531 (2025)
5. Bach, K., et al.: Case-based reasoning meets large language models: a research manifesto for open challenges and research directions. Working paper, hal-05006761 (2025)
6. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023)
7. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union (2024)
8. MIT AI Risk Initiative: Mapping the AI governance landscape: April 2026 update. airisk.mit.edu (2026)
9. Slattery, P., Saeri, A.K., Grundy, E.A.C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., Thompson, N.: The AI risk repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. arXiv:2408.12622 (2024)
10. Xie, X., Ho, J.W.K., Murphy, C., Kaiser, G., Xu, B., Chen, T.Y.: Testing and validating machine learning classifiers by metamorphic testing. Journal of Systems and Software 84(4), 544–558 (2011)
11. Adornetto, C., Mora, A., Hu, K., Izquierdo Garcia, L., Atchade-Adelomou, P., Greco, G., Alonso Pastor, L.A., Larson, K.: Generative agents in agent-based modeling: overview, validation, and emerging challenges. IEEE Transactions on Artificial Intelligence 6(12), 3165–3183 (2025)
12. Bogavelli, T., Bamgbose, O., Gauthier Melançon, G., Riols, F., et al.: Evaluating robustness of large language models in enterprise applications: benchmarks for perturbation consistency across formats and languages. arXiv:2601.06341 (2026)
13. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: behavioral testing of NLP models with CheckList. In: Proc. 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902–4912 (2020)
14. Riasat, N.: When stability fails: hidden failure modes of LLMs in data-constrained scientific decision-making. ICLR 2026 Workshop (I Can't Believe It's Not Better). arXiv:2603.15840 (2026)
15. Pirouzkhah, S., Serbout, S., Bacchelli, A.: Consistent or sensitive? Automated code revision tools against semantics-preserving perturbations. arXiv:2602.14595 (2026)
16. Kostić, B., Fallon, C., Risch, J., Löser, A.: Same meaning, different scores: lexical and syntactic sensitivity in LLM evaluation. arXiv:2602.17316 (2026)
17. Xu, S., et al.: RobustFlow: towards robust agentic workflow generation. arXiv:2509.21834 (2025)

18. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications 7(1), 39–59 (1994)
19. López de Mántaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M.L., Cox, M.T., Forbus, K., Keane, M., Aamodt, A., Watson, I.: Retrieval, reuse, revision and retention in case-based reasoning. The Knowledge Engineering Review 20(3), 215–240 (2005)
20. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics 18(1), 183–190 (1988)
21. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)