Grounding Before Collapse

Formulation Gaps, Verifier Gates, and the Limits of Self-Contained Judgment

Flyxion
Independent Researcher

September 2026

Abstract

Two recent papers, one empirical, one formal, independently name a failure mode this corpus has previously described only qualitatively: a system commits to an output before the informational state it occupies actually justifies commitment. Ge et al. document this as premature formulation in language-model agents performing operations-research modeling and propose a persistent gap ledger, Dynamic Gap Search, as a partial remedy [2]. Chen et al. prove, for a constructed class of environments, that a judge restricted to its own generated text cannot reliably license the kind of commitment at stake, regardless of how sophisticated that judge becomes, unless a signal from outside the text channel is supplied [1]. Read through this corpus's own admissibility vocabulary, both papers describe instances of collapse, the resolution of an admissible family of continuations into one realized output, firing without the grounding that would license it. Collapse itself is not the defect; unwarranted collapse is. This essay states each formal and empirical claim at the strength its source actually supports, distinguishes several kinds of missingness a two-state gap ledger cannot represent, works a concrete ambiguous-pair example, separates the informational necessity of grounding from the further conditions needed for exact convergence, and closes with the proposition the formal result actually licenses together with an explicit caution that grounding is a solution to an information-channel problem, not a guarantee of correctness on its own.

1 Introduction

A system commits to an output before its informational state actually justifies commitment. Ge et al. name this premature formulation and document it in language-model agents tasked with interpreting an underspecified operations-research request: a plausible-sounding formulation is produced and declared ready before the space of task-relevant alternatives has actually narrowed to one [2]. Chen et al. name a structurally related failure at a different level of a system: a self-contained gate, a judge that evaluates only the text it and its counterpart have generated, is asked to decide whether to commit an update to a persistent memory, and the paper proves that such a gate can be made to fail this decision reliably, not merely occasionally, whenever the update's correctness depends on a fact the text channel does not carry [1].

This corpus has a name for the general operation both failures involve. Collapse is the resolution of an admissible family of continuations, $\Gamma(A_t)$, into one realized output, $\gamma^*_{[t,t+\Delta t]}$, and the recurring diagnostic question this corpus has asked of any system that performs collapse is whether the resolution was licensed by evidence adequate to the commitment being made, or merely by the appearance of adequacy. Read in this vocabulary, Ge et al.'s premature formulation is collapse firing while $\Gamma(A_t)$ still contains more than one task-relevant class. Chen et al.'s gate failure is collapse,

in this case the acceptance of a memory update, firing on a signal that cannot discriminate the environments whose difference is exactly what the commitment needs to track. Both are failures to warrant collapse, and reading them together against a single vocabulary is more informative than treating them as unrelated findings from unrelated subfields.

That reading needs to be stated at exactly the strength the two papers support, no more, and it needs one clarification before anything else: collapse itself is not the failure. The sections that follow work through both papers' actual definitions, theorems, and reported figures, correcting several places where an appealing extension of what they say would outrun what they actually establish, and adding a small amount of new formal apparatus, an internal-evidence/external-grounding distinction, a two-dimensional horizontal/vertical admissibility test, where doing so sharpens the source papers' own results rather than substituting for them.

2 Collapse is not failure

Every completed formulation and every accepted memory update requires some reduction from alternatives to a single commitment. If collapse simply named that reduction, it would be a term for something systems must do constantly and successfully in order to function at all, and treating it as inherently suspect would make the vocabulary useless for distinguishing the cases this essay is actually interested in. The operation

$$\text{COLLAPSE}(\mathcal{A}_t) \to a_t$$

is legitimate exactly when one of two conditions holds: either the distinctions remaining among the members of $\mathcal{A}_t$ are irrelevant to the task at hand, so that committing to any one of them serves equally well, or a discriminating warrant, evidence capable of telling the admissible alternatives apart with respect to the property that actually matters, selects a_t specifically. Failure is collapse that satisfies neither condition: a reduction performed while task-relevant alternatives remain both live and undistinguished.

Both papers examined here document a specific instance of this failure rather than an indictment of collapse as such. Ge et al. study collapse in the formulation stage, before task-relevant alternatives have been reduced to one equivalence class; their concern is squarely with cases where the first legitimating condition, irrelevance, does not yet hold and no discriminating warrant has been sought [2]. Chen et al. study collapse in an acceptance gate, where task-relevant alternatives, whether a proposal is corrective or harmful, remain live but the gate lacks any signal capable of telling them apart, so that the second legitimating condition, a discriminating warrant, cannot be met no matter how the gate is built [1]. Keeping this distinction in view throughout prevents collapse from becoming a synonym for error and keeps the essay's actual target, unwarranted collapse, precisely in focus.

3 Formulation completeness as task-relative equivalence

Ge et al. define formulation completeness at a decision point P_t as the condition that every plausible completion c drawn from the bounded space of completions $\mathcal{B}(P_t)$ induces the same formulation structure,

$$|\{\phi(P_t, c) : c \in \mathcal{B}(P_t)\}| = 1,$$

their equations 3.1 and 3.2 taken together [2]. This is, exactly, an equivalence-class criterion, and it is worth making the equivalence relation explicit rather than leaving it implicit in the cardinality

condition, since doing so is what lets the comparison with this corpus's own machinery be drawn cleanly rather than suggestively. Define

$$c_1 \sim_{P_t} c_2 \iff \phi(P_t, c_1) = \phi(P_t, c_2).$$

Formulation completeness is then the statement that $\mathcal{B}(P_t)$ collapses under $\sim_{P_t}$ to a single class, $|\mathcal{B}(P_t)/\sim_{P_t}| = 1$. This is the same general shape as the equivalence-class criterion this corpus's recovery formalism uses to define task-relative sameness, applied here at the coarsest possible task, the task of having interpreted the request at all rather than any more specific downstream task.

Describing this condition as a "trivial holonomy" case, as an earlier pass at this comparison did, is a genuinely useful interpretive move, since it places formulation completeness inside the same family of concepts as this corpus's holonomy-based account of recovery. But this needs to be marked plainly as this essay's extension of the comparison, not as terminology or apparatus Ge et al. themselves supply; the paper states and uses the cardinality condition above and does not frame its own criterion in holonomy language at all. With the equivalence relation made explicit, the exact correspondence holds only after the completion space $\mathcal{B}(P_t)$, the task projection $\phi(P_t, \cdot)$, and the induced equivalence $\sim_{P_t}$ have all been fixed; described more cautiously and more accurately, formulation completeness is the coarsest discrete analogue of trivial holonomy, a binary special case rather than an unqualified instance.

The comparison with this corpus's graded recovery formalism can then be drawn cleanly. Ge et al.'s criterion is binary: either $|\mathcal{B}(P_t)/\sim_{P_t}| = 1$ and the formulation counts as complete, or it does not and there is, on the paper's own terms, no further structure to say about how incomplete it is, no distance between classes, no measure of how much recovery work remains. This is best read as a deliberate operationalization suited to the paper's own goal, a yes-or-no readiness decision, rather than as a defect; a binary criterion is exactly what a stopping rule needs. This corpus's graded recovery formalism extends the criterion by asking a different, complementary question, namely how far apart non-equivalent completions remain from recoverable agreement, which is additional structure Ge et al. never claimed to build and should not be faulted for lacking.

4 Three kinds of missingness, and why asked is not resolved

Before turning to Ge et al.'s proposed remedy, it is worth distinguishing three ways a formulation-critical gap can be missing from a system's represented state, since the remedy's actual coverage is easiest to assess against this distinction rather than against a single undifferentiated notion of "gap."

$$\begin{aligned} \textsc{Open}(g): &\quad g \text{ is represented but unanswered,} \\ \textsc{Refused}(g, c): &\quad c \text{ was considered and rejected under a rule,} \\ \textsc{Unsurfaced}(g): &\quad g \text{ never entered the system's represented state at all.} \end{aligned}$$

Dynamic Gap Search, the first stage of Ge et al.'s proposed framework InterOPT, maintains a persistent ledger of exactly the first of these three: a newly detected gap enters the ledger with status Open, and the currently bindable gaps are formally

$$O_t = \{\ell \in \mathcal{M}_t : \text{status}(\ell) = \textsc{Open}\}$$

in the paper's own notation [2]. Once a question addressing an open gap has been posed, the paper states plainly that its bound entry moves from Open to Asked, "marking that the gap was queried, not resolved," in close paraphrase of the paper's own language. Asked is therefore a record that

another conversational turn occurred, not a record of what that turn accomplished; a user may misunderstand the question, answer only partially, introduce a fresh ambiguity in the course of answering, or reject the question's presupposition entirely, and the ledger's two-state system has no way to represent any of these outcomes differently from a cleanly resolved answer. A more adequate transition system, one this corpus's own bookkeeping would require, would extend the second arrow into several distinguishable outcomes,

$$\textsc{Open} \to \textsc{Asked} \to \begin{cases} \textsc{Resolved}, \\ \textsc{PartiallyResolved}, \\ \textsc{Refused}, \\ \textsc{Superseded}, \end{cases}$$

exposing the gap between interaction history, that a turn happened, and epistemic progress, that the admissible formulation family actually narrowed as a result.

Neither of the ledger's two real statuses is equivalent to this corpus's own REFUSE bookkeeping, in which a candidate x remains in the space of possibilities $P(\mathcal{S})$ while being excluded from the currently admissible set via a positive counterfactual signal, $\rho_r(x, t) > \varepsilon$. Open carries no such signal at all, since an open gap has not yet been evaluated in any sense; Asked carries only a record that a query was attempted, with no signal one way or the other about its outcome. Ge et al.'s ledger, moreover, cannot represent Refused or Unsurfaced as separate outcomes, and it cannot record a gap it never discovered by definition, since Unsurfaced gaps are, by construction, invisible to the mechanism that populates the ledger. Both limitations the paper itself is explicit about: an empty open set,

$$O_t = \varnothing \;\not\Rightarrow\; P_t \text{ is formulation-complete},$$

does not establish that the specification is actually complete, and the paper's own architecture is explicit that Stage 2 never blocks a readiness declaration merely because open entries remain, precisely because Stage 1 may have missed a gap in the first place. Naming this gap between what the ledger tracks and what a fuller REFUSE-adequate ledger would need to track is more informative than any direct equivalence between the two, since it identifies exactly the missing states, and the missing detection guarantee, that would need to be added for the ledger to do the bookkeeping this corpus's own admissibility apparatus requires.

5 Premature collapse, quantified

Ge et al. report that in the open, free-form setting, InterOPT, the paper's own proposed two-stage framework, declares readiness in 98.8 percent of 500 runs, one hundred cases at five repeated runs each, while the paper's stopping audit flags 203 of those runs, 40.6 percent, as premature [2]. The same open-setting condition records an average of 0.560 silent assumptions per run, the paper's term for a gap that was never surfaced as a question at all and was instead resolved implicitly by the system's own default completion. It is worth being precise about what these figures characterize: they describe InterOPT's own behavior in the open, free-form protocol specifically, not a baseline the paper's proposed method improves upon. The paper's own remedy for premature formulation still declares readiness prematurely on this evidence in a substantial share of runs, which is a stronger and more useful fact than a demonstration that some unremedied baseline behaves this way would have been.

These figures are a genuine and valuable quantitative benchmark for premature formulation within OR-Clarify's evaluation setting specifically, one hundred operations-research clarification

cases, a fixed masking procedure, a fixed tested model, a fixed judge. They should not be read more broadly than that. It would overstate the paper's contribution to treat 40.6 percent, or any function of it, as a general base rate for premature collapse across agent architectures, task types, or formulation domains; the paper does not claim this and this essay should not claim it on the paper's behalf. What the figures do supply, and what earlier, purely qualitative treatments of this failure mode in this corpus lacked, is a demonstration that premature formulation is not a rare or contrived pathology confined to weak or unremedied systems; even a framework purpose-built to track formulation gaps still declares readiness prematurely close to two-fifths of the time in this setting, with the accompanying silent-assumption count indicating that a meaningful share of the relevant gaps are never even raised as questions before that declaration is made.

6 Readiness as a decision, not a feeling

The 98.8 percent readiness rate is worth isolating as its own point rather than folding directly into the premature-collapse discussion, because it exposes a distinction the raw prematurity figure does not make explicit on its own. Define

$$R_t = \mathbf{1}[\text{agent declares } \texttt{READY\_TO\_MODEL}]$$

and

$$C_t = \mathbf{1}[P_t \text{ is formulation-complete}]$$

using Ge et al.'s own completeness criterion from Section 3 above. A readiness declaration, $R_t = 1$, is a behavioral output the system produces; formulation completeness, $C_t = 1$, is a fact about whether the admissible formulation space has actually narrowed to one equivalence class. The empirical problem the paper's figures document is precisely that R_t and C_t diverge, and diverge substantially: the system declares $R_t = 1$ in 98.8 percent of runs while the audit's own assessment of something close to C_t agrees in only a bare majority of them. Fluency, conversational closure, and the simple fact that a plausible-sounding formulation has been produced can all raise R_t without changing the equivalence classes C_t is actually tracking, and a readiness gate should accordingly be evaluated by its calibration against formulation completeness, how often R_t and C_t actually agree, rather than by how decisive or confident its declaration sounds. This is a general point about any system that emits a stopping decision as a learned or generated behavior rather than as the output of a check against an independent criterion, and it recurs, in a different guise, in the discussion of grounded gates below.

7 Internal evidence and external grounding

Before turning to Chen et al.'s impossibility theorem, the distinction the theorem trades on needs to be defined carefully enough that "external" does not come to mean, misleadingly, "outside the computer." A deterministic proof checker, a test harness, a simulator, a repository's current state, or a realized reward signal can all be computed within the very same machine running the system under evaluation while remaining external, in the sense that matters here, to the sigma-algebra generated by the system's own text.

Let Z_t denote the generated textual evidence available up to time t, and let E denote the task environment, whatever fact or state the correctness of a proposed commitment actually depends on. A self-contained gate has the general form

$$G_t = G(Z_{\leq t}, \xi),$$

where ξ is the gate's own internal randomness; its output is a function of the text and nothing else. A grounded gate additionally observes some signal

$$Y_t = h(E, a_t),$$

a function of the actual environment and the action or proposal under consideration, whose distribution can differ between two environments that induce exactly the same textual evidence $Z_{\leq t}$. Grounding, on this definition, is not a matter of physical location, computational substrate, or even which process performs the evaluation; it is a matter of whether the evaluating signal's distribution is sensitive to a fact the text channel does not carry. This is the precise sense in which Chen et al.'s "environment-dependent signal" should be understood, and it is the sense this essay uses throughout what follows [1].

8 The impossibility of self-contained gates

Chen et al.'s Theorem 4, titled Self-Gating Impossibility in the paper, is the central formal result and deserves to be walked through in full [1]. The construction fixes disjoint proposal classes $C_0, C_1 \subset C$ with $P(C_0 \mid m) = P(C_1 \mid m) = \mu \in (0, \frac{1}{2}]$ for every reachable memory m, all other proposals treated as inert, and fixes $\kappa \in (0, 1)$ together with two error-update functions, $f_0(v) = (1 - \kappa)v$ and $f_1(v) = \kappa + (1 - \kappa)v$. In environment ε^+, an accepted proposal from class C_a applies f_a to the current error; in environment ε^-, the roles of C_0 and C_1 are exactly swapped, so that the class which is corrective in one environment is harmful in the other, while the text-generation law and the proposal kernel are held identical across both environments.

A worked instance makes the construction concrete. Suppose two candidate memory proposals both contain the identical textual recommendation, "increase the resource cap." In environment ε^+, the hidden cap is currently set too low, so accepting the proposal reduces error. In environment ε^-, the same nominal cap is already binding for an unrelated safety reason, so accepting the identical proposal increases error. Because the two environments produce the same transcript by construction, a text-only gate has no basis for treating the proposal differently across them; a simulator, an execution probe, or any other signal of the form $Y_t = h(E, a_t)$ from the previous section would reveal which environment obtains, but nothing in $Z_{\leq t}$ does.

A self-contained gate, in Chen et al.'s Definition 1, is any possibly randomized, history-dependent acceptance rule measurable with respect to the generated text process and its own internal randomness only; this class includes text-only LLM-as-judge systems of arbitrary sophistication. Because the two constructed environments are, by design, indistinguishable at the level of text and internal randomness, any such gate's accepted class-label sequence can be coupled identically across both environments: the gate cannot behave differently in the two environments because it has no information that differs between them. The paper's proof sketch couples both environments with shared proposal and gate randomness, so that the accepted class-label sequence is identical under ε^+ and ε^-, and a reflection identity between f_0 and f_1 then yields the stated bound. The theorem's conclusion is

$$\max\{e_T^{\varepsilon^+}, e_T^{\varepsilon^-}\} \geq e_0$$

for every self-contained gate and every horizon T, with strict inequality whenever $e_0 < \frac{1}{2}$ and the gate accepts at least one proposal from $C_0 \cup C_1$ with positive probability by time T. By contrast, the paper shows that a free-form rule that accepts everything drifts to $e_T^{\varepsilon} \to \frac{1}{2}$ in both environments, while a grounded gate that directly observes the true error V^{ε} accepts only the corrective class and satisfies $e_T^{\varepsilon} = e_0(1 - \kappa\mu)^T \to 0$ in both. No increase in the gate's text-processing capacity changes

the impossibility side of this picture, since the obstruction is informational rather than computational: the text contains no signal capable of breaking the coupling between the two environments, so a more capable judge applied to the same channel gains nothing.

The theorem's scope needs to be stated with equal care, since an unqualified summary is precisely where the result is most likely to be inflated. Theorem 4 is a minimax result over the constructed class of environment pairs the proof builds; it does not prove that introspection is useless in general, that every correct decision requires external tool use, or that textual proof checking is impossible. The paper's own Remark 1 explicitly and correctly notes that textual self-evaluation remains useful when the transcript itself certifies correctness, a fully checkable mathematical proof being the clearest case, since the discriminating fact is then, by construction, already present in $Z_{\leq t}$ and no external signal is required to recover it. The proper boundary the theorem draws is

$$\begin{aligned}&\text{same observable evidence} + \text{different correct commitments}\\ &\quad\Rightarrow \text{no uniformly reliable evidence-restricted gate;}\end{aligned}$$

when the evidence itself contains a complete certificate, the two environments a proof of this shape would need to construct are no longer relevantly indistinguishable, and the theorem simply does not apply. What Theorem 4 actually proves, at exactly this strength, is that when a proposal's correctness depends on a fact about the environment the text channel does not carry, no amount of model capacity applied to the text can substitute for a signal that actually depends on that fact.

9 Necessity, acceptance, and convergence

Chen et al.'s Section 3 establishes three logically separate levels that an unqualified reading risks compressing into one, and they are worth stating as a chain,

$$\begin{aligned}\text{environment-separating information} &\rightarrow \text{strict verifier-gated acceptance}\\ &\rightarrow \text{convergence under additional assumptions,}\end{aligned}$$

with a distinct piece of machinery doing the work at each link [1].

The first link is Theorem 4 itself, already stated in full above: over the constructed class of ambiguous-pair environments, an environment-separating signal is informationally necessary for any gate to avoid the stated error floor, and this necessity holds regardless of the gate's text-processing capacity.

The second link is Definition 3, Chen et al.'s specific implementation choice, Stochastic Reflective Memory Ascent (SRMA), which commits a candidate memory update only when a fixed, deterministic evaluation protocol certifies a strict decrease in verifier risk,

$$R_i(\tilde{m}^e_{t+1}) < R_i(m^e_t).$$

This rule is one way of satisfying the requirement Theorem 4 establishes, and it is structurally the same shape as this corpus's REFUSE criterion, $\rho_r(x,t) > \varepsilon$, applied to the acceptance of a memory update rather than to neural or behavioral admissibility. Theorem 4 proves the necessity of *some* environment-separating signal; it does not prove that this exact strict-decrease rule is the uniquely necessary form such a signal must take, and the paper does not claim otherwise.

The third link is Theorem 5, and here the paper's own assumption structure needs to be reported precisely rather than folded together with grounding in general. Theorem 5's convergence conclusion, that the verifier risk $R_t \rightarrow 0$ almost surely at an order-tight geometric or polynomial

rate depending on a parameter β, is proved under Assumptions 5 and 6 specifically: Assumption 5, Non-Degenerate Corrective Mass, requires the acceptance probability to satisfy $p_t \geq c_1 R_t^{\beta}$ whenever risk is positive, and Assumption 6, Proportional Accepted Decrement, requires each accepted update to reduce risk by a fraction proportional to the current risk. Grounding, in the sense of Theorem 4, is what makes the evaluation protocol's risk signal environment-sensitive in the first place; Assumptions 5 and 6 are the further, independent conditions that make risk convergence follow from that grounded signal. A fourth condition, Assumption 4, Verifier Calibration, $V_i(m_e) \leq LR_i(m_e)$, is separately what the paper needs to transfer convergence of the grounded risk functional R_t into convergence of the actual task sub-optimality V_t; it is not itself part of what drives $R_t \to 0$, but rather what certifies that a vanishing verifier risk corresponds to vanishing task error. Grounding alone, absent Assumptions 5, 6, and 4, secures the necessity result of Theorem 4 but not the convergence and calibration guarantees of Theorem 5; treating "grounded" as though it automatically meant "convergent to zero task error" would credit the theorem's conclusion to a premise substantially weaker than what the theorem actually requires.

10 The persistent-harm floor, correctly scoped

Chen et al.'s Theorem 3 establishes a positive error floor for free-form, ungated reflection,

$$\liminf_{T\to\infty} e_T \geq \frac{\nu}{\gamma} > 0,$$

and it is tempting, having established the impossibility result of Theorem 4, to read this floor as the general consequence of any collapse that proceeds without grounding [1]. That reading overstates the theorem. The floor in Theorem 3 holds specifically under Assumption 3, Persistent Harmful Commitment, which requires $\gamma \in (0, 1]$ and $\nu \in (0, \gamma]$ such that $\mathbb{E}[V_{t+1} \mid \mathcal{F}_t] \geq (1 - \gamma)V_t + \nu$ on every reachable state; the paper's own Assumption 2, the weaker one-sided drift condition used for the upper bound in Theorem 2, is explicitly insufficient on its own to establish this lower floor, since Assumption 2 only bounds drift from above and says nothing about a persistent lower rate of harmful commitment. Unconditional commitment alone, without the persistence and rate conditions Assumption 3 specifies, is insufficient to establish a universal positive floor.

The corrected statement, and the one this essay adopts, is accordingly narrower than "ungrounded collapse necessarily converges to nonzero error." It is instead that persistent harmful commitment, in the specific sense Assumption 3 defines, provably yields a nonzero asymptotic error floor, while grounding, together with the calibration and non-degenerate corrective-mass conditions catalogued in the previous section, permits convergence all the way to zero error. The distinction locates the actual work each ingredient does: grounding alone addresses Theorem 4's impossibility; grounding combined with Assumptions 4 through 6 addresses convergence; and it is specifically the combination of ungroundedness with a persistent, rate-bounded pattern of harmful acceptance, not ungroundedness on its own, that Theorem 3 shows must converge to a nonzero floor.

11 Grounding can fail too

Everything so far has argued that grounding is necessary for reliable collapse across text-indistinguishable environments. It does not follow, and should not be allowed to be implied by the preceding sections' emphasis, that grounding is sufficient for correctness. A verifier, however genuinely external to the text channel in the sense of Section 6, can still be stale relative to a changing environment,

misspecified relative to the property it is meant to certify, manipulable by an agent that learns to satisfy the verifier's letter without satisfying its intent, only weakly correlated with the actual task objective, or vulnerable to optimization pressure that exploits exactly the gap between the verifier's proxy and the true target, the general Goodhart-type failure mode. Grounding solves an information-channel problem, whether the evaluating signal's distribution is sensitive to the fact that matters; it does not automatically solve the further and separate problems of calibration, specification, or robustness to optimization against the verifier itself. Chen et al.'s own Assumption 4, Verifier Calibration, is precisely an acknowledgment of this gap: calibration is stipulated as a separate condition rather than derived from grounding, and the paper is explicit that its guarantee concerns verifier risk specifically, noting that on incomplete test suites the theorem's conclusion "concerns verifier risk only" rather than true task utility [1].

The resulting asymmetry is worth stating precisely rather than leaving implicit, since it is what gives the essay's eventual conclusion its correct shape:

$$\text{grounded} \;\not\Rightarrow\; \text{correct},$$

while, restricted to the ambiguous-pair setting Theorem 4 actually constructs,

$$\text{not environment-separating} \;\Rightarrow\; \text{not uniformly reliable}.$$

Grounding is necessary in the narrow, proven sense the theorem establishes; it was never claimed, by the paper or by this essay, to be sufficient in general.

12 Two collapses, one structure

With each formal claim now stated at its correct strength, the correspondence this essay set out to draw can be stated precisely, and stated as what it is: a theoretical reconstruction offered by this essay, not a claim either source paper makes about its relation to the other or to this corpus's vocabulary.

Ge et al.'s failure is, in this corpus's terms, a horizontal failure: collapse fires while the admissible family $\Gamma(A_t)$ still contains more than one task-relevant class, before the completions in $\mathcal{B}(P_t)$ have actually been reduced to a single equivalence class under $\sim_{P_t}$ [2]. Chen et al.'s failure is, by contrast, a vertical failure: collapse, here the acceptance of a proposed memory update, is licensed by a gate whose available signal cannot discriminate between environments in which the same proposal would warrant opposite decisions [1]. The two failures can be made into a small, two-dimensional formal test rather than left as a loose analogy. Let

$$H_t = |\{\phi(P_t, c) : c \in \mathcal{B}(P_t)\}|$$

measure unresolved horizontal multiplicity, using it at least in its binary form, $H_t = 1$ against $H_t > 1$, and let $V_t \in \{0, 1\}$ indicate whether the acceptance channel currently in use is capable of discriminating environments in which the same candidate commitment would carry different consequences. Justified commitment, on this reading, requires both conditions to hold at once,

$$H_t = 1 \quad \text{and} \quad V_t = 1.$$

A system can fail horizontally by committing to a formulation while $H_t > 1$, which is Ge et al.'s documented failure mode; it can fail vertically by accepting a proposal through a non-discriminating gate even when $H_t = 1$, the candidate has been clearly and uniquely specified, which is Chen et

al.'s proven failure mode; and, in principle, it can fail on both axes at once, a case neither paper studies directly but which the two-dimensional test makes visible as a distinct cell rather than an unclassified residual.

The value of stating the correspondence this way, and of the (H_t, V_t) classification it licenses, is that it makes visible a structural kinship, the same underlying operator, collapse, failing for evidential reasons at two different levels of a system, formulation and memory update, that the two papers' own vocabularies, developed independently and for different purposes, do not make visible on their own. Neither paper frames its failure mode in terms of admissible families, collapse operators, or a shared two-axis test, and the synthesis should be read as this essay's contribution rather than as a result either source paper states about itself.

13 What neither paper reaches

This corpus's own tier-3 question, whether an external signal contributes irreducible discriminatory work or whether a sufficiently rich internal process could in principle recompute it, needs careful placement here, since within Theorem 4's own constructed setting the question is already closed rather than open: the theorem's environments are built so that the discriminating fact is, by design, absent from the text channel altogether, and no degree of self-contained sophistication can recover information that is not there to recover. Treating the tier-3 question as unresolved *within that construction* would misread what an indistinguishable-pair proof establishes; a "sufficiently powerful" text-only judge is no better positioned than a weak one when the two environments it must tell apart are, by the proof's own hypothesis, identical at the level the judge can see.

The genuine open question sits one level up, in whatever actual experimental regime a system like SRMA is deployed in rather than in the idealized ambiguous-pair environments Theorem 4 builds to prove its impossibility result. Chen et al.'s own experiments, a hidden-cap resource contest, Overcooked with an exact breadth-first-search value table, and 500 SWE-bench instances with the repository's own test harness as the grounded verifier, are precisely the kind of real deployment regime where it remains a live and unsettled empirical question whether the grounded signal is doing discriminatory work that is genuinely irreducible, in the sense that no feature already implicit in the generated text and the model's own internal state could have been extracted to substitute for it, or whether these regimes are simply not as adversarially constructed as Theorem 4's proof requires, so that some of the grounded signal's apparent necessity in practice is an artifact of not having tried hard enough to extract an internal substitute. The paper's own SWE-bench numbers are suggestive on exactly this point, though only suggestive: on the Kimi K2.5 backbone, the free-form multi-agent system that commits every proposed reflection ungated resolves 58.4 percent of instances, markedly worse than a single non-coordinating mini-SWE-agent baseline on the identical backbone (70.8 percent, external reference), while the grounded SRMA system resolves 72.2 percent; the controlled DeepSeek runs show the same direction, 68.2 percent for the single-agent baseline against 71.4 percent for grounded SRMA. Ungated multi-agent coordination, on this evidence, does not merely fail to help; it actively hurts relative to not coordinating at all, which is a striking result if it holds up, but the experiments were not designed to isolate how much of that gap the constructed impossibility theorem's specific logic explains, as opposed to a more ordinary difference in evaluation quality between a text-only judge and an executable test harness. Settling the sharper question would require an intervention neither paper as summarized here reports: holding a proposal's text-level presentation fixed while manipulating access to the grounded signal directly, and checking whether performance in the system's actual evaluation regime tracks the grounded signal's presence in a way a sufficiently careful internal-only baseline cannot match.

Absent that intervention, SRMA's grounded gate should be described as causally sufficient by construction, since it is simply given the environment-dependent signal it needs, rather than as having demonstrated that such access is the only route available in principle for problems of the kind it is tested on. In this respect the paper's grounded architecture has, for the narrow class of problems it addresses, gotten further in practice than this corpus's own wave-based hypothesis has gotten for the analogous claim about neural computation, where the tier-3 question, whether phase or field structure does causally irreducible work rather than work a sufficiently rich rate-based description could in principle recover, remains open rather than settled, and remains open for exactly the same reason: the decisive intervention has not yet been run.

A further limitation concerns what Ge et al.'s own experimental design does not test. The paper documents that InterOPT declares readiness while gaps remain unresolved and that this correlates with subsequent audit findings of prematurity, but it does not report an intervention that holds the fluency or confidence of the agent's output fixed while independently varying whether the gap ledger has actually been driven to completeness, which would be the more decisive test of whether recovery from a premature formulation tracks genuine gap resolution or merely tracks how confident the output sounds, the same R_t versus C_t divergence isolated in Section 6 above. Such an intervention is not present in the paper as summarized here and would be the natural next experiment for establishing that the ledger's OPEN and ASKED statuses are doing genuine discriminating work rather than merely correlating with an independently confident-sounding output.

14 Conclusion

The strongest proposition the two papers jointly support, stated at the strength their proofs and findings actually license rather than at the strength an appealing generalization might suggest, is the following.

> When observationally indistinguishable evidence states require different commitments, no gate restricted to that evidence can reliably license collapse across them. A discriminating signal must enter from outside the shared evidence class.

This captures Chen et al.'s Theorem 4 without universalizing it into a claim the paper does not make about self-evaluation in general; the proposition is conditioned, as the theorem is, on the evidence states in question actually being indistinguishable to the gate and actually requiring different commitments, and it says nothing about cases, the checkable-proof case foremost among them, where the transcript itself already carries the discriminating fact [1]. Nor does it say that a grounded signal, once supplied, guarantees correctness; Section 11 above stated the opposite asymmetry explicitly, and that caution should be read as inseparable from the proposition rather than as an afterthought to it.

Ge et al. supply the complementary empirical half of the picture: a demonstration, quantified in one experimental setting and, tellingly, exhibited by the paper's own proposed remedy rather than by an unremedied baseline alone, that systems in practice declare readiness before the publicly available formulation space has actually been reduced to a single task-relevant equivalence class, and that a substantial share of the relevant gaps are never even raised as questions before that declaration is made [2]. Read together, and read at the strength each result actually supports, the two papers converge on a single, narrower, and more defensible claim than either an uncorrected synthesis or either paper read in isolation would suggest: collapse, in this corpus's sense, is warranted only where the evidence available to the gate performing it is adequate to the distinction the commitment actually depends on, where that adequacy is judged along both the horizontal

axis of unresolved formulation multiplicity and the vertical axis of discriminating grounding, and where neither model capacity, confident presentation, nor the mere fact of having consulted an external signal is a substitute for that adequacy actually holding.

References

[1] Yihang Chen, Yuxiang Chen, Yuxuan Huang, Meng Fang, Weilin Luo, and Jun Wang. Bilevel Coordinated Reflection: A Game-Theoretic Approach to Multi-Agent LLM Systems. arXiv:2609.02750 [cs.AI], 2026. https://arxiv.org/abs/2609.02750

[2] Sihan Ge, Yichen Lin, Chenyu Zhou, Jianghao Lin, Tao Yao, and Dongdong Ge. Ask Before You Optimize: Dynamic Pre-Formulation Clarification for Interactive Optimization. arXiv:2609.05258 [math.OC], 2026. https://arxiv.org/abs/2609.05258