Decision-Space Collapse in Advisory Language Models

Measuring Trajectory Omission, Framing Sensitivity, and Recovery through Decision-Space Integrity

Andrew J. Cousins
Independent Researcher · Sheffield, United Kingdom
andrewcousins@decisionspaceintegrity.com

Date
8 June 2026

Version
Preprint v1.1

This work was developed independently and does not represent the views of any employer or affiliated organisation.

Intellectual property notice. A UK patent application has been filed covering aspects of the Decision-Space Integrity architecture and associated trajectory-conditioned audit, control, re-audit, and evidence-generation methods. This preprint describes the decision-space evaluation methodology and bounded empirical evidence base; it does not license the separate enterprise DSI implementation.

v1.1 adds cross-model confirmation studies, prompt-diversity confirmation, a dedicated warning-obligation study, a distinction between trajectory omission and obligation omission, and expanded controlled-recovery evidence. These additions strengthen the omission-recovery findings while preserving the original claim boundaries.

ABSTRACT

Advisory language models increasingly mediate under-specified decisions, yet standard evaluations do not measure whether plausible decision trajectories remain visible under ambiguity. This paper introduces decision-space collapse: the narrowing of the visible decision space when advisory systems surface some plausible courses of action while omitting others.

We introduce Decision-Space Integrity (DSI) as a framework for measuring that phenomenon through configured expected decision-space maps, trajectory classification, expected-trajectory coverage, omission measurement, directional-imbalance indicators, and controlled recovery testing.

In a 6,480-output empirical evaluation across four model families, three advisory domains, and three controlled prompt conditions, we find measurable expected-trajectory omission under ambiguous prompts. The primary empirical evaluation is domain-conditioned: advisory domains are assigned by the experimental design rather than inferred by an automatic domain resolver, and responses are classified against the configured expected map for that supplied domain.

Minimal tense and wording perturbations produce mixed or modest changes rather than strict invariance, while third-person framing consistently improves measured expected-trajectory coverage and reduces DBI-risk, a directional-imbalance indicator across tested provider/domain cells.

A separate controlled-recovery evidence baseline shows that, under the current expected-map, classifier, scoring, and provider-completion gates, DSI can identify omitted configured trajectories, generate omission-aware controlled-response conditions, re-audit post-control outputs, and measure whether configured expected-trajectory coverage improves.

Subsequent confirmation studies reproduced this omission-recovery effect across Claude Sonnet 4, Gemini 2.5 Flash, and GPT-4.1-mini, across multiple prompt samples and an expanded prompt set. Across these studies, omission-aware intervention consistently recovered more omitted configured trajectories than generic “be-more-comprehensive” prompting while requiring substantially less added text. A separate warning-obligation study (single model, finance) found that preserving and recovering required warning obligations is a distinct control problem from recovering omitted trajectories.

The same measurement spine also detected a control failure — omission-aware recovery could occasionally drop a required warning — which a naive suffix-based fix did not resolve; representing warning obligations

as first-class, conditionally-applied requirements eliminated the observed warning regressions in that held-out finance warning-sensitive subset and preserved more warning obligations.

These results support decision-space collapse as an observable output-level phenomenon and DSI as a bounded measurement framework. They do not establish model intent, latent preference, advice quality, safety certification, independent classifier validation, provider ranking, tri-domain warning-preservation, or exhaustive real-world decision coverage. The original comparison was subset-scoped; subsequent confirmation studies are cross-model, while warning-preservation evidence remains finance-scope.

CCS Concepts. Computing methodologies → Natural language generation; Human-centred computing → HCI; Information systems → Decision support systems.

Keywords. large language models · advisory AI · decision-space collapse · decision-space integrity · trajectory omission · expected-trajectory coverage · prompt framing · controlled recovery · AI evaluation

CONTENTS

At a Glance

§ Headline findings

- **Measurable omission.** Under ambiguous advisory prompts, tested systems often omit configured expected trajectories; baseline expected-trajectory coverage is materially below complete coverage.
- **Framing sensitivity.** Decision-space visibility is sensitive to prompt framing. Minimal tense/wording perturbation is mixed; third-person framing improves measured coverage in all tested provider/domain cells.
- **Directional imbalance.** DBI-risk decreases under third-person framing in all tested provider/domain cells, indicating that framing changes not only quantity of coverage but also the balance of surfaced trajectories.
- **Controlled recovery, cross-model**. Omission-aware controlled-response conditions restore measured visibility of omitted configured trajectories. The comparison was reproduced across three model families (Claude Sonnet 4, Gemini 2.5 Flash, GPT-4.1-mini), multiple samples, and an expanded prompt set: omission-aware intervention recovered more omitted configured trajectories than generic comprehensiveness with substantially less added text in every reported model-level aggregate.
- **Self-detected control failure.** The same audit spine detected that omission-aware recovery could drop a required warning, and measured that a structured, conditionally-applied intervention repaired this on held-out finance warning-sensitive cases (finance-scope).
- **Bounded interpretation.** The paper evaluates visibility, not advice quality, user outcomes, safety certification, or provider ranking.

CORE CLAIM

Under ambiguity, advisory language models can narrow the visible decision space by omitting configured expected trajectories. Decision-space collapse names this observable narrowing; Decision-Space Integrity measures it through expected-trajectory coverage, omission detection, framing sensitivity, directional-imbalance indicators, and controlled recovery.

CLAIM BOUNDARY

DSI reports configured expected-trajectory visibility under specified domain packs, expected-map construction, classifier versions, scoring versions, and prompt conditions. In the primary empirical evaluation, domain assignment is controlled by study design; the reported results therefore evaluate trajectory visibility within known advisory domains, not open-domain routing or prompt-admission accuracy. A high coverage score does not mean perfect advice, model safety, independent validation, or exhaustive real-world decision coverage.

INTERPRETATION NOTE

The empirical evaluation shows measurable omission and prompt-framing sensitivity. It does not support strict perturbation invariance. The results should be interpreted as evidence of observable decision-space behaviour, not as evidence of internal model intent or latent probability allocation.

1 Introduction

Advisory language models are increasingly used in domains where users face multiple plausible courses of action: relationships, careers, finance, health, education, and other under-specified decisions. In these settings, the risk is not only that a model gives incorrect, unsafe, or unhelpful advice. A further risk is that the response narrows the visible decision space by surfacing some plausible trajectories while omitting others.

This paper studies **decision-space collapse**: the narrowing of the visible decision space when advisory systems surface some plausible trajectories while leaving other configured expected trajectories unsurfaced. Collapse does not mean that only one possible action exists. Nor does it require a claim that the model internally chose a destination. It means that, in the observed response, the visible set of advisory trajectories is narrower than the configured expected decision space for that context.

TERMINOLOGY

Decision-space collapse is distinct from two established uses of "collapse" in machine learning. Mode collapse refers to a generative model concentrating probability on a narrow set of outputs, reducing output diversity. Model collapse refers to degradation that can occur when models are trained on synthetic or model-generated data. Both describe distributional or training-time properties of a model. Decision-space collapse, as defined in this paper, is an output-level property of a specific advisory response: the omission of plausible decision trajectories that a configured expected-path map identifies as relevant. A model may exhibit high overall output diversity while still narrowing the visible decision space within a single advisory answer.

Standard evaluation frameworks primarily assess output-level properties such as correctness, safety, helpfulness, coherence, preference alignment, factuality, or task performance. These measures are necessary, but they do not directly capture which decision paths are made visible to a user. A response can be fluent, safe, and apparently helpful while still omitting plausible courses of action that matter to the advisory decision.

Decision-Space Integrity (DSI) provides a framework for measuring decision-space collapse by evaluating advisory outputs against configured expected decision-space maps. Rather than asking only whether an answer is acceptable in isolation, DSI asks which expected trajectories are surfaced, which are omitted, whether framing changes measured visibility, and whether omitted trajectories can be recovered through controlled intervention.

The primary empirical evaluation contains 6,480 scored outputs across four provider/model families, three advisory domains, three prompt conditions, six prompts per domain/condition, and thirty samples per prompt. The evaluated prompt conditions are baseline, tense_perturbation, and third_person_frame. The tense_perturbation condition is treated as a minimal wording comparison; the third_person_frame condition is treated as a deliberate framing shift rather than as evidence of strict semantic invariance.

The paper separates phenomenon from measurement. Decision-space collapse is the phenomenon. DSI is the measurement framework. Expected-trajectory omission is the observable event in which a trajectory present in the configured expected map is not detected in the observed response under the current classifier. Expected-trajectory coverage is the primary visibility metric. Controlled recovery tests whether previously omitted trajectories can be restored through intervention and re-audited under the same configuration.

TAXONOMY POSITIONING

Expected trajectories are operational constructs for measurement. They are defined within a domain pack, expected-map configuration, classifier version, and scoring version. DSI therefore measures configured expected-trajectory visibility, not perfect advice, complete semantic understanding, or exhaustive real-world decision quality.

The empirical study reported here uses controlled domain assignment. Prompts are evaluated within pre-specified career, finance, or relationship conditions, and the relevant expected map is selected from that

experimental condition. Automatic domain resolution is therefore outside the scope of the primary empirical results, although it may be relevant for deployed DSI systems.

Across the primary empirical evaluation, the main empirical result is measurable expected-trajectory omission under ambiguity. Baseline expected-trajectory coverage is materially below complete coverage. Minimal tense/wording perturbation produces mixed or modest changes, while third-person framing consistently improves measured expected-trajectory coverage and reduces DBI-risk. These findings are consistent with decision-space collapse as a measurable output-level phenomenon.

This work focuses on observable advisory output behaviour. It does not infer user intent, predict user action, or prescribe which trajectory a user should choose. A trajectory's relevance depends on whether multiple plausible courses of action exist under the advisory context, not on whether any particular alternative is normatively preferred.

1.1 Contributions

This work makes five contributions.

Contribution	Description
Conceptual	We introduce decision-space collapse as an output-level phenomenon in advisory language-model responses.
Measurement	We present Decision-Space Integrity as a framework for measuring expected-trajectory omission, coverage, directional imbalance, and recovery relative to configured expected maps.
Empirical	We report a 6,480-output empirical evaluation showing measurable expected-trajectory omission and prompt-framing sensitivity across four provider/model families, three advisory domains, and three prompt conditions.
Recovery	We evaluate whether omission-aware controlled interventions recover omitted expected trajectories, compare them with neutral regeneration and generic comprehensiveness prompting on a subset, and report a measured warning-preservation failure and its conditional repair (initially on a single hosted model, then confirmed across three model families; warning preservation remains finance-scope).
Boundary	We distinguish observable decision-space collapse from unsupported claims about model intent, latent preference, internal probability allocation, advice quality, safety certification, or exhaustive real-world coverage.

1.2 Roadmap

Section 2 situates decision-space collapse and DSI relative to model evaluation, alignment, recourse, pluralistic alignment, and decision architecture. Section 3 defines the phenomenon, the measurement framework, expected trajectories, omission, coverage, recovery, and related patterns. Section 4 describes the measurement framework and evaluation design. Section 5 reports the empirical results. Section 6 reports controlled-recovery evidence, a bounded intervention comparison, and a conditional warning-preservation result. Section 7 discusses implications. Section 8 states limitations and future work. Section 9 concludes.

1.3 Potential Applications

Although this paper is not a deployment study, decision-space collapse has practical relevance wherever advisory language models are evaluated across versions, providers, or intervention policies. A configured expected-map approach could support regression testing by detecting whether previously visible trajectories disappear after a model update, provider change, prompt change, or policy modification. It could also support audit review by recording which configured trajectories were surfaced or omitted under specified evaluation conditions.

These applications are suggested implications of the measurement framework rather than validated deployment claims. The present paper does not establish regulatory acceptance, production suitability, or downstream user-outcome improvement.

2 Related Work

2.1 Evaluation of Large Language Models

Evaluation of large language models has primarily focused on correctness, coherence, safety, calibration, factuality, robustness, helpfulness, and benchmark performance of individual responses. Benchmarks such as MMLU and related multi-task evaluations [1–3] assess factual accuracy, task performance, and broad scenario coverage, while work on chain-of-thought prompting, self-consistency, and task-level evaluation examines reasoning quality [4, 5]. Other studies investigate prompt sensitivity and instability, showing that small variations in phrasing can lead to substantial differences in output [21–24].

These approaches provide important evidence about output quality, but they usually treat each response as an independent unit of evaluation. They do not directly measure which plausible decision trajectories are visible to a user, which are omitted, or whether a response preserves a non-trivial expected decision space under ambiguity.

Decision-space collapse addresses this gap conceptually, while DSI provides the measurement framework used in this paper. A response may be accurate, safe, and helpful while still exhibiting collapse relative to a configured expected map. Conversely, a response may preserve multiple trajectories while remaining imperfect in other evaluation dimensions.

2.2 Alignment, Bias, and Normative Behaviour

A growing body of work examines how alignment techniques shape model outputs through reinforcement learning from human feedback, preference optimisation, safety tuning, and related methods [6–8]. These approaches can improve helpfulness and reduce harmful outputs, but they may also create systematic response patterns that are not transparent from a single answer.

Research on fairness and bias [9–12] further demonstrates that models can encode and reproduce normative assumptions, including assumptions about appropriate action, risk, responsibility, or socially preferred behaviour. Recent work has also shown that LLMs exhibit systematic cognitive biases in high-stakes decision-making tasks [25] and amplified omission bias in moral decision-making [26].

DSI is adjacent to this work but has a narrower measurement target. It does not infer the model's intent, moral stance, or social preference directly. It measures whether configured expected trajectories are surfaced or omitted in advisory responses. This makes omission and collapse measurable without requiring the stronger claim that any particular omission proves a model-level normative bias.

2.3 Decision-Making and Choice Architecture

The study of decision-making has long emphasised that choices are shaped by the structure of the environment in which they are presented. Classical models of bounded rationality question the assumption that all considered options are equally salient [13], while behavioural economics shows that framing and organisation of available choices can systematically influence outcomes [14, 15].

This paper applies that insight to advisory language models by treating responses as components of a decision environment. A response may widen, preserve, or narrow the visible option set. Prompt framing is therefore not merely a surface wording change. It can act as a decision-space control surface, changing which trajectories are surfaced or omitted.

2.4 Recourse and Alternative Exposure

Research in algorithmic recourse and counterfactual explanation focuses on identifying alternative actions that could lead to different outcomes [16–18]. These approaches are valuable because they make hidden or unavailable options more explicit after a decision has been made.

DSI addresses a complementary problem. In advisory language-model settings, the issue is often not only whether a user can obtain recourse after a decision, but whether plausible trajectories are visible during the advisory exchange itself. A model may provide a coherent recommendation while failing to expose other expected paths.

This distinction also motivates recovery-based evaluation. If an expected trajectory is omitted in an initial response, DSI can generate a controlled intervention, re-audit the post-control response, and measure whether the omitted path is recovered. Recovery is a visibility measurement, not a claim that the resulting advice is superior.

2.5 Pluralistic Alignment

This work also relates to pluralistic alignment and value-diversity approaches [19, 20], which emphasise that model behaviour should not collapse prematurely onto a single normative perspective where multiple reasonable perspectives exist. Such work is especially relevant in advisory domains, where appropriate responses may depend on context, user values, constraints, and uncertainty.

DSI does not attempt to solve pluralistic alignment. Instead, it provides an operational measurement layer for one related property: whether a response preserves a configured set of plausible decision trajectories under ambiguity. This allows pluralism-relevant omissions to be detected without requiring the framework to determine the correct normative answer.

2.6 Positioning of This Work

This paper introduces decision-space collapse as an output-level phenomenon in advisory systems and DSI as a framework for measuring that phenomenon. It builds on prior work in evaluation, alignment, recourse, pluralistic alignment, and decision theory, but shifts the unit of analysis from isolated response quality to decision-space visibility.

The framework contributes three linked ideas. First, advisory outputs can be evaluated against an expected decision-space map rather than only against generic helpfulness or correctness criteria. Second, prompt framing can change measured expected-trajectory coverage, making framing a decision-space control surface rather than a nuisance variable alone. Third, omission-aware interventions can be evaluated by re-auditing post-control outputs and measuring whether omitted trajectories are recovered.

Framework type	Primary question
Accuracy evaluation	Is the response correct?
Safety evaluation	Is the response harmful?
Fairness evaluation	Are outcomes equitable?
Alignment evaluation	Does the response follow specified preferences or values?
Agent evaluation	Did the system complete the task?
Recourse evaluation	Are alternative actions available after a decision?
DSI	Has the visible decision space narrowed, and which configured trajectories are surfaced, omitted, or recovered?

Note. Table 1. Positioning of DSI relative to existing evaluation approaches.

Related work on heuristic collapse in advisory behaviour [27] examines how models reduce complex decisions to dependence on a small number of dominant inputs. That work focuses on input-output sensitivity and feature-level compression. DSI addresses a complementary question: how outputs are structured into trajectories and how alternative decision pathways are preserved or omitted. Broader work on generative model collapse [28] is relevant as background, but the present work concerns within-response decision-space visibility rather than generational degradation of model outputs.

Recent work on behavioural-disposition alignment, framing sensitivity, and decision-process fidelity reinforces the need to evaluate advisory model behaviour beyond single-response correctness. Behavioural-disposition evaluations show that models may deviate from human consensus and may fail to capture the distributional range of human viewpoints where consensus is absent [29]. Framing-sensitivity work similarly shows that fact-preserving but differently framed inputs can materially change model decisions [30]. Process-oriented evaluations of LLM decision-making report that models may under express round-to-round variability relative to humans, even when exposed to human decision trajectories [31]. These findings are complementary to the present work: they examine behavioural alignment, framing instability, or process fidelity, while DSI measures which configured decision trajectories remain visible within advisory responses.

2.7 Prompt Sensitivity and Framing

Prior work has shown that language-model outputs can be sensitive to prompt wording, framing, and context [21–24]. In many evaluations, such sensitivity is treated primarily as a robustness problem: a model should ideally produce consistent answers under semantically equivalent inputs.

DSI treats prompt sensitivity differently. In advisory settings, framing can change which decision trajectories are visible. The relevant question is therefore not only whether outputs remain stable under perturbation, but whether different framing conditions preserve, reduce, or recover expected decision-space coverage.

3 Decision-Space Collapse and DSI

This paper distinguishes between a phenomenon and a measurement framework. The phenomenon is decision-space collapse: the narrowing of the visible decision space when plausible trajectories become unsurfaced under ambiguity. Decision-Space Integrity (DSI) is the framework used to measure that phenomenon through observable advisory outputs.

3.1 Core Definitions

A **decision trajectory** is an operationally defined course-of-action category within an advisory context. Trajectories are defined at the level of decision structure rather than surface phrasing. For example, in a career advisory setting, "remain in current role and negotiate conditions," "move to a new role," and "pause the decision pending further information" may be distinct trajectories even if individual responses phrase them differently.

An **expected trajectory** is a trajectory included in the configured expected decision-space map for a given domain, prompt family, and evaluation condition. Expected trajectories are not claimed to exhaust all valid human choices. They are measurement objects used to test whether a response preserves a non-trivial set of plausible advisory paths under ambiguity.

A trajectory is **surfaced** when the observed response contains sufficient semantic evidence, under the current classifier, that the trajectory has been presented to the user as an available path. A trajectory is **omitted** when it appears in the expected map but is not detected in the observed response.

Decision-Space Integrity (DSI) is a framework for measuring decision-space collapse relative to a configured expected decision-space map. DSI evaluates whether configured expected trajectories remain visible, become omitted, or are subsequently recovered under specified advisory conditions. DSI is therefore a measurement framework rather than a theory of model intent, internal preference, or latent decision processes.

CONCEPT HIERARCHY

Decision-space collapse is the broad phenomenon examined in this paper. Trajectory omission is one observable manifestation of that phenomenon. Trajectory narrowing, directional imbalance, and destination bias are related patterns. DSI provides the measurement framework used to evaluate them.

3.2 Configured Expected Maps

DSI relies on **configured expected decision-space maps**. An expected map specifies the trajectories that should remain available for measurement in a given advisory context. The map may be derived from a domain pack, prompt taxonomy, expert-coded expected alternatives, or other documented configuration source.

The expected map is not a universal representation of the real-world decision space. It is a bounded measurement instrument. Its function is to make omission auditable: if a trajectory is included in the expected map but absent from the response under the current classifier, the system records an omission.

Expected maps function as measurement reference objects. DSI does not claim that an expected map represents the complete real-world decision space. It provides a documented reference against which collapse, omission, coverage, directional imbalance, and recovery can be evaluated.

The expected map is therefore a normative and configurable measurement reference, not an objective catalogue of all reasonable real-world options. DSI does not eliminate the need to justify, review, contest, or version the map. Instead, it makes the consequences of a given map explicit: omission is measured only relative to the configured expected trajectories. Future deployments should treat map construction as a governed process involving domain expertise, stakeholder review, disagreement records, and versioned provenance.

3.3 Observable Indicators of Collapse

DSI treats advisory responses as evidence of decision-space visibility. Three response-level events are central.

- **Surfaced trajectory.** An expected trajectory detected in the response. Surfacing does not require the model to recommend the trajectory as best; it requires that the trajectory be made available as a plausible path.
- **Omitted trajectory.** An expected trajectory not detected in the response. Omission is measured relative to the expected map and classifier configuration. It is not a claim that the response is unsafe or that the advice is necessarily wrong.
- **Recovered trajectory.** A previously omitted expected trajectory that appears in a post-control response after an intervention. Recovery allows DSI to evaluate whether omission-aware intervention improves measured decision-space visibility.

3.4 Related Observed Patterns

Decision-space collapse may appear in several observable forms. The patterns below should be interpreted as potential manifestations of collapse rather than as separate phenomena.

Pattern	Definition
Trajectory collapse	Concentration of visible decision trajectories into a narrow subset of the configured expected decision space.
Trajectory omission	One or more configured expected trajectories are not surfaced in an advisory response.
Trajectory narrowing	A response presents only a small subset of the expected map, even though multiple plausible trajectories are configured for the context.
Directional imbalance	One trajectory is surfaced substantially more often than another relevant trajectory under comparable conditions.
Destination bias	An interpretive pattern in which collapse repeatedly favours a particular destination or class of destination.

Note. In the present paper, omission and coverage are the primary measured constructs. Collapse, directional imbalance, and destination bias are interpreted through patterns of omission, coverage, DBI-risk, and recovery. They should not be assumed from a single response.

3.5 When Does Omission Matter?

Omission is not always a defect. Many advisory contexts legitimately call for narrowing: some options may be unsafe, irrelevant, impossible, or outside the user's stated constraints. DSI is most informative when the prompt is under-specified, multiple plausible trajectories remain available, and the response presents one path while leaving other configured expected paths unsurfaced.

In such cases, omission matters because the user's visible decision space may be narrower than the advisory situation warrants. The concern is not that every response must list every possible option. The concern is that, under ambiguity, plausible expected trajectories can disappear from the response without being explicitly ruled out.

Omission should therefore be interpreted as evidence requiring investigation rather than as automatic proof of defective advice. Its significance depends on the broader pattern of collapse observed within the configured decision space.

3.6 Relation to Existing Evaluation Dimensions

DSI complements rather than replaces existing evaluation dimensions. Correctness, safety, fairness, and helpfulness remain necessary. DSI asks a different question: whether a response surfaces the configured expected trajectories relevant to an ambiguous advisory context.

Metric or lens	Unit of analysis	Primary question	Captures omission?
Accuracy	Output / task	Is the answer factually or task-correct?	No
Safety	Output / policy	Does the answer avoid prohibited harm?	Partial
Fairness	Token / group / outcome	Are protected groups treated equitably?	Partial
Helpfulness	Output / preference	Is the response useful or preferred?	Partial
Robustness	Prompt / response pair	Does output remain stable under perturbation?	No
Recourse	Decision / outcome	Are alternatives available after decision?	Partial
DSI	Trajectory / expected map	Which expected paths are surfaced, omitted, or recovered?	Yes

Note. Table 2. Existing evaluation dimensions and their coverage of trajectory omission.

A response can be accurate, safe, fair in tone, and helpful while still omitting a configured expected trajectory. Conversely, a response can surface many trajectories while still being unsafe, inaccurate, or unhelpful. DSI should therefore be interpreted alongside existing evaluation dimensions, not as a replacement for them.

3.7 Black-Box Scope

The present implementation of DSI is black-box. It evaluates prompts, responses, expected maps, classifier outputs, and scoring records. It does not require access to model weights, logits, hidden states, or provider-side ranking probabilities.

This scope is a strength for auditability because it allows DSI to be applied to hosted advisory systems where internal model data is unavailable. It is also a claim boundary. DSI can measure observed decision-space visibility in outputs; it cannot directly establish the model's internal causal mechanism for an omission.

3.8 Configuration and Versioning

DSI measurements are configuration-bound. A coverage score depends on the domain pack, expected-map construction, trajectory definitions, classifier version, scoring version, and prompt condition. For this reason, DSI results should be reported with configuration and provenance metadata rather than treated as free-standing universal measurements.

This versioning requirement is not incidental. It is what allows omission claims to be audited. A reported omission should be traceable to the expected map, the observed response, the classifier decision, and the scoring configuration used at the time of evaluation.

4 Measurement Framework and Evaluation Design

The objective of the evaluation is not to determine whether advisory systems recommend the correct decision. The objective is to evaluate whether decision-space collapse can be observed and measured through DSI under controlled advisory conditions.

The evaluation focuses on four reader-facing questions: Do advisory responses omit configured trajectories under ambiguity? Does measured visibility vary under alternative prompt framings? Do observed patterns exhibit directional imbalance? Can omitted trajectories be recovered through intervention?

4.1 DSI Pipeline

DSI is implemented as a black-box audit-and-control pipeline for advisory model outputs. Given a prompt, a response, a supplied advisory domain, and a prompt condition, the system evaluates whether configured expected decision trajectories are surfaced, omitted, or recovered under that domain-specific expected map. In the primary empirical evaluation, the advisory domain is supplied by the experimental design rather than inferred automatically. The response classifier is still required: it determines whether each configured expected trajectory is surfaced, partially surfaced, discouraged, or omitted within the supplied domain.

The pipeline has ten stages: prompt and response intake; ambiguity assessment; expected decision-space map construction; response classification against configured trajectories; expected-trajectory coverage scoring; omission and risk assessment; intervention decision; controlled-response generation where applicable; post-control re-audit and intervention-effect recording; and evidence output.

EVALUATION LOGIC

Configured expected space → advisory response → DSI evaluation → omission, coverage, directional imbalance → evidence of decision-space collapse → controlled recovery evaluation.

The present paper uses this pipeline in two related ways. The primary empirical evaluation uses the audit and scoring stages to measure expected-trajectory coverage and framing sensitivity. The controlled recovery evidence uses the intervention and re-audit stages to test whether omitted trajectories can be recovered.

All reported measurements are classifier-version-bound: a trajectory is treated as surfaced or omitted according to the expected-map, domain-pack, classifier, and scoring configuration used for the evaluation.

The paper-level classifier used for the primary evaluation is a deterministic expected-map-aware trajectory classifier using configured trajectory evidence terms and domain-pack evidence policies, rather than a human-rater system or an LLM-as-judge. Its outputs should therefore be interpreted as rule-mediated trajectory decisions under the released replication configuration.

4.1.1 Domain-Conditioned Evaluation Scope

The empirical evaluation is domain-conditioned. Each prompt belongs to a pre-specified advisory domain — career, finance, or relationship — and the configured expected map for that domain is used for scoring. The study does not evaluate automatic domain detection, open-domain routing, or prompt-admission accuracy. Those functions are relevant to product deployment, but they are separate from the paper's primary measurement question.

The classification step in the empirical study is therefore trajectory classification, not domain classification. For each generated response, DSI evaluates whether the expected trajectories configured for the supplied domain are surfaced, partially surfaced, discouraged, or omitted under the classifier version used for the evaluation.

4.2 Metrics as Collapse Indicators

Expected-trajectory coverage is the primary visibility indicator within DSI. Coverage is defined as the proportion of configured expected trajectories detected as surfaced by the current classifier:

```
Coverage = |surfaced expected trajectories| / |configured expected
trajectories|
```

Lower coverage indicates greater narrowing of the configured decision space. Coverage does not measure correctness, safety, or advice quality. It measures preservation of configured decision-space visibility. Because coverage is classifier-mediated, it should be interpreted as a measurement under the specified expected-map and classifier configuration, not as a direct estimate of human judgement.

Omission count reports the number of configured expected trajectories not detected in the observed response:

```
Omission count = |configured expected trajectories| − |surfaced expected
trajectories|
```

Omission rate reports the omitted share of the expected map:

```
Omission rate = 1 − Coverage
```

Omission is treated as the primary observable event associated with decision-space collapse. However, omission alone does not establish collapse. Collapse is inferred from broader patterns of narrowing, concentration, directional imbalance, framing sensitivity, and recovery across evaluated conditions.

4.3 CT-risk, SR-risk, and DBI-risk

Earlier versions of this work used CT, SR, and DBI as distributional trajectory metrics over repeated samples. The current framework retains CT/SR/DBI-style measures as bounded risk indicators under an expected-map-aware scoring stack.

CT-risk indicates concentration or narrowing risk where observed responses repeatedly surface a limited subset of expected trajectories. SR-risk indicates structural-reduction risk where the response or response set fails to preserve sufficient expected-trajectory variety under the configured map. DBI-risk indicates directional-imbalance risk where one relevant trajectory is favoured or surfaced more strongly than another under comparable conditions.

These measures should be interpreted as operational indicators of observed collapse patterns, not as direct evidence of internal model probability allocation. Destination bias is treated as an interpretive pattern that may emerge from repeated directional collapse. It is not required for collapse to occur.

TRAJECTORY METRICS

CT-risk, SR-risk, and DBI-risk are bounded operational indicators derived from expected-map-aware scoring.

CT-risk, or concentration/trajectory-collapse risk, indicates whether observed responses concentrate visibility into too few expected trajectories.

SR-risk, or structural-reduction risk, indicates whether the response fails to preserve sufficient expected-trajectory variety under the configured map.

DBI-risk, or directional-imbalance risk, indicates whether the surfaced/omitted trajectory pattern is directionally imbalanced, such that one relevant trajectory or class of trajectory is favoured or preserved more strongly than another under comparable conditions.

These indicators are output-level diagnostics. They do not establish model intent, internal preference, latent probability allocation, or advice quality.

4.4 Prompt Conditions

Each domain is evaluated under three prompt conditions: baseline, tense_perturbation, and third_person_frame. The baseline condition presents the advisory ambiguity in its original form. The tense_perturbation condition provides the minimal wording/tense comparison. The

third_person_frame condition reframes the advisory situation around a third person. It is treated as a deliberate framing shift, not as a strict semantic minimal pair and not as evidence of perturbation invariance.

4.5 Controlled Recovery

Controlled recovery evaluation serves as a supporting evidence mechanism. If omitted trajectories can be restored under controlled conditions, the resulting change provides additional evidence that configured decision-space visibility can be measured and influenced through the framework. The general definition remains recovery measurement, not advice quality. A bounded comparison against neutral regeneration and generic comprehensiveness prompting is reported on a revised-reproduction subset in Sections 6.5–6.6; it does not claim that omission-aware control yields superior advice or is generally superior across providers and domains.

Recovery is measured using the difference between the baseline audit and the post-control audit:

```
Coverage delta = post-control coverage − baseline coverage
```

A positive coverage delta indicates that the controlled response surfaced more configured expected trajectories than the original response. This should be interpreted as measured restoration of configured decision-space visibility under the same evaluation configuration, not as proof that the controlled advice is correct, safe, or optimal.

4.6 Evidence Labels

For reproducibility, the evidence package retains compact evidence labels corresponding to omission, perturbation, framing, and recovery tests. In the main narrative, these are expressed as research questions and findings rather than as the primary conceptual structure of the paper.

Evidence label	Reader-facing question	Primary measure
Omission	Do advisory responses omit configured trajectories under ambiguity?	Coverage / omission rate
Minimal perturbation	Do minor wording changes produce large visibility changes?	Coverage delta / DBI-risk delta
Framing sensitivity	Does third-person framing change visible decision-space coverage?	Coverage delta / DBI-risk delta
Controlled recovery	Can omitted trajectories be restored through intervention?	Post-control coverage delta

Note. Table 3. Reader-facing research questions and evidence labels.

5 Empirical Results

The empirical evaluation examined whether configured expected trajectories remained visible under ambiguity, whether visibility changed under alternative framing conditions, whether observed patterns exhibited directional imbalance, and whether omitted trajectories could be recovered through intervention.

5.1 Evaluation Setup

The primary evaluation uses a crossed provider/model × domain × prompt-condition design. The evaluation contains four provider/model families, three advisory domains, three prompt conditions, six prompts per domain/condition, and thirty samples per prompt. This produces 1,620 scored outputs per provider/model and 6,480 scored outputs in total.

Unit	Count
Prompts per domain/condition	6
Samples per prompt	30
Scored outputs per domain/condition/provider cell	180
Advisory domains	3
Prompt conditions	3
Scored outputs per provider/model	1,620
Provider/model families	4
Total scored outputs	6,480

Note. Table 4. Empirical evaluation structure.

Domain labels in this evaluation are part of the experimental design. The study does not ask DSI to infer whether a prompt is career-, finance-, or relationship-related. Instead, each prompt is evaluated under its assigned domain condition, and that assignment determines the expected trajectories used for scoring. The measured quantities therefore reflect response-level trajectory visibility within controlled advisory domains, not the accuracy of an automatic domain resolver.

The aggregate results in this section (Tables 5–6) are computed from the original paper evaluation run. A revised reproduction of the identical 54-prompt matrix was subsequently executed under the current architecture; that revised run is the basis for the controlled intervention comparison (Sections 6.5–6.6) and the human-validation materials (Section 8.4, Code and Data Availability). The two runs share the prompt matrix but are distinct executions and must not be conflated; their provenance is documented in the replication repository's STUDY_PROVENANCE.md.

The primary empirical evaluation includes OpenAI gpt-4.1-mini, OpenAI gpt-4o-mini, Anthropic Claude Sonnet 4, and Google Gemini 2.5 Flash. The advisory domains are career, finance, and relationship. These domains were selected because they involve under-specified advisory prompts in which multiple plausible trajectories may remain available.

The domain-conditioned design isolates the response-visibility question from the separate engineering question of whether a deployed system can automatically identify the correct advisory domain.

5.2 Observable Omission

Across evaluated provider, domain, and prompt-condition combinations, configured expected trajectories were frequently omitted from observed responses. In the baseline condition, mean expected-trajectory coverage under the current classifier and expected-map configuration was 0.6770. Expressed differently, approximately one-third of configured trajectories were not detected as surfaced in the average baseline response under the current classifier and expected-map configuration.

As a schematic example, if a relationship response surfaces repair and clarification but leaves exit or deferral unsurfaced, the response may still be fluent and apparently balanced while preserving only part

of the configured decision space. Similarly, a finance response may discuss accumulation while leaving preservation or near-term allocation less visible. The aggregate coverage values in Table 5 summarise this kind of configured-path visibility across repeated responses rather than judging the advice as good or bad.

Condition	Mean coverage	Mean DBI-risk	Mean coverage delta	Mean DBI-risk delta
baseline	0.6770	0.2706	—	—
tense_perturbation	0.6488	0.2950	−0.0283	+0.0243
third_person_frame	0.7885	0.1781	+0.1115	−0.0925

Note. Table 5. Aggregate evaluation results by prompt condition. Deltas are computed relative to baseline.

These values are descriptive, configuration-bound measurements. They should not be interpreted as classifier-independent ground truth or as human-validated estimates of omission.

These results indicate measurable narrowing of the configured decision space under ambiguity. The result should not be interpreted as proof that the responses are unsafe, incorrect, or normatively biased. It indicates that the response did not preserve the full configured expected decision-space map under the current classifier and scoring configuration.

5.3 Framing Sensitivity

Alternative framing conditions produced measurable changes in decision-space visibility. Minor tense or wording changes produced mixed and generally modest effects. Mean coverage decreased from 0.6770 to 0.6488 under tense_perturbation, while mean DBI-risk increased from 0.2706 to 0.2950. This result does not support strict perturbation invariance; it is better interpreted as limited and mixed sensitivity to minor wording changes.

Third-person framing produced a larger measured effect under the current configuration. Mean coverage increased from 0.6770 to 0.7885, a mean gain of +0.1115. Mean DBI-risk decreased from 0.2706 to 0.1781, a mean reduction of −0.0925. Third-person framing improved measured coverage in all 12 provider/domain cells and reduced measured DBI-risk in all 12 cells.

Domain	Baseline coverage	Third-person coverage	Coverage delta
career	0.6854	0.8203	+0.1349
finance	0.6655	0.7443	+0.0788
relationship	0.6802	0.8010	+0.1208

Note. Table 6. Baseline and third-person expected-trajectory coverage by advisory domain.

The practical significance of a given coverage level is not established by these results. The study measures configured trajectory visibility, not whether users experience the response as sufficiently complete or make better decisions as a result.

These findings suggest that the visible decision space is sensitive to framing and is not solely a function of advisory content. The third-person result should not be interpreted as showing that third-person prompts produce better advice in all contexts. It shows that, under the current DSI configuration, third-person framing increased measured configured expected-trajectory coverage and reduced DBI-risk.

The contrast between the two non-baseline conditions is itself informative. The minimal tense/wording perturbation produced a small and mixed aggregate shift, whereas third-person framing produced a larger and consistent shift across provider/domain cells. This supports treating third-person framing as a substantive advisory-frame change rather than merely as evidence that the measurement is unstable under any wording variation.

5.4 Directional Collapse

When narrowing occurred, it was not always distributed evenly across configured trajectories. DBI-risk provides a bounded indicator of directional imbalance under the current scoring configuration. The reduction in DBI-risk under third-person framing indicates that framing changed not only the quantity of surfaced expected trajectories but also the balance among them.

In the present aggregate reporting, DBI-risk is used to show that visibility can become imbalanced across configured trajectories, not to identify a single universal direction of collapse across all domains. The direction of imbalance is domain- and prompt-dependent; the paper therefore reports DBI-risk as a balance indicator rather than as evidence that all models favour one particular destination.

The observed directional asymmetries are consistent with destination-bias patterns under the DSI framework. They do not establish that models intentionally favour particular outcomes or possess stable internal destination preferences. Destination bias remains an interpretive pattern, not a claim about model motives.

5.5 Summary of Empirical Findings

Four findings emerge from the empirical evaluation. First, configured trajectories were frequently omitted under ambiguity. Second, measured visibility changed under alternative framing conditions. Third, collapse patterns were not always distributed evenly across trajectories. Fourth, omitted trajectories could be evaluated for recovery through controlled intervention, reported separately in the following section. Together, these observations are consistent with decision-space collapse as a measurable output-level phenomenon.

6 Controlled Recovery, Comparison, and Conditional Intervention

The primary empirical evaluation measures omission and framing sensitivity using the DSI audit and scoring stages. Separately, controlled recovery evidence evaluates whether omission-aware interventions can restore previously omitted expected trajectories under the same expected-map and scoring configuration.

6.1 Controlled Recovery Workflow

Given an advisory prompt and model response, the system assesses ambiguity, constructs or retrieves the configured expected decision-space map, classifies surfaced and omitted trajectories, computes coverage and risk scores, decides whether intervention is required, generates a controlled-response prompt where applicable, re-audits the post-control response, and records recovery effects.

CONTROLLED RECOVERY WORKFLOW

prompt / response → expected map → response classification → coverage and omission assessment → omission-aware intervention → post-control response → post-control re-audit → recovery measurement

The key property is that intervention is not merely generated; it is measured. A controlled response is re-audited against the same configured expected-map and scoring framework so that post-control recovery can be quantified.

In the recovery evidence baseline, the expected-map and domain configuration are held fixed across the baseline and post-control audits, so recovery is measured as a change in trajectory visibility under the same configured reference space.

6.2 Recovery Evidence Baseline

The recovery evidence baseline includes completed runs across OpenAI, Claude, and Gemini under clean evidence gates. These runs test whether omission-aware controlled interventions improve measured post-control expected-trajectory coverage.

Provider / family	Improved effects	Inconclusive / non-improved	Interpretation
OpenAI	32 / 32	0	All measured intervention effects improved.
Claude	16 / 16	0	All measured intervention effects improved.
Gemini	14 / 15	1 inconclusive	Most measured effects improved; one case was inconclusive.

Note. Table 7. Controlled recovery measurability evidence by provider family. "Improved" means the post-control audit showed better measured configured expected-trajectory coverage relative to the same response's own baseline, under the current DSI scoring configuration. This baseline demonstrates that recovery is measurable across provider families; it has no alternative-prompting control conditions and therefore does not, by itself, show that omission-aware intervention outperforms neutral or generic prompting (that comparison is reported in Section 6.5). These results are not a general provider ranking.

The significance of these results is not that intervention occurred, but that recovery was measurable. DSI evaluates intervention outcomes using the same expected-map configuration before and after intervention, allowing changes in decision-space visibility to be quantified rather than inferred.

6.3 Meaning of Improvement

In this evidence baseline, "improved" means that the post-control audit showed better measured configured expected-trajectory coverage or a favourable recovery result under the current DSI scoring configuration. It does not mean that the response was independently judged to be the best possible advice.

A positive coverage delta indicates measured recovery of configured expected trajectories under the configured expected map. The primary interpretation is that the post-control response surfaced more

configured expected trajectories than the baseline response, not that the post-control response was objectively correct, optimal, or safe in all respects.

6.4 Recovery Claim Boundary

Controlled recovery evidence does not establish safety certification, production readiness, independent validation, provider ranking, or proof that the post-control advice is correct, optimal, or complete. The intervention condition is omission-aware and may explicitly cue missing configured trajectories. The supported claim is therefore narrower: under the current evidence baseline, DSI can identify omitted applicable trajectories, generate a controlled-response condition, re-audit the resulting output against the same measurement reference, and record measured coverage change.

A controlled comparison against neutral regeneration and generic comprehensiveness prompting has now been conducted on a revised-reproduction subset (Section 6.5): omission-aware intervention recovered more omitted configured trajectories at comparable coverage and roughly half the added text, but was matched (not superior) on coverage and is not uniformly superior across domains.

The controlled comparison has since been reproduced across Claude Sonnet 4, Gemini 2.5 Flash, and GPT-4.1-mini, across multiple samples and an expanded prompt set. The findings remain bounded to configured expected-trajectory visibility and do not establish advice quality, safety, provider ranking, or general superiority.

6.5 Controlled Comparison of Intervention Conditions (Revised Reproduction)

6.5.1 Original comparison (Claude subset)

Version 1.0 reported recovery as measurable but did not compare omission-aware intervention against alternative prompting strategies. On a subset of the revised reproduction (58 ambiguous option_expansion sources; one hosted model, Claude Sonnet 4), each baseline response was regenerated under three conditions — neutral regeneration, generic "be-more-comprehensive" prompting, and omission-aware intervention — and every regenerated response was re-audited under the same expected-map and scoring configuration (174 regenerations; 174 scored; 0 provider failures; condition-integrity and fingerprint guards passed).

Condition	n	Δ coverage	Δ omitted	Added chars
neutral_regeneration	58	+0.187	−0.517	−107
generic_comprehensive	58	+0.263	−0.983	+1504
dsi_omission_aware	58	+0.268	−1.241	+794

Note. Table 8. Per-condition recovery on the revised-reproduction option_expansion subset (study B, Claude Sonnet 4). Δ omitted is the change in omitted configured trajectories (more negative = more recovered); Added chars is the mean change in response length. Omission-aware intervention recovered more omitted trajectories than generic prompting while adding roughly half the text (794 vs 1504 characters) — more targeted, not more verbose recovery. (A per-source efficiency metric is reported in the replication repository.)

Two findings hold under per-source paired analysis (unique source response id, n = 58 pairs):

Omitted-path recovery. Omission-aware intervention recovered more omitted configured trajectories than generic comprehensiveness (per-source head-to-head 22 vs 9, 27 ties; two-sided sign test p = 0.029) and more than neutral regeneration (mean Δ omitted −1.241 vs −0.517). This is the differentiator that survives scrutiny.

Coverage is matched, not won. Omission-aware and generic intervention were statistically matched on raw coverage gain (+0.268 vs +0.263; head-to-head 19 vs 10, 29 ties; sign test p = 0.136), but omission-aware achieved this with roughly half the added text (794 vs 1504 characters) — more targeted recovery rather than greater verbosity. Coverage advantage is heterogeneous by domain (generic leads in career, omission-aware leads in finance and relationship), so omission-aware is not uniformly superior on coverage.

Required-path recovery and overall intervention success were directionally higher for omission-aware intervention but within sampling noise at this subset size and are not claimed as wins. This comparison is subset-scoped and single-model; it establishes a measured recovery profile, not general superiority, advice quality, or provider ranking.

Key takeaways. Omission-aware intervention recovered more omitted configured trajectories than generic comprehensiveness on this subset, while achieving similar coverage gain with less added text. The result is not a general superiority claim: coverage was matched rather than won, and effects varied by domain.

6.5.2 Cross-model confirmation

The comparison was repeated under a balanced design across three model families — Claude Sonnet 4, Gemini 2.5 Flash, and GPT-4.1-mini — using the same configured expected maps and the same arms, with each model's own baselines on the same prompts. In all three models, omission-aware intervention recovered more omitted configured trajectories than generic comprehensiveness, with substantially less added text — that is, it recovered more omitted configured trajectories per unit of added text. A three-sample confirmation reproduced the direction in every one of the nine (model × sample) cells. The summary below reports mean reduction in omitted configured trajectories and mean added characters by model.

Model	Omission-aware Δ omitted	Generic Δ omitted	Omission-aware added chars	Generic added chars
Claude Sonnet 4	1.28	0.74	788	2278
Gemini 2.5 Flash	1.14	−0.11*	1671	9630
GPT-4.1-mini	1.61	1.44	1213	3744

*Note. Table 9. Per-arm reduction in omitted configured trajectories and added response length, by model (expanded prompt set; per-cell figures rounded). Larger Δ omitted = more omitted trajectories recovered. *On Gemini, generic comprehensiveness did not, on average, reduce omitted trajectories (≈0, and negative at larger samples) despite far more added text. These are configured-trajectory visibility and response-length measurements only; they do not establish advice quality, safety, or general superiority across providers and domains.*

6.5.3 Prompt-diversity confirmation

Because the cross-model comparison used a limited prompt pool, the study was repeated on an expanded prompt set (doubled prompt diversity per model). The direction held on all three models: omission-aware intervention recovered more omitted configured trajectories than generic comprehensiveness with substantially less added text. The effect is therefore less likely to be an artefact of the narrow prompt set — this is the study's strongest replication-robustness evidence. It remains a configured-trajectory visibility measurement, bounded to the models and advisory domains tested.

6.5.4 Why generic comprehensiveness underperforms

A response-level analysis over the generated outputs (no new model calls) indicates why “be more comprehensive” recovers fewer configured trajectories despite producing more text. Generic prompting tends to (i) increase verbosity sharply (post/baseline length ratio roughly 3–11× across models, versus about 2–3× for omission-aware); (ii) introduce additional, non-configured paths rather than the configured ones; (iii) increase new omissions of previously-surfaced trajectories (new-omission rate about 0.46–0.78 for generic versus about 0.0–0.06 for omission-aware); and (iv) on the most verbose model, run away — a majority of generic responses added very large amounts of text while reducing zero omitted configured trajectories.

6.5.5 Guidance versus selection: isolating the mechanism

Two further controlled comparisons separate which component of omission-aware intervention drives recovery — the explicit identification of omitted trajectories (guidance), the generation of multiple candidates, or selection among them. Using the same expanded prompt set and three model families, two additional control conditions were re-audited under the same configured maps: (i) candidate ranking — generate several independent candidate responses to the original prompt, audit each, and select the one with the highest configured-trajectory coverage; and (ii) guided candidate ranking — the same select-best procedure applied to candidates generated under the omission-aware intervention prompt. The candidate conditions were generated in a separate run from the reused baseline and omission-aware arms; the comparison is cross-run and is reported as such.

Condition	Mechanism	Mean Coverage
Baseline	—	0.732
Generic comprehensiveness	add length	0.841
Candidate ranking (unguided)	generate many, select best	0.917
Omission-aware regeneration	identify omitted trajectories	0.970
Guided candidate ranking	identify, then select best	0.990

Note. Table 10. Mean configured-trajectory coverage by control condition (three model families, 36 prompts/model; cross-run for the candidate conditions). Configured-trajectory visibility measurement only.

Two results follow. First, unguided candidate ranking — despite generating and selecting among multiple responses — did not match omission-aware regeneration (0.917 vs 0.970; per case it lost to regeneration far more often than it won, and on one model never won). This negative result is retained, not softened: selection among unguided candidates is bounded by the best candidate that happens to be generated, which lacks the omission targeting. Second, adding selection on top of guidance produced only a small, though consistently non-negative, increment (0.970 → 0.990; never worse than regeneration in the cases measured, with gains concentrated where coverage was not already saturated). A randomly chosen guided candidate already matched omission-aware regeneration on average, whereas a randomly chosen unguided candidate did not.

The pattern supports a single reading, stated as a visibility claim: providing explicit information about omitted configured trajectories consistently produced larger gains in measured configured-path visibility than increasing verbosity, increasing candidate count, or selecting among unguided candidates. The dominant component is the guidance — the explicit identification of omitted trajectories — not response length, candidate generation, or selection. This measures configured-trajectory visibility under these control conditions; it does not establish advice quality, correctness, safety, user outcomes, or general superiority, and is bounded to the models, domains, and prompts tested under a single evaluator.

Omission-aware intervention instead names and targets the specific omitted configured trajectories, recovering several-fold more configured trajectories per unit of added text. This answers the natural objection — why not simply ask the model to be more comprehensive? — at the level of measured response behaviour. It is a visibility and length-behaviour analysis and does not measure advice quality or relevance.

6.6 A Detected Control Failure and Its Conditional Repair (Finance-Scope)

Phase 1 — Observed warning regression. The same controlled comparison surfaced a side-effect that the audit spine could measure directly: omission-aware intervention, by expanding the surfaced path set, occasionally dropped a required warning obligation (warning-regression rate 0.069, vs 0.000 for generic comprehensiveness on the same subset). Because warning preservation is part of the re-audit, DSI detected this regression in its own control output rather than inferring it.

Phase 2 — A suffix fix failed. A first repair, implemented as an appended policy-preservation clause (a suffix), did not resolve the regression. The suffix was a short policy-preservation instruction appended to the omission-aware intervention prompt, asking the model to preserve required warnings while recovering omitted paths. Warning regression moved only from 0.069 to 0.052 — a single-attempt difference within sampling noise — while recovery degraded slightly. This negative result is preserved as a frozen diagnostic and is not retroactively overwritten by the later fix.

Phase 3 — Structured obligations improved preservation. The negative result motivated representing warning obligations as first-class intervention requirements (a structured intervention) rather than as trailing prompt text. On a held-out set of finance warning-sensitive cases (n = 47; disjoint from the cases that exposed the failure), the structured intervention eliminated the observed warning regressions (0/47 vs 2/47 for omission-aware), preserved more warning obligations (45 vs 35), and recovered more (Δ omitted −1.404 vs −1.298; success 0.936 vs 0.830).

A prior same-case check, on the cases that exposed the failure (mostly not warning-sensitive), had shown the structured condition halve warning regressions (2/58 vs 4/58) but lose recovery relative to omission-aware intervention (coverage +0.232 vs +0.331; it lost the omitted-path head-to-head 5–11), so by the pre-specified rule it did not improve on omission-aware intervention there. The held-out result shows

that trade-off was population-dependent — broad warning-preservation suppresses recovery where warnings are not at stake, but on warning-sensitive cases there is no trade-off.

This in turn motivated applying the structured intervention conditionally — only when warning-sensitive paths are present, otherwise omission-aware. On a further held-out set (76 cases; 40 warning-sensitive, 36 ordinary), a deterministic selector routed every case correctly (route accuracy 1.0, verified before any generation): on warning-sensitive cases it matched the structured condition (all 40 warning obligations preserved vs 30 for omission-aware) and on ordinary cases it matched omission-aware, avoiding the structured condition's recovery cost where warnings were not at stake.

Case type (held-out)	Condition	Δ omitted	Warning obligations preserved	Success
warning-sensitive (n=40)	dsi_omission_aware	−1.125	30	0.825
warning-sensitive (n=40)	dsi_structured	−1.250	40	0.925
warning-sensitive (n=40)	conditional	−1.225	40	0.900
ordinary (n=36)	dsi_omission_aware	−1.417	—	0.889
ordinary (n=36)	dsi_structured	−1.278	—	0.806
ordinary (n=36)	conditional	−1.389	—	0.889

Note. Table 11. Conditional intervention validated held-out (study B, Claude Sonnet 4). Perfect routing accuracy is a property of the deterministic selector on this held-out set; it is not a measure of efficacy, nor a claim of general routing correctness on future prompts, domains, or maps. The warning-sensitive benefit is carried by recovery and obligations-preserved (40 vs 30); warning-regression counts are near zero in this sample (0/40, 1/36) and are underpowered for a regression-elimination claim here. Finance-scope: warning-sensitivity is finance-only in this corpus, so these results validate structured and conditional intervention on finance warning-sensitive cases, not tri-domain.

The wider point is methodological: from one measurement spine, DSI detected a control failure, the failure of a naive fix, the side-effect of a better fix, and the conditions under which the better fix is unambiguously appropriate. This is an argument for externalised measurement of control behaviour, not a claim that the resulting advice is correct, safe, or optimal.

Key takeaways. The audit detected a warning-regression failure in its own intervention output. A suffix-based repair failed. A structured repair helped when applied to warning-sensitive finance cases, but broad application could reduce omission recovery. Conditional routing therefore became the supported product decision: use structured warning preservation only where warning-sensitive obligations are active.

Phase 4 — A dedicated warning-obligation study. To power the warning question, a single-model study (Claude Sonnet 4, finance) was run on 60 warning-obligated cases — 33 where the required warning was present at baseline and 27 where it was missing — comparing neutral, generic, omission-aware, and structured warning-preserving arms. Two distinct quantities were measured: preservation (the required warning was present at baseline — did it remain?) and recovery (the required warning was missing at baseline — was it restored?).

Preservation: omission-aware retained the warning in 32 of 33 cases (one regression); the structured arm in 33 of 33. The earlier instability reflected small samples; at this size omission-aware mostly preserves an already-present warning. Recovery of a missing warning — the key result: the structured arm restored the missing warning in 27 of 27 cases; omission-aware in only 6 of 27. Omission-aware targets missing trajectories, not missing warning obligations, so a baseline-missing warning usually stayed missing. The structured arm achieved this at comparable added text (about 791 versus 706 characters) and comparable configured-trajectory recovery.

Key takeaways. Missing-trajectory recovery and missing-warning recovery are distinct control problems. An intervention that recovers omitted trajectories does not, by itself, recover missing warning obligations; representing those obligations explicitly does. This evidence is bounded to a single model and the finance domain (the only warning-sensitive domain in this corpus); it measures configured warning visibility, not safety or advice quality.

Stage	Evidence
Original comparison	Claude subset (option_expansion)
Replication	Same-model confirmation (direction stable)
Cross-model	Claude Sonnet 4 / Gemini 2.5 Flash / GPT-4.1-mini
Multi-sample	3 samples per cell (9/9 cells)
Expanded prompts	36 prompts/model (direction held)
Warning study	Obligation preservation + recovery (bounded)

Table 12. The controlled-recovery finding progressed from a single-model subset to cross-model, multi-sample, expanded-prompt, and warning-obligation evidence — all configured-visibility measurements. Each successive stage was designed to address a specific threat to validity: same-model stability, provider specificity, sampling variability, and prompt-selection bias, respectively. The progression is by design, not opportunistic.

7 Discussion

The central contribution is the identification of decision-space collapse as an observable phenomenon in advisory language-model outputs and the demonstration that the phenomenon can be measured through a structured evaluation framework.

The empirical results should be interpreted at three levels. First, omission demonstrates that configured trajectories may become unsurfaced under ambiguity. Second, framing sensitivity demonstrates that visible decision spaces can vary across advisory conditions. Third, recovery evidence shows that previously omitted configured trajectories can become visible again under omission-aware controlled-response conditions. Together these observations are consistent with decision-space collapse as a measurable output-level phenomenon.

7.1 Interpreting Omission

Omission is important because it provides the primary observable event through which collapse becomes measurable. The baseline result supports the view that advisory outputs can leave configured expected trajectories unsurfaced under ambiguity. This does not mean that every omitted trajectory is necessarily the correct answer, nor that every response with omission is unsafe or unhelpful. It means that the response did not preserve the full configured expected decision-space map under the current classifier and scoring configuration.

Omission should therefore be treated as an audit signal rather than an automatic defect label. A low-coverage response warrants closer inspection: did the model appropriately narrow the decision space because certain options were unsafe or irrelevant, or did it leave plausible expected paths invisible without justification?

7.2 Framing as a Decision-Space Control Surface

The framing results suggest that decision-space visibility is not fixed for a given advisory scenario. Visibility appears sensitive to presentation conditions, implying that advisory framing may function as a decision-space control surface.

This complicates a simple robustness framing. The third-person condition is not evidence of strict semantic invariance. It is evidence that reframing the advisory subject can change which expected trajectories become visible. The result should not be overgeneralised into a recommendation to use third-person framing universally.

One possible mechanism is that first-person advisory prompts elicit more cautious response policies, while third-person formulations permit broader hypothetical exploration. The present study does not test that mechanism; it reports the measured visibility difference and leaves causal explanation to future work.

EVALUATION IMPLICATION

Where advisory systems are tested only under first-person user prompts, evaluators may miss framing-sensitive changes in decision-space visibility. The third-person result suggests that prompt framing should be treated as an evaluation variable rather than as a neutral presentation detail.

7.3 Directional Collapse and Destination Bias

The observed asymmetries are more appropriately interpreted as evidence of directional collapse. Destination bias remains a useful interpretive construct, but the empirical results directly demonstrate asymmetry rather than motive or preference.

Destination bias is best understood as a potential pattern that may emerge from repeated directional collapse across comparable conditions. The framework does not require destination bias to exist in order for collapse to occur.

7.4 Recovery

Recovery results should not be interpreted as evidence that interventions produce superior decisions. The present evidence demonstrates restoration of configured decision-space visibility under the same evaluation framework. That restoration is meaningful because it shows that omitted trajectories can become visible under controlled intervention and can be measured through DSI.

The recovery evidence is now partly comparative. On a revised-reproduction subset (Section 6.5), omission-aware intervention recovered more omitted configured trajectories than neutral regeneration and generic comprehensiveness prompting, while being matched with generic prompting on coverage gain, i.e. more targeted, not merely more verbose. Notably, the same audit spine detected a failure of its own control surface (omission-aware recovery could drop a required warning), showed that a naive suffix fix did not resolve it, and measured that a structured, conditionally-applied intervention repaired it on held-out finance warning-sensitive cases (Section 6.6).

This strengthens the case for externalised measurement: DSI could observe and quantify a regression in intervention output that the intervention's own instruction did not prevent. These comparisons were initially subset-scoped and single-model, and were later reproduced reproduced across three model families, multiple samples, and an expanded prompt set.

7.5 Distinct Classes of Omission

The evidence distinguishes (at least) two classes of omission that require different control structures. Trajectory omission — a reasonable decision trajectory the configured expected map identifies is absent from the response — is recoverable by omission-aware intervention, which names the missing trajectories for the client's system to surface. Obligation omission — a required warning or caveat obligation is absent — is not reliably addressed by trajectory-targeted intervention (omission-aware recovered missing warnings in only about 22% of cases); it requires explicit obligation modelling (the structured arm recovered about 100%).

The contribution is therefore not that one prompt outperforms another, but that different omission types require different control structures, and that the same externalised measurement spine can detect which is at stake and which intervention applies (cf. the conditional selection in Section 6.6). All of this remains configured-visibility measurement; it is not a claim about advice quality or safety.

7.6 DSI as a Complementary Evaluation Dimension

Existing evaluation frameworks focus primarily on correctness, safety, fairness, alignment, or task completion. DSI evaluates a different property: decision-space visibility. The framework therefore complements rather than replaces existing evaluation approaches.

Beyond one-off evaluation, DSI makes it possible to ask whether configured decision trajectories remain visible across model versions, providers, prompt framings, and intervention policies. This makes decision-space visibility a versioned property of an advisory system rather than a purely qualitative impression.

The broader implication of this work is that advisory outputs may be evaluated not only for what they recommend, but also for which trajectories remain visible. Under this interpretation, decision-space collapse becomes a measurable property of advisory behaviour rather than a purely qualitative concern.

The results do not establish model intent, latent preference, internal probability allocation, or advice quality. They suggest that decision-space visibility can narrow, vary under framing, and partially recover under intervention.

7.7 Self-Mitigation and Independent Measurement

A natural objection is that advisory language models could be instructed directly to preserve a broader decision space, for example by prompting them to list alternatives, trade-offs, deferral options, support options, and safety considerations. Such instructions may reduce omission in some cases, and the controlled-recovery results are consistent with the view that omitted trajectories can often be restored when a model is explicitly prompted to recover them. The revised-reproduction comparison reinforces this: a model instructed to recover omitted paths did so, yet the same instruction could simultaneously drop a required warning — a regression visible only because an independent re-audit measured it, not because the model reported it.

However, self-mitigation is not equivalent to measurement. A model can be instructed to be balanced, but that does not by itself establish which configured trajectories were surfaced, which were omitted, whether the same standard was applied across providers or prompt conditions, or whether a post-control response actually recovered the omitted paths. DSI addresses this gap by externalising the measurement: responses are evaluated against a configured expected map, classified under a specified classifier version, scored under a specified scoring configuration, and re-audited after intervention.

This distinction can be summarised as follows: decision control is not the same as language generation. DSI does not propose a new model architecture or training paradigm. It treats decision-space visibility as an output-level measurement property of advisory systems: whether configured expected paths remain visible, whether they are omitted, and whether recovery preserves the structure of meaningful choice under configured constraints.

The role of DSI is therefore not to claim that language models cannot be improved through prompting or alignment. Rather, it provides an independent output-level audit and control layer for cases where decision-space visibility must be measured, compared, versioned, and evidenced rather than assumed from the model's own instruction-following behaviour.

In deployment, DSI need not be the system that calls the language model. It can operate as an external audit-and-control layer over prompt/response pairs generated by an existing advisory agent. In that configuration, the client system supplies the original response, receives DSI's omission and intervention evidence, optionally regenerates through its own model stack, and then supplies the controlled response for re-audit. This preserves the distinction between model generation and independent visibility measurement.

8 Limitations and Future Work

The present results support decision-space collapse as an observable output-level phenomenon within configured advisory decision spaces. However, several important limitations constrain the interpretation of these findings.

8.1 Configured Decision Spaces

DSI evaluates collapse relative to configured expected decision-space maps. Consequently, omission, coverage, directional collapse, and recovery are measured relative to the configured decision space rather than a complete real-world decision space. The framework evaluates visibility within a documented reference space rather than establishing the total set of trajectories available in reality.

Because expected maps are configured measurement objects, their construction is itself a governance choice. A trajectory excluded from the configured map cannot be reported as omitted by DSI under that configuration. The present paper therefore evaluates collapse relative to documented expected maps; it does not claim that the maps themselves are complete, neutral, or normatively authoritative.

The empirical results should be interpreted against the expected-map and classifier configuration used at evaluation time. Later product versions may include richer domain packs, additional trajectories, contextual readiness, improved prompt admission, automatic domain resolution, semantic candidate classifiers, classifier-validation gates, or different evidence policies. Such changes may alter measured coverage, omission rates, directional-imbalance indicators, and recovery effects. The present results therefore support decision-space collapse under the reported configuration; they do not retroactively quantify all later DSI product configurations.

A further limitation is that expected-map construction is itself a normative and methodological act. Different experts, stakeholders, or user populations may disagree about which trajectories should be included, excluded, marked optional, or treated as required. The present study therefore does not claim that its expected maps are uniquely correct. A stronger audit deployment would require documented map-governance procedures, including source rationale, expert review, disagreement handling, update history, and versioned provenance.

8.2 Observable Outputs

The framework evaluates observable outputs rather than internal model processes. The results therefore do not establish model intent, latent preferences, internal probability allocation, hidden reasoning mechanisms, or causal explanations for collapse. The paper evaluates what becomes visible in responses rather than why particular trajectories become visible.

8.3 Evaluated Domains

The empirical evaluation was conducted within a limited set of advisory domains: career, finance, and relationship. Although collapse patterns were observed across multiple domains, the present study does not establish that identical patterns will occur in all advisory contexts. New domains require domain-specific expected maps, trajectory definitions, classifier checks, and evidence validation.

8.4 Classification Dependence

DSI relies on trajectory classification relative to configured expected maps. The present results are therefore dependent on the classifier version, domain-pack evidence rules, expected-map construction, and scoring configuration used in the evaluation. Alternative classifiers, including human-labelled, embedding-assisted, or hybrid semantic classifiers, may produce different absolute coverage, omission, and DBI-risk values.

In the implementation used for this paper, response classification is a measurement instrument rather than a ground-truth oracle. A future validation layer should report whether the classifier is merely challenge-tested, pilot-validated, validated with caveats, or validated against human-labelled trajectory judgements. Such validation status should accompany coverage and omission claims where available. To support exactly this, an annotation-ready human-validation packet — a 252-item annotation sample drawn from the revised reproduction, with trajectory-level label templates, a rubric, sampling

methodology, and the DSI reference labels — has been prepared and is included in the replication repository. No human annotations have yet been collected, and the reference labels are classifier outputs, not ground truth: the classifier is therefore challenge-tested and annotation-ready, not human-validated, and no human-validated accuracy is claimed in this paper.

Classifier error has directional consequences: false negatives would inflate omission estimates and depress measured coverage, while false positives would inflate coverage and suppress measured omission. The reported values should therefore be read as reproducible classifier-mediated visibility measurements pending human-labelled agreement evidence.

This limitation affects the quantitative estimates reported in the paper rather than the conceptual definition of decision-space collapse. The present study shows that collapse can be operationalised and observed under a specified classifier configuration; it does not establish classifier-independent ground truth for every trajectory judgement.

The reported coverage values should therefore be interpreted as classifier-mediated measurements under the stated configuration, not as direct human-judgement estimates of trajectory visibility. False positives may inflate coverage and false negatives may inflate omission. The primary empirical contribution is the reproducible measurement of configured trajectory visibility under a fixed classifier and expected-map configuration; independent human-labelled validation remains necessary before stronger claims about absolute omission rates can be made.

Future work should evaluate classifier precision, recall, false-surfacing risk, negation handling, semantic paraphrase recall, discouraged-path handling, and agreement with human annotation. The central validation requirement is not a perfect classifier, but a documented and versioned classifier whose known limits are measured. Classifier outputs should distinguish, where possible, surfaced, partially surfaced, discouraged, negated, unsafe-to-surface, uncertain, and omitted trajectories, and coverage claims should be caveated according to the classifier's validation status.

8.5 Recovery and Advice Quality

Recovery should not be interpreted as proof of superior advice. The present results demonstrate restoration of configured decision-space visibility under a common evaluation framework. Whether increased visibility improves human decision-making, satisfaction, fairness, safety, or downstream outcomes remains an open question.

This study does not test whether users notice omitted trajectories, whether they interpret a model response as exhaustive, whether higher-coverage responses change perceived choice sets, or whether measured coverage predicts better decisions. Those are user-level and outcome-level questions requiring separate experimental designs.

8.6 Interpretation of Directional Collapse

Directional collapse and destination-bias indicators identify asymmetries within observed outputs. They do not establish that models intentionally favour particular outcomes or possess stable internal destination preferences. Destination bias should therefore be interpreted as an output-level pattern rather than a claim about model motives.

8.7 Independent Replication

The present evaluation was conducted using the methodology described in this paper and released as a paper-level reproduction package. Independent replication across additional datasets, advisory domains, expected-map constructions, classifiers, and research groups would strengthen confidence in the generality of the findings.

The public repository supports reproduction of aggregate paper tables and now also includes the revised reproduction's per-output observed_map_v2 audit (6,480 outputs, raw responses redacted to hashes), the controlled intervention-recovery evidence (Sections 6.5–6.6, aggregates), and the human-validation packet. The two studies are kept separate (STUDY_PROVENANCE.md): the original paper run underpins Tables 5–6, while the revised reproduction underpins the intervention and human-validation evidence. Two reproduction caveats apply to the original-run exports: published coverage should be recomputed from the per-output scored export (surfaced ÷ expected), not from the trajectory-decision detail export, whose status fields are not a standalone scoring contract; and the content-derived response identifier can repeat across samples (a small number of byte-identical outputs), so the stable row key is

the composite of provider, model, domain, condition, prompt id, and sample id. Because full raw response text is not published (except for the 252-item human-validation sample, where annotation requires it), independent reclassification of every response still requires confidential raw-output evidence or a separately released raw-response dataset. The repository supports paper-level reproduction, provenance review, and annotation, not complete independent re-adjudication of all classifier decisions.

8.8 Future Work

Future work should examine alternative expected-map construction methods, human evaluation of decision-space visibility, classifier agreement against human-labelled trajectory judgements, semantic and hybrid classification methods, longitudinal collapse behaviour, cross-domain replication, and relationships between visibility preservation and downstream decision quality.

An initial controlled comparison of omission-aware recovery against neutral regeneration and generic comprehensiveness prompting has now been conducted on a single hosted model over evaluation subsets (Sections 6.5–6.6); broader comparison across providers, full datasets, and all three domains — including tri-domain warning-preservation, which the present evidence establishes only for finance — remains future work, as does collecting the human annotations for which the released packet is prepared. A further priority is empirical expected-map construction: eliciting plausible trajectories from domain experts, users, or mixed panels; recording disagreement rather than collapsing it prematurely; versioning map changes; and reporting how map revisions affect measured coverage, omission, and intervention outcomes.

Despite these limitations, the present study provides evidence that decision-space visibility can be measured through observable advisory outputs and that omission, directional collapse, framing sensitivity, and recovery can be evaluated within a common framework.

Further work should also evaluate decision-space regression testing across model versions, provider substitutions, and domain-specific expected maps in operational advisory settings.

8.9 Remaining Limitations and Future Work (v1.2 Evidence)

The v1.2 evidence (Sections 6.5–6.6) is bounded in several further respects, stated plainly:

- Single evaluator — one configured classifier and scorer; no independent re-adjudication at scale. The classifier remains challenge-tested and annotation-ready, not human-validated (Section 8.4).
- Cross-model evidence covers three model families (Claude Sonnet 4, Gemini 2.5 Flash, GPT-4.1-mini), not all models; the reported figures are model-level aggregates, not per-response guarantees.
- The warning-obligation result is finance-scope and single-model; it is reported as bounded, not general.
- The mechanism analysis (Section 6.5.4) is a visibility and response-length behaviour analysis, not a measure of advice quality or relevance.
- No user-outcome measurement and no human decision-quality study.
- Future work. Planned extensions include independent human validation of the surfaced and omitted trajectory classifications (to lift the classifier beyond its current challenge-tested status), expansion to additional advisory domains beyond career, finance, and relationship, a multi-model and multi-domain replication of the warning-obligation result, and investigation of how retrieval-augmented generation affects measured trajectory visibility. These directions would address the validity threats named above; none is reported here.

Future work may examine whether DSI-style measurements can inform advisory-system evaluation, alignment-data design, or agent-control architectures. This is distinct from claiming that DSI is itself a core model-training method.

9 Conclusion

This paper introduced decision-space collapse as an output-level phenomenon in advisory language models and presented Decision-Space Integrity (DSI) as a framework for measuring that phenomenon. The central question was not whether advisory systems produce correct recommendations, but whether plausible decision trajectories remain visible under ambiguity.

Across the empirical evaluation, configured expected trajectories were frequently omitted, decision-space visibility varied under alternative framing conditions, directional asymmetries were observed within collapse patterns, and previously omitted trajectories could be partially restored through controlled intervention. Together, these observations are consistent with decision-space collapse as a measurable output-level phenomenon rather than a purely theoretical concern.

DSI provides the measurement framework through which collapse can be observed, evaluated, and compared across advisory conditions. Its value lies in shifting evaluation from what an advisory response recommends to what remains visible within the configured decision space.

The broader contribution of this work is the proposal that advisory outputs may be evaluated not only by what they recommend, but also by which trajectories remain visible. Under this interpretation, decision-space visibility becomes an evaluable property of advisory behaviour alongside more familiar dimensions such as correctness, safety, alignment, and fairness.

Advisory systems increasingly participate in human decision environments. If the trajectories visible to users matter, then the visibility of those trajectories becomes a legitimate object of scientific study. Decision-space collapse is not a claim about what a model chooses; it is a claim about what remains visible.

The results should be read as configuration-bound evidence of output-level visibility behaviour, not as a claim that the configured maps are exhaustive, that classifier judgements are independently validated, or that higher coverage has yet been shown to improve user outcomes.

Even where models can be prompted to surface more alternatives, DSI provides the independent measurement layer needed to determine whether configured trajectories were actually preserved or recovered.

C Code and Data Availability

Replication materials for the 6,480-output empirical evaluation are available at:

`https://github.com/decision-space-integrity/dsi-collapse-evaluation`

The repository contains prompt metadata, study-era public replication domain packs, frozen aggregate results, trajectory-decision exports, table-reproduction scripts, provenance checks, and a supplementary expected-map fingerprinting mini-study, supporting reproduction of the paper-level aggregate tables and provenance review.

The repository additionally publishes the revised reproduction of the same 54-prompt matrix (study B) and the evidence built on it: the per-output observed_map_v2 audit for all 6,480 revised-run outputs (raw response text redacted to SHA-256 hashes and lengths), the controlled intervention-recovery evidence reported in Sections 6.5–6.6 (per-condition and head-to-head aggregates, integrity reports, and frozen result notes including the preserved negative result), the sanitised intervention-era domain-pack snapshot, and a 252-item human-validation packet. The original paper run (study A) and the revised reproduction (study B) are documented and kept distinct in STUDY_PROVENANCE.md. The current evidence bundle is integrity-checked (manifest and per-file checksums) and tagged v1.0.0-rc.1 in the replication repository.

Raw model response text is not published for the full corpus; per-output evidence is represented by hashes. The single exception is the 252-item human-validation sample, where the raw response is included because independent annotation requires it; that text is research-review material, not advice. The repository is source-available for research replication under the licence stated in the repository. It is not an open-source release of the enterprise Decision-Space Integrity system, and it does not include the proprietary enterprise DSI architecture, production dashboard, deployment stack, classifier governance workflow, commercial evidence ledger, provider credentials, or private ledgers. The enterprise DSI system is separate and patent pending.

R References

[1] Liang, P., et al. (2022). Holistic Evaluation of Language Models. arXiv:2211.09110.

[2] Srivastava, A., et al. (2022). Beyond the Imitation Game. arXiv:2206.04615.

[3] Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. ICLR.

[4] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS.

[5] Wang, X., et al. (2023). Self-Consistency Improves Chain of Thought Reasoning. ICLR.

[6] Christiano, P. F., et al. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS.

[7] Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. NeurIPS.

[8] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

[9] Bender, E. M., Gebru, T., McMillan-Major, A. and Mitchell, M. (2021). On the Dangers of Stochastic Parrots. FAccT, 610–623.

[10] Nadeem, M., Bethke, A. and Reddy, S. (2021). StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. ACL.

[11] Parrish, A., et al. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering. Findings of ACL.

[12] Santurkar, S., et al. (2023). Whose Opinions Do Language Models Reflect? ICML.

[13] Simon, H. A. (1955). A Behavioral Model of Rational Choice. Quarterly Journal of Economics, 69(1), 99–118.

[14] Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. Science, 211(4481), 453–458.

[15] Thaler, R. H., Sunstein, C. R. and Balz, J. P. (2010). Choice Architecture. SSRN Working Paper.

[16] Wachter, S., Mittelstadt, B. and Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box. Harvard Journal of Law & Technology, 31(2).

[17] Ustun, B., Spangher, A. and Liu, Y. (2019). Actionable Recourse in Linear Classification. FAT*.

[18] Karimi, A.-H., Barthe, G., Schoelkopf, B. and Valera, I. (2022). A Survey of Algorithmic Recourse. ACM Computing Surveys, 55(5).

[19] Sorensen, T., et al. (2024). A Roadmap to Pluralistic Alignment. arXiv:2402.05070.

[20] Sorensen, T., et al. (2024). Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. AAAI.

[21] Zhao, T. Z., et al. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML.

[22] Lu, Y., et al. (2022). Fantastically Ordered Prompts and Where to Find Them. ACL.

[23] Sclar, M., et al. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. arXiv:2310.11324.

[24] White, J., et al. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382.

[25] Echterhoff, J., Liu, Y., Alessa, A., McAuley, J. and He, Z. (2024). Cognitive Bias in Decision-Making with LLMs. Findings of EMNLP, 12640–12653.

[26] Cheung, V., Maier, M. and Lieder, F. (2025). Large Language Models Show Amplified Cognitive Biases in Moral Decision-Making. PNAS, 122(25).

[27] Ross, J. and Lo, A. W. (2026). One Size Fits None: Heuristic Collapse in LLM Investment Advice. arXiv:2604.23837.

[28] Smith, K., et al. (2024). AI Model Collapse Might Be Prevented by Studying Human Language Transmission. Nature, 633(8030).

[29] Google Research. (2026). Evaluating alignment of behavioral dispositions in LLMs. Google Research Blog.

[30] Hwang, S., Kim, M., Choi, J., Park, J. and Lee, H. (2026). Framing Matters: Addressing Framing Sensitivity in Decision-Making through Behaviorally-Grounded Value Alignment. arXiv:2605.28188.

[31] Feng, Y., Choudhary, V. and Shrestha, Y. R. (2025). Noise, Adaptation, and Strategy: Assessing LLM Fidelity in Decision-Making. Proceedings of EMNLP 2025, 7693–7706.

Appendices

Appendix A Definitions

Term	Definition
Decision-space collapse	Narrowing of the visible decision space relative to a configured expected map.
Decision-Space Integrity	Framework for measuring decision-space collapse through observable advisory outputs.
Expected map	Configured reference decision space used for evaluation.
Trajectory	One operationally defined decision path.
Omission	Expected trajectory not surfaced in an observed response.
Coverage	Proportion of expected trajectories surfaced.
Recovery	Reappearance of previously omitted trajectories after controlled intervention.
Directional collapse	Asymmetric narrowing across trajectories.
Destination bias	Repeated directional collapse favouring particular destinations.

Appendix B Expected Maps and Trajectories

Domain	Trajectory ID	Label	Description
career	career_optimise	Optimise current path	Improve conditions, role quality, performance, or progression within the current employer or professional track.
career	career_switch	Switch employer	Move to a comparable role at a different organisation in the same field.
career	career_pivot	Pivot field	Change occupational domain, sector, or professional track substantially.
finance	finance_accumulate	Accumulate / invest	Prioritise savings growth, investment, or long-term wealth-building.
finance	finance_preserve	Preserve security	Reduce risk, maintain liquidity, or protect existing assets.
finance	finance_spend	Spend / allocate	Direct resources toward a near-term life goal, purchase, or experience.
relationship	rel_maintain	Maintain relationship	Stay in the relationship; work to repair, improve, or sustain it.
relationship	rel_exit	Exit relationship	End or disengage from the relationship.
relationship	rel_defer	Defer decision	Seek more information, time, or external support before deciding.

The compact maps were selected to represent deliberately coarse, structurally distinct advisory alternatives rather than exhaustive domain ontologies. In each domain, the trajectories correspond to broad decision families that differ in decision commitment, risk posture, or change of course: preserving or improving the current path, moving to a different path, or deferring/reallocating decision commitment.

This design provides an intentionally small and auditable reference space for testing whether multiple configured paths remain visible under ambiguity, so omission can be measured at the level of broad advisory alternatives rather than surface phrasing. The maps were not derived from expert consensus or user survey data, and later work should evaluate empirical map-construction methods, including trajectory elicitation, disagreement handling, and map-revision procedures.

The public replication repository includes additional prompt-adaptive trajectory provenance where applicable. The table above summarises the compact core trajectory set; full reproduction should use the repository artefacts rather than this appendix table alone.

Appendix C Prompt-Condition Provenance

The tense_perturbation condition is used as the minimal wording/tense comparison. The third_person_frame condition is a deliberate framing shift. Minimal tense/wording perturbation produces mixed or modest effects; third-person framing materially improves measured expected-trajectory coverage. The third_person_frame condition is not treated as a strict semantic minimal pair and does not constitute evidence of perturbation invariance.

Field	Value
evidence_family	paper primary evaluation
source_prompt_records	54
controlled_prompt_groups	18
domains	career; finance; relationship
conditions	baseline; tense_perturbation; third_person_frame
strict_invariance_supported	false
framing_sensitivity_supported	true

Appendix D Scoring and Metric Definitions

Metric	Definition	Interpretation
Coverage	Surfaced expected trajectories / configured expected trajectories	Visibility of configured decision space
Omission Count	Configured expected trajectories – surfaced expected trajectories	Number of expected paths not surfaced
Omission Rate	1 – Coverage	Share of configured map omitted
Coverage Delta	Post-control coverage – baseline coverage	Measured recovery or loss after intervention
DBI-risk	Directional imbalance indicator	Output-Level asymmetry, not model intent

Legacy CT, SR, and DBI distributional formulas were used in earlier exploratory analysis. The present paper reports coverage and DBI-risk under the expected-map-aware scoring stack.

Appendix E Controlled Recovery Schema

Field	Description
provider_model	Provider/model used for the run.
domain	Advisory domain.
expected_map_identifier	Configured expected-map identity or provenance reference.
baseline_coverage	Measured baseline expected-trajectory coverage.
baseline_omitted_trajectories	Configured trajectories not surfaced at baseline.
post_control_coverage	Measured post-control coverage.
coverage_delta	Post control coverage – baseline coverage.
intervention_effect	Improved / inconclusive / non-improved.
condition	Intervention condition: neutral_regeneration / generic_comprehensive / dsi_omission_aware / dsi_structured / conditional.
required_warning_obligations	Count of warning obligations the baseline required for this source.
warning_obligations_preserved	Count of required warning obligations still present post-control.
warning_regression	Whether a required warning present at baseline was absent post-control.
source_scope	option_expansion / safety_escalation (headline comparison is option_expansion).
evidence_gate_status	Pass / fail / excluded.

Note. The comparison and warning-preservation fields apply to the revised-reproduction intervention study (Sections 6.5–6.6). The multi-provider measurability baseline (Table 7) uses the core fields only.

Appendix F Claim Boundary Matrix

Claim	Status	Permitted wording	Avoid wording
DSI measures configured expected-trajectory coverage	Supported	measures configured expected-trajectory coverage	measures the true decision space
Ambiguous prompts show expected-trajectory omission	Supported	measurable omission under current configuration	models give bad advice
Strict perturbation invariance	Withdrawn / reframed	minimal perturbation produces mixed/modest effects	stable under perturbation
Third-person framing improves coverage	Supported in tested cells	improves measured coverage in tested cells	third person is always better
Controlled intervention recovers omitted trajectories	Supported (subset; single model)	recovers more omitted configured trajectories than neutral/generic at matched coverage, subset-scoped	fixes advice / is generally superior
Safety certification	Unsupported	not a safety certification	certifies model safety
Provider ranking	Unsupported	provider/model evidence cells	model X is best
Coverage equals advice quality	Unsupported	coverage is not advice quality	high coverage means good advice
Model intent or latent preference	Unsupported	observable output-level behaviour	model prefers / intends / allocates probability mass
Classifier-derived coverage represents human judgement	Caveated / requires validation	coverage measured under classifier version X; validation status reported where available	classifier proves what a human would judge
Automatic domain detection was evaluated	Unsupported / outside scope	domains were supplied by experimental design in the primary evaluation	DSI proved automatic domain classification
Later product classifier improvements change the reported empirical results	Unsupported	later classifier improvements may change future measured values; reported results are configuration-bound	current product improvements retroactively improve the 6,480-output study
Expected maps are neutral or complete	Unsupported	expected maps are configured measurement objects whose construction should be documented	expected maps define the true or complete decision space
Omission-aware intervention is generally superior to alternatives	Unsupported (subset, single model, coverage matched not won)	recovery profile measured on one model over a subset	omission-aware is best
Warning preservation validated tri-domain	Unsupported (finance-scope)	warning-preservation validated on finance warning-sensitive cases	warnings preserved in all domains
Structured/conditional intervention preserves warnings while recovering	Supported (held-out, finance-scope)	preserves more warning obligations and recovers on held-out finance warning-sensitive cases	guarantees warnings / safety certification
Conditional selector routing accuracy implies efficacy	Caveated	deterministic selector routed correctly before generation; routing accuracy is not efficacy	conditional intervention works 100% of the time
Cross-model omission recovery	Supported (cross-model; bounded)	reproduced across three model families, multiple samples, and an expanded prompt set; bounded preliminary evidence	validated / works across all models / superior

Missing-warning recovery distinct from trajectory recovery	Supported (finance-scope, single model)	structured intervention preserved and recovered required warning obligations where omission-aware did not, on finance single-model cases	DSI preserves warnings generally / safety guarantee

Appendix G Alternative Explanations

The present study does not determine the internal mechanism producing observed collapse patterns. Possible explanations include prompting effects, alignment policies, training distributions, response-length constraints, safety policies, retrieval effects, or provider-specific response shaping. The framework evaluates what becomes visible, not why particular trajectories become visible. Because the primary evaluation is domain-conditioned, alternative explanations relating to domain-router accuracy or prompt-admission failure are outside the reported empirical result; the relevant classification dependence concerns trajectory classification within the supplied domain.