Cross-Pollination and Crop Rotation: A Field Guide for Agentic Systems Research

Adrian de Wynter
The University of York
Correspondence: adrian.dewynter@york.ac.uk

Abstract

Contemporary agentic systems blend generative AI models with software components (e.g., tools, memory, orchestration). This makes them capable, but failure-prone. In this work we build a 'map' of where the field is concentrated or converging, and where it could improve or cross-pollinate, with respect to *open problems and negative results*. We conduct a systematic meta-review of recent agentic systems literature, leveraging the fact that most contemporary articles list their limitations and motivate themselves via an open problem. We label papers by their methodological practices (e.g., human and LLM evaluators, calibration) and characterise their reported negative results by type, claimed cause, and proposed fix. Our central finding is a systematic *rigour-prevalence mismatch*: the negative types most frequently reported across the field–underlying model capabilities, measurement, and workflow robustness–are addressed with weaker evidence than less-prevalent types such as safety or multimodality. We further show that subareas with higher methodological rigour also exhibit tighter type-cause-fix relationships, suggesting that rigour and problem convergence co-evolve. We close with concrete recommendations for diversifying research directions and strengthening evaluation practices in this field.

1 Introduction

The study of agentic systems is not new, nor constrained to AI. Some could argue that AI is a subfield of agentic system research (Wooldridge, 2009; Shoham and Leyton-Brown, 2008); or, depending on the definition of the 'I' in AI, its entire goal (Russell and Norvig, 2003). Nonetheless, the consensus is that an agentic system must be autonomous to a degree, yet not necessarily fully 'intelligent' or knowledge-forward. Hence, much of agentic system development involves its integration with other platforms, rather than solving open problems in AI (Etzioni 1996; also Section 2.1).

That said, adding easily 'programmable' (primarily via in-context learning) generative AI systems to software platforms has further entangled both fields. When discussing an artificial agent (e.g., a chatbot), it is now common to refer to it as the model *with* extra capabilities, such as code generation, and requirements–say, decision-making.

This entanglement has driven major changes in how we think of AI, and expanded the applications of agentic systems. It also expanded the error surface, and it is now harder to isolate and mitigate system errors from their core components. This is a problem, since these could range from minor, such as an airline having to refund customers due to hallucinations on their chatbots (Yagoda, 2024) or preparing and releasing unscientific research (Esau et al., 2025); to grave, like doling out dangerous advice in official sites (Landymore, 2026) and deleting production code (Sharwood, 2025); to fatal, leading to deaths (Duffy, 2025; Payne, 2024).[1]

Although a survey could consolidate the field's research, we have a different drive. Almost all scientific work requires a motivation–i.e., an open problem–and, at major conferences, a limitations statement. This means that there exists a considerable amount of 'negative knowledge' in agentic systems literature, such as null and non-generalisable results, brittleness, missing capabilities, to name a few. These are rarely discussed *in relation to one another*, or investigated holistically as boundaries mapping the field's frontiers. Doing so could reveal where the field is concentrated, as well as areas where it could improve, focus, or cross-pollinate.

In this work we review the recent (roughly a full year's) literature on agentic systems. We classify each paper by various characteristics (e.g., use of LLMs-as-judges, presence of statistical significance tests), and their **negative results**–explicitly stated shortcomings and open problems. Then we

[1] See Vectara (2026) for further examples.

carry out statistical analyses to relate these negative results and their components (i.e., type, claimed causes, suggested fixes) to an 'evidence strength score', and to their domains of research.

Our central finding is what we term a *rigour-prevalence mismatch*. The negative-result types most commonly reported across the agentic systems literature–issues with the underlying model's capabilities, with measurement, and with workflow robustness–are precisely those addressed with the weakest evidence (few statistical tests, uncalibrated evaluators, little human verification). Conversely, less-prevalent types such as safety and multimodality are routinely studied with stronger methodology. We argue that this mismatch is not incidental: it marks where the field has the most to gain from methodological investment, and where new research directions are likelier to shift the frontier.

1.1 Contributions

Our goal is to map the field's open problems and their interrelations, not to probe the shortcomings of agentic systems. Hence, our work provides:
1. A systematic evaluation of recent literature on agentic systems, with the state of the field decomposed by our various labels. We find that **the primary research motivation in this field is related to the underlying model's capabilities, measurement, or infrastructure**. Broadly, we find that the field tends to interrelate causes and fixes regardless of domain, with the possible exceptions of AI safety and security, and HCI and design.
2. A statistically-motivated **review of the relationships between negative results, their causes and proposed fixes**, and domains and research strength. Amongst our findings, we note that **certain fixes are related to the paper's strength**: for example, safety and governance and knowledge graphs tend to be more rigorous than prompt engineering.
3. **A rigour-prevalence map** identifying the negative-result types with highest community interest and weakest evidence base. Workflow robustness, underlying capabilities, and runtime/infrastructure are the most actionable targets for the field: widely reported but under-measured, and plausible drivers for broader classes of **agentic system failure**.

We provide recommendations around maintaining rigour to better understanding the cause-fix relationships in research. We hope our map of the field enables further advances through diversified or improved research directions and practices.

2 Background

2.1 Related work

We focus on meta-evaluations which cover negative results, as well as efforts and progress by the field to improve known pathologies. Refer to Wooldridge (2009) for a primer on (early) agentic systems as a foundational work; and Liu et al. (2025) and Mohammadi et al. (2025) for surveys focusing on generative AI-based agentic systems.

We focus on negative results because a common finding in meta-science is that literature overfocuses on positive results; work on the former is thus often thin. This may stem from sociological reasons (e.g., publication bias, institutional incentives; Ioannidis 2005; Nosek et al. 2015; Smaldino and McElreath 2016), which are not necessarily intentional but still lead to non-reproducible work. For example, Ioannidis (2005) notes that most research results fail to generalise to larger sample sizes, or when the experiment has been set up to succeed. Similar results were found by (Nuijten et al., 2016; Smaldino and McElreath, 2016; Raff, 2019; Gundersen and Kjensmo, 2018) in various fields, including AI.

However, in AI there has been a shift towards stricter reproducibility practices, especially in major conferences (see, e.g. AAAI 2026; Carpuat et al. 2026; NeurIPS 2026). This directly impacts our work's viability since, as noted, the limitations-section requirement lets us gather negative results as defined here. Likewise, there has been a recent shift towards providing surveys on explicit open problems in various subfields of AI, such as LLMs (Song et al., 2025) and mechanistic interpretability (Sharkey et al., 2025), which in turn form a major inspiration for our work.

Agentic system literature does contain explicit negative findings. These are common in benchmarks, often showing shortcomings in tool usage (Guo et al., 2025; Schick et al., 2023), state loops and other planning and reasoning tasks (Greshake et al., 2023; Liu et al., 2024; Valmeekam et al., 2022; Yao et al., 2023), debugging code (Jimenez et al., 2024), and generalisation (Deng et al., 2023; Zhou et al., 2024). Due to this, we incorporate specific requirements for papers as part of our quality assessment (Section 5). This includes the need of statistical tests to mitigate stochasticity (Henderson et al., 2018) and appropriate calibration and use of automated (LLM) evaluators in conjunction with humans (De Wynter, 2025). Even though, as noted,

our goal is not to evaluate the state of the field with respect to the quality of research practices, these are (or should be) an integral part of any strong scientific contribution.

2.2 Scope: what *is* an agentic system?

We focus on generative AI-powered agentic systems. We will review components (typically an LLM, but also, say, tool integration), system-level research (e.g., topologies, error resilience), and platform-level considerations (e.g., engineering and hardware considerations). These are often mixed, and thus we do not constrain ourselves to a single domain or application.

3 Methodology

We crawled and annotated a dataset of recent papers focused on agentic systems. Our annotation focused on evaluating their experimental practices and their **negative results**. Here, a 'negative result' is **an explicitly documented shortcoming, failure mode, or open problem** in the articles.

For the labelling and cleaning of the data we used an ensemble of three LLMs-as-judges with human verification. Description and statistics for the judges and further methodological considerations–including longer versions of the label definitions–are in Appendix A, and prompts in Appendix D.

3.1 Corpus collection

We queried Semantic Scholar (Kinney et al., 2023) using the query 'agent workflow llm tool', and limiting our crawl to between 18 May 2025 and 1 January 2026. This range focuses our work on open problems being actively worked on, with recent fixes, and recently peer-reviewed. For every paper we queried the arXiv API[2] to obtain its corresponding preprint. To avoid noise, we enforced an exact match between the Semantic Scholar's record and the one from arXiv, deduplicated our corpus by comparing titles and identifiers, and filtered those not present at a major venue. We then used our judge ensemble to label the preprints as one of 'article', 'survey', 'platform', or 'book'. As per our definitions (Appendix D), we kept scientific articles and surveys, discarding platforms (papers not evaluating agentic systems) and books (a catch-all term for how-to guides, book chapters, and opinion pieces). The final corpus has 274 articles.[3]

[2] https://info.arxiv.org/help/api/index.html

[3] Artefacts are at https://github.com/adewynter/agentic_systems_field_guide.

Label	Description
statistical tests	Presence of stat. sig. tests
open LLMs*	Closed, open, or both, LLMs
human evaluators	Humans in the evaluation process
automatic evaluators	LLMs in the evaluation process
calibrated evaluators	Calibrated LLMs-as-judges
metrics†	A list of the metrics used
primary domain†	The paper's primary domain
negative	It discusses negative results
whose*	From whom come these results
negative type†	A list of negative types
causes†	List (per type) of claimed causes
fixes†	List (per type) of proposed fixes w.r.t. causes

Table 1: Abbreviated table of our core label sets. Precise definitions and supporting sets are in Appendix A. Sets marked with † are free-form; these with * are non-binary sets. For 'whose' and 'open LLMs', the sets are {self, other, both} and $\{0, 1, 2\}$ (r. closed, open, both).

3.2 Annotation

We used three primary label sets: methodology, domain, and negative-result characterisation. The first classifies the papers based on criteria of interest to our research questions, such as the presence of statistical tests or which type of evaluators (human, LLMs) were used. The domain label set focuses on which fields are the primary and secondary focus of the work, as well as whether the papers contain negative results. The negative-result characterisation set–only for papers with negative results–comprises a list of types (what kinds of negative result appear?), plus causes, proposed fixes, and whys (rationales) for every type. Every model was also requested to provide a summary of the results. A condensed table of the label sets is in Table 1, and a full table in Appendix A.

Most methodology labels were binary or ternary, and we aggregated these with a majority vote. The open-ended responses, however, required further processing. For this, we followed Pham et al. (2024) and began began by clustering the responses into fine-grained categories ('minor keys') with the two main LLMs by providing a list of suggested seed categories, along with the summaries and whys from the papers. Then we clustered the minor keys again into coarser categories ('major keys') with the LLMs and another set of suggested categories. We then performed a manual review on the taxonomies, merging and splitting clusters into pragmatically logical units (e.g., separating *human*

factors into *human dependence* and *human-centred design*). Then we used the LLMs to reassign entries into minors and majors, and then reviewed these once more. See Appendix A.2 for the major and minors per taxonomy, and Figure 5 for a taxonomic breakdown of the dataset.

3.3 LLMs and metrics used

The LLMs used for filtering and annotation were GPT-5 (Open AI, 2025b) and Claude Opus 4.6 (Anthropic, 2026). Whenever these disagreed, a third model (GPT-4.1 Open AI 2025a) was used as tiebreaker. We calibrated this strategy with a t-test on a subset of the data, to average 93.1% agreement across binary labels (SD 5.0), with per-label agreement ranging from 86.7% (calibrated evaluators) to 100% (negative-result presence). We further measured their agreement with Gwet's AC1 (Gwet, 2010), Jaccard similarity, or Krippendorff's α, depending on the metric. Our choice of AC1 was because this metric has the advantage of being interpretable at the same scale as the more-commonly used Fleiss' or Cohen's κ, supporting multiple annotators by default, and being less sensitive to skewed distributions (Cicchetti and Feinstein, 1990). In this case, the agreement reported (0.89 ± 0.09) is considered good. We used Jaccard for free-form responses. After normalisation, the average Jaccard distance between models for metrics and primary domain was 0.71 ± 0.10. More details for LLM calls and calibration are in Appendix A. Statistical tests as part of our analysis will be introduced in their respective sections.

4 Corpus composition

4.1 Corpus analysis

Most of the dataset (273 out of 274) contained some form of negative results. In raw statistics, every paper contains an average of 2.6 causes per type. The most frequent negative types relate to the model's *underlying capabilities* (brittleness, reasoning capabilities, etc.) at 16%, *measurement* problems (including benchmarks; 13%), and *runtime and infrastructure* problems (11%; Figure 1). The least frequent are *governance*, *interaction quality*, and *miscellanea*. The most called-out causes relate to other *model limits* (20%), a catch-all for issues such as stochasticity, architecture, or hallucinations. Although most other majors link to model limitations, the remaining causes connect to the environment: *tooling and API* failures (e.g., coverage and runtime fragility) take up the second spot (11%) and *limitations on evaluation* the third (11%). The three lowest are *RAG*, *limited model generalisability*, and *prompt dependence*. The most prevalent negative result types (and research) per domain is *scientific discovery*, followed by *agentic systems* as a whole, and *AI and security*. The least explored are *travel and transport*, *spreadsheets and structured data*, and *games and procedural generation*.

4.2 In-depth analysis

When plotting the co-occurrence of negative types, we observed a considerable skew towards issues within *measurement* and *underlying capabilities*, and less attention to *interaction quality* and *governance*. Although this aligns with the previous observations, the co-occurrence matrix also showed a prevalence of *error types* and *underlying capabilities* co-occurring with most other types. We also measured the co-occurrence of fixes and causes, and observed quantifiable relationships between these (χ^2 = 1553.7; Figure 6). For example, as measured by the matrix' standardised residuals, the dependence was driven (expectedly) by the cause-fix pairs *(scalability cost, cost control)* (R= 14.940) and *(evaluation limitations, evaluation benchmarks/datasets)* (R= 14.906). However, unexpected drives were also brought about by *(dependence on human factors, dynamic workflow adaptation)* (R= 4.9), and *(prompt dependence, prompt engineering)* (R= 6.7).

Running a correspondence analysis (CA) on domains versus negative types showed, on average, close proximity between the negative types and the domains, primarily clustered around the origin ($\chi^2 = 12147.6$, $p < 0.001$, inertia: $[37.8, 17.0]$); with average coordinates of (0, 0) for domain and types. The clearest outliers were *AI safety and security* (both domain and type), and *HCI and design* (domain) and *multimodality* and *interaction quality* (types). The biplot and details on this and the previous section are in Appendix B.

5 A field guide for agentic research

In this section we aim to find the boundaries on the field's open problems, beginning by how a paper's robustness relates to its claims (Section 5.1), and the link between domains and negative types as a composition, not as a co-occurrence (Section 5.2). Then we determine the consensus of the community on the triples (type, cause, fix) (Section 5.3);

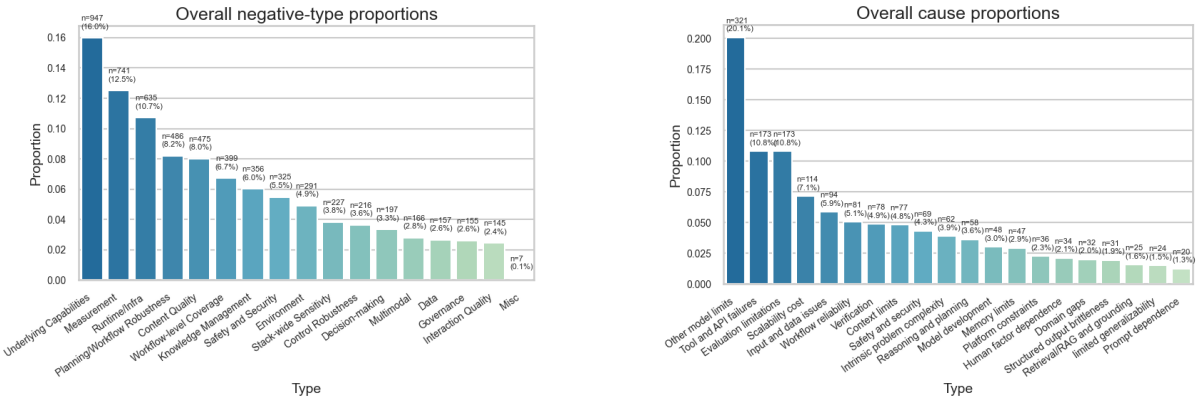


Figure 1: Overall proportions in the corpus of negative types (left) and causes (right), as aggregated by their majors. For negative types, most issues were related to the *underlying model*'s intrinsic capabilities, or *measurement*. For causes, although *other model limits* (a catch-all term for causes not specified in the remaining majors) took up the first spot, the rest related to environmental issues, such as *tool and API failures* and *scalability cost*.

and use that to find areas of opportunity as open problems and evaluation practices (Section 5.4).

5.1 Evidence strength-based analysis

We aim to determine (1) the rigour with which failures are reported, in aggregate and across domains; and (2) whether the proposed fixes are different with stronger evidence. For this we create a deterministic metric based on the methodology label set. Let $M = \{$ human evaluators, automatic evaluators, statistical tests, calibrated evaluators, $\mu\}$, where μ is zero if the metric set is empty, one otherwise. Then the *evidence strength* $\mathcal{E}$ of a paper is given by the sum $\mathcal{E} = \sum_{m \in M} m$. Intuitively, a paper with stronger arguments will have a higher $\mathcal{E}$ (up to 5). We also build a separate, ordinal metric to account for variability on various measurement practices, the *evidence level* $\mathcal{L}$:

$$\mathcal{L} = \begin{cases} 0 & \text{if } \mathcal{E} = 0 \text{ or (human + auto} = 0 \\ & \qquad \text{and } \mu = 0) \\ 1 & \text{if human + auto} > 0 \text{ and } \mu = 0 \\ 2 & \text{if } \mu = 1 \text{ and stat + calibrated} = 0 \\ 3 & \text{otherwise, i.e., } \min(3, \mathcal{E}) \geq 3 \end{cases} \tag{1}$$

The design of $\mathcal{L}$ allows us to denote a paper as speculative ($\mathcal{L}$ = 0), qualitative ($\mathcal{L}$ = 1), quantitative ($\mathcal{L}$ = 2), and controlled ($\mathcal{L}$ = 3). It is needed because $\mathcal{E}$ is a **coarse proxy** for methodological rigour, not a measurement of research quality. Its components capture distinct practices characteristic of the scientific method (statistical testing, type and calibration of evaluators, and metric reporting),

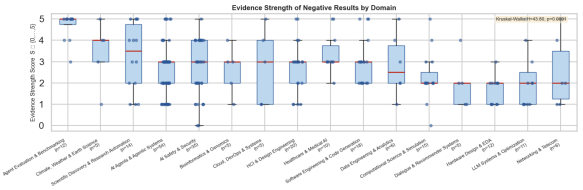


Figure 2: Box plot of evidence strengths per domain. The works with the highest average $\mathcal{E}$ were agentic research (4.7); climate, weather, and earth sciences (3.4); and scientific discovery. The lowest were LLM systems and optimisation (2.1) and telecoms (2.5).

and we weight them equally for interpretability. See Section 9 for further discussion on this. Due $\mathcal{E}$'s nature as a proxy, we use $\mathcal{L}$ as our primary metric. This allows us to circumvent potential issues with $\mathcal{E}$ by collapsing it into four tiers (speculative, qualitative, quantitative, controlled). We use $\mathcal{E}$ only where its continuous range is useful for ranking.

Aggregated per domain, the papers had a mean $\mathcal{E}$ of 2.81, with the highest strength coming from general agentic evaluation and benchmarking ($\mathcal{E}$ = 4.7±0.7) and climate, weather, and earth sciences ($\mathcal{E}$ = 3.4±1.5). The lowest were LLM systems and optimisation ($\mathcal{E}$ = 2.1±1.4) and networking and telecoms ($\mathcal{E}$ = 2.5±1.6). These differences are statistically significant under a Kruskal-Wallis H-test (H=43.8, p<0.0001). See Figure 2 for a plot.

Ehen looking at the proposed fixes with respect to $\mathcal{L}$, we observed a (qualitative) difference: for example, proposed fixes such as knowledge graphs and evaluation benchmarks appeared in high-evidence papers ($\mathcal{L} \geq 3$); while proposed fixes around memory and standardisation often appeared in low-evidence papers (Figure 3).

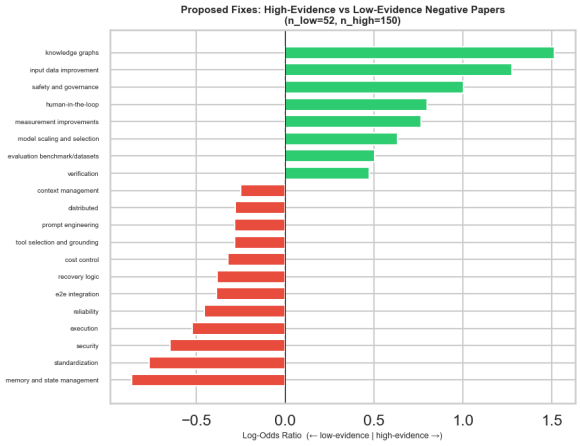


Figure 3: Log-OR plot relating fixes with a paper's $\mathcal{L}$. We consider a high-evidence paper anything $\mathcal{L} \geq 3$, and low-evidence ≤ 1. Proposed fixes such as the use of knowledge graphs and human-in-the-loop were often seen in high-evidence papers. Prompt engineering, standardisation, and memory/state management fixes appeared in low-evidence papers.

5.2 Negative-type overlap

Section 4's results only provide skew, but no information as to which methodological practices associate with which negative types as a mixture.

Given that every paper from our work has a multi-label set of negative types, we focus then on their composition (as a proportion) and fit a Dirichlet regression model with bootstrap over these. We found significant ($\chi^2 = 459.2$, $p < 0.001$) associations between underlying capabilities and the use of human annotations (β= -1.75, CI: [-4.226, -0.339]); measurement and calibrated evaluators (β = 1.49; [0.424, 10.049]); and measurements in software engineering and scientific discovery (β = 1.21, [0.747, 1.738]; 0.86, [0.224, 1.903]). In addition, we used a one-versus-rest logistic model to determine the predictability (instead of frequency) of each negative type. While some types are highly predictable (e.g., safety and security and measurement), others are notoriously hard to detect, such as these relating to stack-wide sensitivity. Both results' figures are in Figure 8, in the Appendix.

5.3 Diagnostic Consensus

Based on the results from the previous section, we next attempted to determine which negative types, or rather, the triple (type, causes, fixes), have consensus across the community.

For this we computed the entropy $H(\Pr[\cdot|t])$ for the distributions of causes ($\Pr[c|t]$) and fixes ($\Pr[f|t]$), where c, f, t are causes, fixes, and types, respectively. We found, in terms of causes, the lowest normalised entropies were in safety and security, measurement, and multimodality ($H = 0.66, 0.68, 0.72$), and the highest in stack-wide sensitivity, interaction quality, and decision making ($H = 0.94, 0.91, 0.91$). With respect to fixes, the lowest entropies were in safety and security, measurement, and knowledge management ($H = 0.70, 0.77, 0.80$); and highest in stack-wide sensitivity, interaction quality, and control robustness ($H = 0.92, 0.92, 0.91$). See Table 6 for results.

We likewise computed the cause-fix normalised mutual information (NMI), with the expectation that specific causes would lead to specific fixes. In terms of (normalised) mutual information, the three lowest categories were underlying capabilities, content quality, and planning and workflow robustness (NMI = $0.14, 0.15, 0.17$). The three highest were miscellaneous causes and fixes, data-related causes, and interaction quality (NMI = $0.69, 0.37, 0.36$).

5.4 'Missingness' of evaluations

Our main finding is a systematic *rigour-prevalence mismatch* across negative types. Combining the prevalence of each type with the methodological rigour of the papers reporting it showed that the types most central to the field's self-reported shortcomings are those its evidence base is least equipped to characterise. In this section we attempt to probe this further by crossing the methodological labels against type prevalence. This is done in a $\mathcal{L}$-free manner, and acts as a secondary check on Section 5.1 that does not rely on $\mathcal{E}$.

We found that, although underlying capabilities of models was present in 87% of the papers, only 49% of these ran statistical tests, and 27% calibrated evaluators. We observed something similar on measurement: called out by 78% of the papers, but only 52% ran statistical tests, and 32% calibrated their judges. On a positive note, 59% of the papers denoting underlying capabilities as an issue (and 62% of evaluations) used human annotations. A table for the full results is in Table 6, and a plot in Figure 4. These results are also observable by checking their Wilson CI at 95% with respect to other label sets (Figure 9).

6 Discussion

6.1 Corpus characterisation

Our first finding was that the corpus was strongly skewed towards negative results. This is perhaps

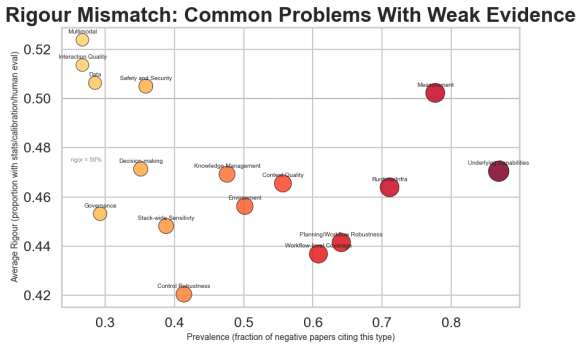


Figure 4: Rigour mismatch between negative types and the methodological label set. As a type goes towards the right, the number of papers considering it as a problem increases. As it goes down, however, its rigour score decreases. In this plot, for example, it can be seen that the community is very interested in workflow-related work (coverage, robustness), but the evidence strength is lacking on these when compared to, for example, interaction quality or measurement.

not surprising: most major venues require a limitations statement, and a scientific work's framing requires an open problem. A pattern for the former was found by De Wynter (2025) for LLM research. Although at a first glance it might not be surprising that a model's underlying capabilities is the most frequent negative type, we note that most agentic systems require more than an LLM to function. In fact, measurement and runtime/infrastructure considerations are the second- and third-most frequent negative types. This is also reflected in the causes. It is also worth mentioning that measurement appeared often as both a negative type and a cause.

The skewness of the data in the biplot towards the centre suggests that both domains and types are tightly correlated–or rather, that most problems arise across all domains. This is supported by domains with specialised problems (e.g., AI safety and security) projecting onto their own axes.

6.2 Deeper results

Our evidence strength-based analysis showed that there is a disparity on measurement practices across various domains. For example, agentic system research; and climate, weather, and earth sciences, had the highest scores. However, other areas were much lower. A deeper analysis on the relationship between fixes and evidence also yielded results such as the fact that prompt engineering and standardisations as proposed fixes typically appeared in low-evidence papers; while works regarding measurement and benchmarking, as well as safety, appeared often in high-evidence papers.

We revisited these results in Section 5.4 by lifting the dependence on $\mathcal{E}$, and found that most papers called out shortcomings on either measurements or underlying model capabilities, yet only about half ran statistical tests and less than a third calibrated their evaluators. This shows that the robustness scores $\mathcal{L}, \mathcal{E}$ are intentionally forgiving, since they only require *some*, but not all, of what it is often considered good scientific practices.

There were several overlaps between negative types and the labels we used in this work. For example, automatic evaluators and human evaluators were almost always used in both underlying capability studies, as well as in knowledge management. The latter also presented robust statistical tests and calibrated evaluators, which, unsurprisingly, is the reason why it ranked atop Figure 3.

If we assume that low entropy implies convergence of the research towards a smaller set of causes and fixes; and high entropy that the problem is highly contested, then we note that several types in agentic systems research have reached some degree of maturity (e.g., safety and security, measurement), and others appear to be underspecified (interaction quality, control robustness, etc). Conversely, high NMI between causes and fixes could be seen as a tighter relation between them: specific causes lead to specific fixes; and low information implies generic fixes regardless of the cause (i.e., the problem remains open). See Appendix C for an example. From our results we found that (modulo miscellanea) data-related causes and interaction quality have the highest NMI. These are relatively tighter cause-fix couplings despite high type entropy, though in absolute terms the NMI remains low. Conversely, lower-valued areas (content quality, planning robustness, underlying capabilities) directly point at open areas of research.

Finally, our rigour mismatch experiment noted that types remain open and with high prevalence across the literature; as well as not meeting a sufficiently-high quality bar. These, we argue, are the core contributions of our work. They directly map their prevalence in the literature with their lack of active, rigorous research. While some types have less prevalence (e.g., multimodality) and have high rigour, it is those which have high prevalence and low rigour which have the most likelihood to impact the entire field. That is because their prevalence means that they are likely *root causes* or components of multiple broader research themes

and issues in the field. Thus, high-$\mathcal{E}$ research in these might impact other areas.

7 Cross-Pollination and Crop Rotation

A field can be over-farmed or left fallow. Its yield and health depends not only on what is planted, but how, and whether it works as an ecosystem. The following recommendations are directions suggested–not prescribed–by correlational observations from our data (see Section 6).

7.1 Soil Health: Rigour and Convergence Co-Evolve

The domains with the lowest entropy over causes and fixes–safety and security, and measurement (Table 6)–are also those with the highest average evidence strength (Figure 2). It is not possible to determine from our data which causes what: whether rigour drives convergence, or convergence enables more rigorous studies, or whether they only reflect field maturity. What we *can* say is that these properties travel together: healthy soil and a settled crop appear jointly. Types exhibiting neither convergence nor rigour (e.g., control robustness ($H = 0.905, 0.910$) or interaction quality ($H = 0.913, 0.915$) are unlikely to converge on their own. Without better evaluation, these can't evolve: their problems aren't necessarily intractable, but without the right measurements it will be difficult to improve upon solutions.

7.2 Crop Rotation: Diversifying Research Targets

Our rigour-prevalence mismatch (Section 5.4) shows the field repeatedly planting the same high-prevalence plots–underlying capabilities, measurement, workflow robustness–in 'depleted soil': widely reported, but under-measured.

This is not a call to abandon these problems, but to rotate. Several high-prevalence types are reported far more often than they are rigorously studied, and several tractable-but-neglected failure modes (e.g., those flagged as opportunities in Table 6). . Rotating effort toward the latter–or toward under-examined failure modes *within* a well-worn type–is more likely to move the frontier than another low-rigour pass over exhausted ground. The expected yield is highest precisely where prevalence and rigour are most misaligned, because those types are probably part of much larger classes of agentic system failure modes.

7.3 Cross-Pollination: Sharing What Works

Specific causes do lead to specific fixes–but reliably only where a domain and type are niche enough to have developed their own understanding (e.g., safety and security). Elsewhere, we were hard pressed to find fixes that logically followed from a stated cause. Consider the entropy and mutual-information results: low type-entropy does not guarantee a tight cause-fix coupling–measurement is among the most converged types ($H = 0.684, 0.770$)–yet has one of the lowest cause-fix NMIs (0.177), meaning the field agrees on *what* is wrong but not on *which fix follows from which cause*. For example, data-related issues have high type-entropy ($H = 0.750, 0.876$) but the highest genuine NMI (0.373), so once a cause is named, the remedy is largely forced. To use our metaphor one last time: these are the plots most worth cross-pollinating. A mature understanding of one area (e.g., how safety work ties fixes to causes) could be transplanted to adjacent, high-entropy areas that lack one. Where many fixes have been proposed yet cause-fix information stays low, the problem is not solved but open–and the motivation behind those fixes might not be well-founded.

8 Conclusion

Agentic systems are in the somewhat uncomfortable border between knowledge- and intelligence-driven objects and their platforms. Hence research in the area is complex and highly non-standardised. In this work we set out to map the boundaries of the field with respect to its open problems, and the internal structure of its research relationships.

We found a systematic *rigour-prevalence mismatch*: the failure modes the field most often reports–underlying capabilities, measurement, and workflow robustness–are precisely those it studies least rigorously. These are the ones where new investment is likely to move the frontier. We also found that rigour and convergence co-evolve: the types with the most settled diagnoses are also those studied with the strongest evidence. These patterns emerged only from viewing the works holistically, and by linking negative types to their rigour. These support a new-old view. Despite the field's focus on improving underlying model capabilities, agentic systems are best seen from a new angle: as objects of study in their own right, rather than software supporting a generative model or a set of prompts.

9 Limitations

Our work is limited in two ways: our reliance on automated metrics (namely, LLMs-as-judges) for extraction and labelling of the work, and our (intentional) selection bias. For the former, we leveraged on LLMs-as-judges to extract and annotate data. To mitigate any potential biases or inaccurate results (e.g., hallucinations during clustering), a manual verification step was performed on all aspects of the work which involved them. We likewise computed measures of statistical significance to evaluate their reliability. The results of this calibration are in Appendix A.

Likewise, our corpus was constructed via a single Semantic Scholar query ('agent workflow llm tool') over a narrow window (18 May 2025 - 1 January 2026). We performed a sanity check to confirm that this window did not introduce an age-citation confound within our sample. Fitting an ordinary least squares (OLS) model of citation counts versus publication date yielded a statistically-significant negative slope ($\beta_1 = -0.004$, SE $= 3.04 \times 10^{-4}$, $R^2 = 0.411$, $p < 0.001$). Rescaled, this corresponds to roughly -1.52 citations per year or -0.42 per 100 days of age difference, which we consider negligible for our analyses; we therefore do not control for age. It is unlikely that controlling for either source of bias would preserve all our findings. The relative under-representation of safety-adjacent work in our corpus might strengthen our rigour-prevalence mismatch result (safety has both low prevalence and high rigour in our sample, and sampling more safety work would reinforce this), but might weaken other conclusions about which domains are converging.

Finally, our evidence-strength score $\mathcal{E}$ quantifies methodological rigour through five binary practices (human evaluation, automatic evaluation, statistical testing, calibration, and metric reporting). However, rigour and evidence strength are not identical. A paper may be rigorous in ways our labels do not capture (e.g., careful ablations, qualitative error analysis, strong theoretical grounding, or well-designed human studies that go beyond mere presence). Conversely, a paper may check all five boxes while remaining methodologically weak. We explicitly treat $\mathcal{E}$, and especially the ordinal $\mathcal{L}$, as a *forgiving* proxy (Section 5.1; hence our statements in the introduction around not seeking flaws in the field). This metric is intended to surface field-level patterns, not to assess individual papers. Our central claims rely on its tiers rather than its exact value.

10 Ethics

To avoid overloading all services, we requested API keys when applicable and implemented timeouts and backoff. Due to licencing (mainly around redistribution of works) and anonymity considerations, the source text will not be released, although we will release reproducible code and annotations to encourage repetitions of this study. LLMs were used to support code-writing (debugging, plotting) and to create LaTeX tables from raw results. All outputs were manually verified for accuracy and corrected for style.

References

AAAI. 2026. Reproducibility checklist.

Anthropic. 2026. Introducing Claude Sonnet 4.6.

Marine Carpuat, Marie-Catherine de Marneffe, Ivan Vladimir Meza Ruiz, Jesse Dodge, Margot Mieskes, and Anna Rogers. 2026. Responsible NLP research. Accessed 2026.

Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High agreement but low kappa: II. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Clare Duffy. 2025. Parents of 16-year-old sue OpenAI, claiming ChatGPT advised on his suicide.

Paul Esau, Nazar Shmatko, and Alex Adam. 2025. GPTZero finds over 50 new hallucinations in ICLR 2026 submissions.

Oren Etzioni. 1996. Moving up the information food chain: deploying softbots on the world wide web. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1322–1326. AAAI Press.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA. Association for Computing Machinery.

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. 2025. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. *Preprint*, arXiv:2403.07714.

Kilem Li Gwet. 2010. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61:29–48.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, and 29 others. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Frank Landymore. 2026. US Government deploys Elon Musk's Grok as nutrition bot, where it immediately gives advice for rectal use of vegetables.

Fan Liu, Jindong Han, Tengfei Lyu, Weijia Zhang, Zhe-Rui Yang, Lu Dai, Cancheng Liu, and Hao Liu. 2025. Foundation models for scientific discovery: From paradigm enhancement to paradigm transition. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*.

Mahmoud Mohammadi, Yipeng Li, Jean-Pierre Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.

NeurIPS. 2026. Neurips paper checklist guidelines. Accessed 2026.

Brian A. Nosek, George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, Christopher D. Chambers, Gilbert Chin, Garret Christensen, Mike Contestabile, Allan Dafoe, Eric Eich, Jeremy Freese, Rachel Glennerster, Daniel Goroff, Donald P. Green, Bradford Hesse, Macartan Humphreys, and 20 others. 2015. Promoting an open research culture. *Science*, 348(6242):1422–1425.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

Open AI. 2025a. Introducing GPT-4.1 in the API.

Open AI. 2025b. Introducing GPT-5.

Kate Payne. 2024. An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. *AP News*.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Stuart Russell and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*. Survey Certification.

Simon Sharwood. 2025. Vibe coding service replit deleted user's production database, faked data, told fibs galore.

Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Paul E. Smaldino and Richard McElreath. 2016. The natural selection of bad science. *Royal Society Open Science*, 3(9):160384.

Peiyang Song, Pengrui Han, and Noah Goodman. 2025. A survey on large language model reasoning failures. In *2nd AI for Math Workshop @ ICML 2025*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Vectara. 2026. Awesome agent failures.

Michael Wooldridge. 2009. *An Introduction to MultiAgent Systems*. John Wiley & Sons Ltd.

Adrian de Wynter. 2025. Awes, laws, and flaws from today's LLM research. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12834–12854, Vienna, Austria. Association for Computational Linguistics.

Maria Yagoda. 2024. Airline held liable for its chatbot giving passenger bad advice - what this means for travellers.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

A Detailed methods

A.1 Annotation rubrics

In Table 2 we show a more expanded version of our annotation rubric (i.e., labels and definitions) along with their categorisation with respect to this work, and the values which they can take. Given readability considerations, we refer the reader to the prompts in Appendix D for the concrete, slightly-less-formal definitions.

A.2 Taxonomies

Samples of the taxonomies (majors only) are in Table 3. These are, in order, from the negative type (17 majors and 151 minors total), causes (r. 20 and 65), and fixes (r. 33 and 92) taxonomies. Similar clusterings were done for papers and metrics. Due to the large number of elements, we refer the reader to the repository for a full depiction.

A.3 Annotation details

We used GPT-5 (5.2; version 2025-11-24), Claude Opus 4.6, and GPT-4.1 (longco; 2025-04-14). The latter was called with temperature set to zero, and–due to its age–as a tiebreaker whenever the other two models disagreed. It was not used for free-form generation. We requested 5,000 maximum tokens for completion. All models were called through the Azure OpenAI API. Data processing was carried out on a consumer-grade laptop.

We performed a calibration for the LLMs by measuring their agreement values. On average, model agreement (as measured by AC1) was considered good to perfect (0.8895 ± 0.0942; Gwet 2010). Per-label breakdowns are in Table 4. Average Jaccard distance under this strategy is 0.7112 ± 0.0954. The t-test on binary labels–other label sets are evaluated in the table–showed an expected performance of 93.1 ± 5.0 on average at a 95% CI versus the aggregate, with CI results of 88.3 ± 8.2, 82.9 ± 4.5, 91.5 ± 3.4, 85.3 ± 4.3, 99.3 ± 1.1, and 100 ± 12 for *statistical_tests*, *human_evaluators*, *automatic_evaluators*, *calibrated_evaluators*, *is_theoretical*, and *negative*, respectively.

B Further corpus characterisation

A figure of the relationships between majors and minors in the corpus may be found in Figure 5. This acts as a bird's eye view for our entire corpus. The relationships between negative types and between causes and proposed fixes are in Figure 6. Likewise, The co-occurrence matrix for the negative types and the full biplot are in Figure 7. The residual computations for the cause-fix pairs are in Table 5.

C Further results

For our study on rigour mismatch, we report the full results in Table 6, and an alternative take is in Figure 9. The Dirichlet and one-versus-rest predictability analyses are in Figure 8. From these we can observe a strong relationship between variables,

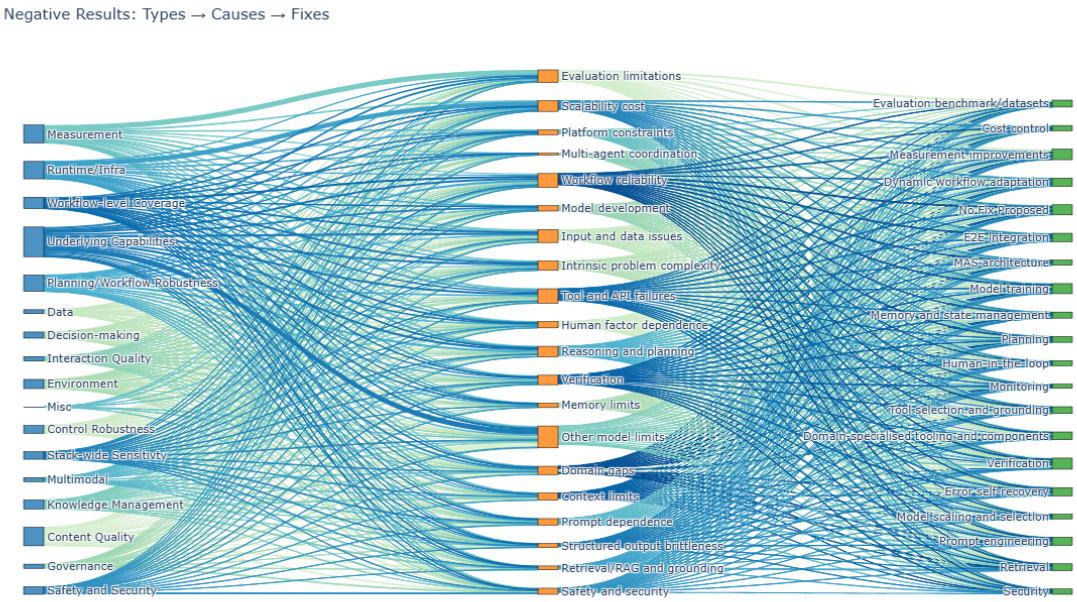


Figure 5: Sankey diagram depicting our main corpus, as defined by the relationships between negative types (left), causes (centre), and fixes (right), aggregated by majors only.

such as how knowledge management is (positively) related to $\mathcal{E}$. The one-versus-rest (logistic) plot illustrates that predictability of each negative type is high in more regular fields, such as safety, but less easy to detect in others (e.g., prompt engineering). Similar observations can be seen from the fully-decomposed results for our entropy analysis from Section 5.3 (Table 6).

C.1 A worked example of entropy and NMI

To make these quantities concrete, consider two contrasting types. *Safety and security* has the lowest entropy over both causes and fixes ($H = 0.657, 0.701$): the literature largely agrees on a small set of failure sources and converges on a correspondingly narrow set of remedies. *Stack-wide sensitivity*, by contrast, sits near the top of both distributions ($H = 0.939, 0.918$): reported causes are diffuse and proposed fixes more so, consistent with a failure mode the field has not yet learnt to localise. Entropy and mutual information capture different things, however, and reading them together is informative. A low-entropy type need not couple its causes to its fixes: *measurement* is highly converged ($H = 0.684, 0.770$) yet has among the lowest cause-fix NMI (0.177)—the field knows *that* measurement is a problem but not *which* fix a given cause implies. The inverse also occurs: *data*-related issues remain contested in their framing (high entropy, $H = 0.750, 0.876$) yet exhibit the highest non-trivial NMI (0.373), indicating that once a specific data cause is identified, the appropriate fix is largely determined. We exclude *miscellanea* from these comparisons, as its high values ($H = 0.958, 0.968$; NMI $= 0.685$) are an artefact of its catch-all composition rather than a substantive finding.

D Prompts

In this section we capture the prompts used in this work. The prompts for classifying papers (article, survey, etc) are in Prompt 1. The prompts for labelling the papers (characterisation) are in Prompts 2 and 3. We split these in two to simplify the problem, and to keep reasoning models from running out of tokens.

Finally, the prompt for characterising agentic workflow articles is in Prompt 4, and the negative type extraction prompt in Prompt 5.

For simplicity, we omit the specifications on output format (JSON).

Class	Label	Description	Values
Filter	type of work	What main contribution this work makes	{article, survey, book, platform}
Methodological	statistical tests	Presence of statistical significance tests	$\{0, 1\}$
	open LLMs	Use of closed-only, open-only, or both (r.) models	$\{0, 1, 2\}$
	human evaluators	Use of humans in the evaluation process	$\{0, 1\}$
	automatic evaluators	Use of LLMs in the evaluation process	$\{0, 1\}$
	calibrated evaluators	If there were LLMs-as-judges, whether these were calibrated	$\{0, 1\}$
	is theoretical	Whether this paper's primary contribution is theoretical	$\{0, 1\}$
	has theoretical background	If the paper is primarily applied, whether it contains a mathematically-grounded, novel, motivation	$\{0, 1\}$
	metrics	A list of the metrics used in the paper	Free-form
Domain	primary domain	The paper's primary domain	Free-form
	secondary domain	The paper's secondary domain	Free-form
	negative	Whether it discusses negative results (as per our definition)	$\{0, 1\}$
	whose	From whom are these negative results claimed to originate	{self, other, both, ϵ}
Negative	negative type	A list of negative types present in the work	Free-form
	causes	A list (per type) of the claimed causes of each negative type	Free-form
	fixes	A list (per type) of the proposed fixes as related to the causes	Free-form
	whys	A list (per type) of the rationales for each fix	Free-form
	summary	A brief summary of the negative results from the work	Free-form

Table 2: Complete list of annotation fields and their labels used in the work. Precise definitions of each (e.g., what does 'calibration' or 'book' mean) are in the prompts (Appendix D). Here ϵ denotes the empty string.

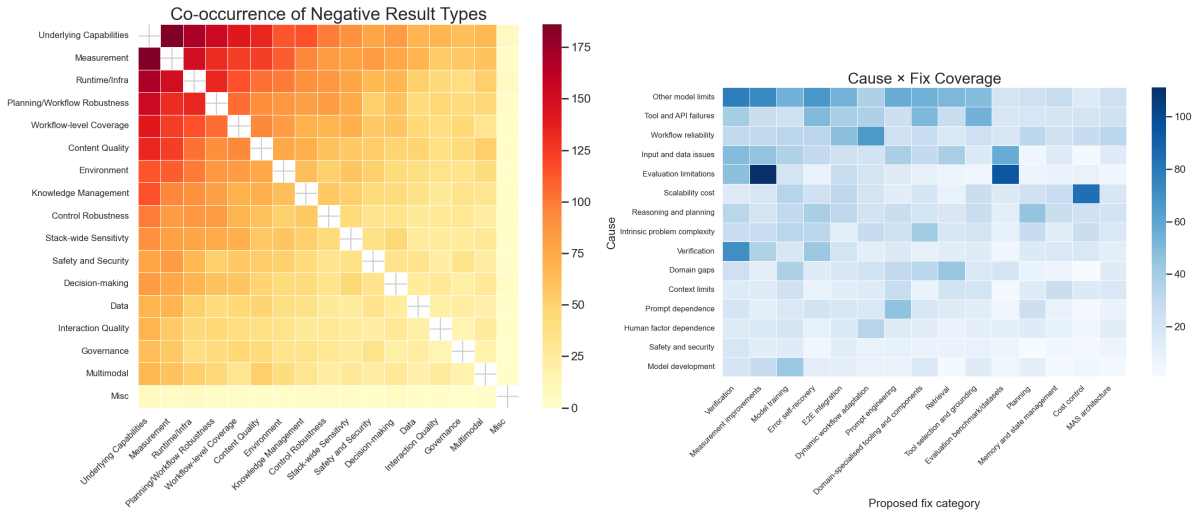


Figure 6: Co-occurrence matrices (as heatmaps) for negative types (left) and between causes and proposed fixes (right). (*Left*): it can be seen from the plot that there is an emphasis from the community to investigate underlying capabilities as they relate to measurement, infrastructure, and other problems. This suggests some level of overlap within negative-types, which we explore later on in Section 5.2. (*Right*): unlike the co-occurrence matrix for types, this matrix is less symmetric and suggests that there is a prevalent theme of various approaches (*prompt engineering, verification, training*, etc.) being proposed as fixes for various model limitations.

Types	Causes	Fixes
Content-level output quality: faithfulness, clarity, verbosity, compilability, generation hallucinations	Other model limits (stochasticity, architecture, effectiveness, incl. multimodal)	Retrieval (augmentation, grounding, evidence, KBs)
Interaction-level output quality: UX, interactivity, personalization, user-facing feedback	Domain gaps (knowledge, expertise)	Knowledge graphs (knowledge representations, ontologies)
Environment realism: tool and sandbox limitations	Prompt dependence (instruction ambiguity, sensitivity, underspecification, misalignment)	E2E integration (pipelines, testing and deployment, hardware, structured workflows)
Measurement: evaluation, benchmarks, metrics, tests, and validation	Reasoning and planning (trajectory complexity, long-horizon fragility, underspecification)	Planning (task decomposition, staging)
Safety and security: attacks, jailbreaks, malicious tool use, verification	Context limits (context window, prompt bloat, information loss during compression, truncation)	MAS architecture (decomposition, specialization, orchestration)
Workflow-level support, scope, and coverage	Memory limits (state tracking, persistence, failures)	Verification (validation, reflection, diagnostics, consistency checks)
Underlying model capabilities: brittleness, collapse, distribution shift, and generalization limits	Tool and API failures (coverage, capability, execution, environment runtime fragility, mapping ambiguity)	Error self-recovery (iterative refinement, self-correction, learning from errors and recovery)
Knowledge management: RAG, memory, context, state persistency	Structured output brittleness (formatting, schema)	Prompt engineering (CoT, templates, structured prompts)
Control robustness: idempotency, recovery, invariants, transactional workflows	Verification (error propagation, validation, proofs and weak self-correction)	Structured outputs (schemas, formatting control, typed interfaces)
Stack-wide configuration sensitivity: prompting, system parameters, tool schemas, templates	Evaluation limitations (benchmark and metric coverage, realism, ground truth)	Context management (bounding, reducing, selecting)
Planning, decomposition, coordination, and workflow robustness	Input and data issues (ambiguity, scarcity, noise, quality, documentation gaps)	Memory and state management (caching, persistence, external memory)
Decision: routing, ranking, and gating	Workflow reliability (rigidity, orchestration brittleness, heterogeneity)	Tool selection and grounding (execution, routing, planning, pruning)
Governance: auditability, compliance, human approval, social impact, and interpretability	Scalability cost (compute, latency, energy, money)	Domain-specialised tooling and components
Runtime, scalability, deployment, and infrastructure constraints	Retrieval/RAG and grounding (KB, corpora) coverage limits	Monitoring (observability, logging, provenance, tracing)
Data quality, bias, and coverage gaps	Human factor dependence (operational and oversight burden, SME dependence)	Evaluation benchmark/datasets (creation, expansion)
Multimodal, voice, and nonverbal limitations	Safety and security (governance, transparency, privacy)	Measurement improvements (metrics, rubrics, judging redesign, ablations, calibration)
Miscellaneous	Multi-agent coordination (communication overhead, complexity, conflicts)	Human-in-the-loop (oversight, expert review)
	Model development (training signal, reward design, calibration)	Determinism / reproducibility controls
	Platform constraints (rate limits, drift, integration, access, hardware)	Security (sandboxing, thread modelling, access control)
	Intrinsic problem complexity (space size, convergence, physical constraints)	Dynamic workflow adaptation (evolution, generation, rewiring, transferrability, reconfiguration)
		Model training (RL, post-training, fine-tuning, reward shaping, distillation)
		Other model improv. (architecture, open., decoding)
		Cost control (scheduling, budgeting, loop control)
		Input data improvement (quality, curation)

Table 3: Majors from our taxonomy. With the exception of Types, these are a sample subset. These are *not* aligned row-wise. Each column must be read independently. Moreover, the majors are meant to be machine-readable, so (for example) having 'memory limits' as a cause of error types in 'governance' does not imply that governance *as a whole* has memory limits as a cause, but instead that there exists at least one paper dealing with a problem where at least one of its causes maps to a minor falling within the 'memory limits' major in causes.

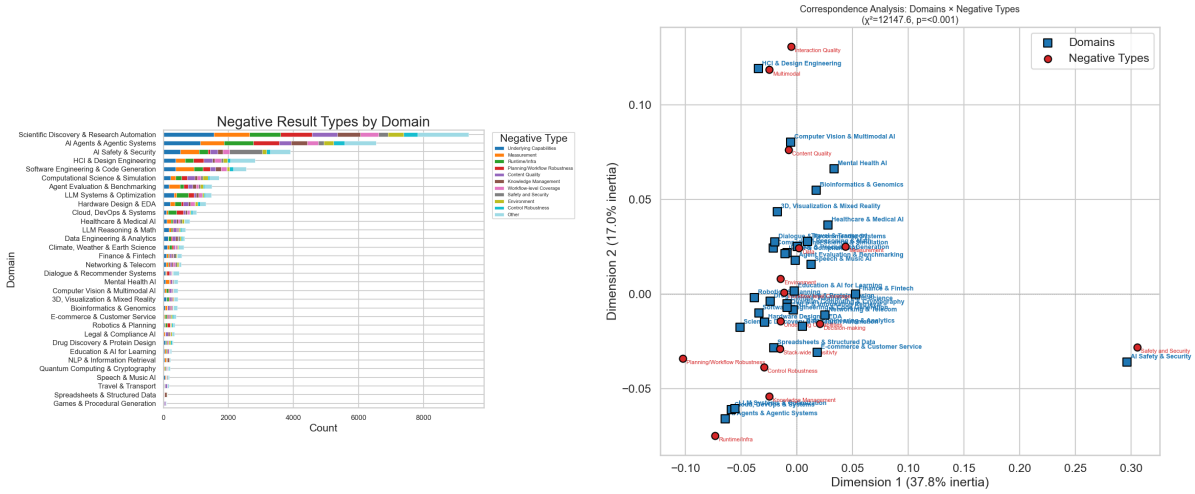


Figure 7: Results per domain (left) and a biplot for negative types and fixes (right). *(Left)*: in terms of results per domain, the domain with the largest amount of negative types (and research) is *scientific discovery*, followed by *agentic systems*. *(Right)*: observe how most entries are concentrated near zero, with only a few outliers towards the right (AI safety and security) and top (HCI and design, multimodality, and interaction quality).

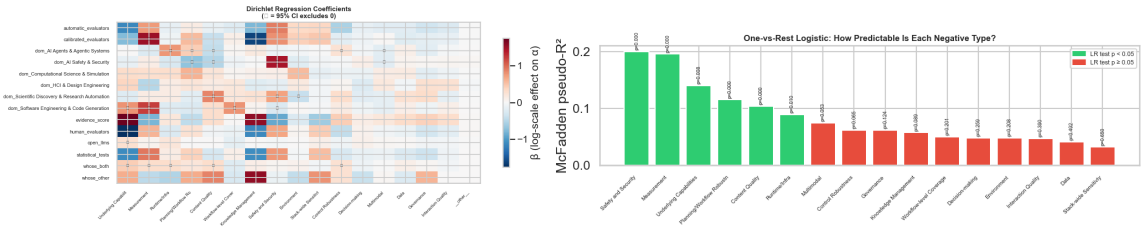


Figure 8: *(Left)* Dirichlet regression over negative-type overlaps. There exist significant associations between some variables, such as in knowledge management being strongly related to $\mathcal{E}$ and *whose* when taking the value of 'other'. These also have a positive tendency to use calibrated evaluators. *(Right)* One-versus-rest logistic model over negative types. Some types, such as these related to safety and security are highly regular. On the other hand, issues related to stack-wide sensitivity and appear to have less predictability.

Label	Pairwise AC1	Other
statistical_tests	0.7438	$\kappa = 0.7350$
human_evaluators	0.7984	$\kappa = 0.7858$
automatic_evaluators	0.9124	$\kappa = 0.9124$
calibrated_evaluators	0.8882	$\kappa = 0.7869$
is_theoretical	0.9981	$\kappa = 0.9556$
negative	0.9963	$\kappa = 0.6648$
whose	—	$\alpha = 0.5446$
metrics	—	$J = 0.6703$
primary_domain	—	$J = 0.8431$
secondary_domain	—	$J = 0.6204$

Table 4: Inter-annotator agreement across the LLM labellers ($N = 274$). We report average pairwise Fleiss' κ and Gwet's AC1 for binary labels; for set-valued labels average pairwise Jaccard similarity (J); and for ordinal scales Kripendorff's α. On average, agreement across models with the consensus was considered good-to-perfect. This can be explained in part by the skewness of our dataset (is_theoretical is omitted from the table due to its extreme imbalance) and the use of a tie-breaker. While Gwet's AC1 is robust to it, it has its limits (e.g., the number of negative papers was 273/274). This is also why we employ dual agreement metrics in this section.

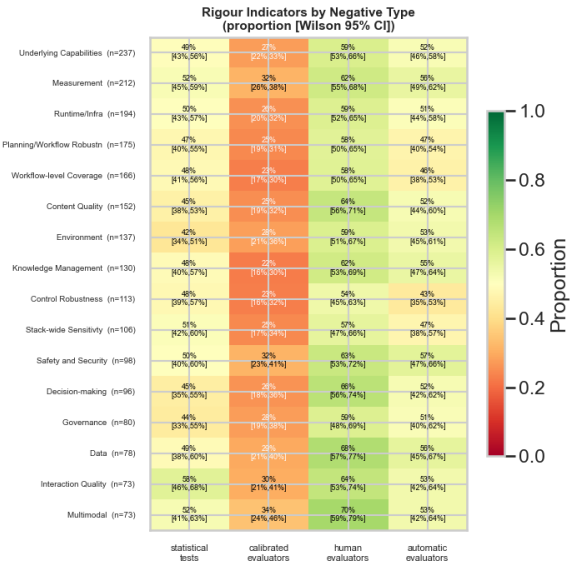


Figure 9: Rigour indicators like the ones from Section 5.4, but leaning on Wilson's CI instead. The results are the same as in the other plots.

The user will provide some text. Your job is to output what type of text it is.

You may pick from the following:
1. article: a scientific article, but not a book chapter, survey, or opinion piece (e.g., these with 'Opinion' in the title). Here we also count technical reports and benchmarks as long as they explicitly evaluate agentic workflows/LLMs using tools.
2. survey: a review of the field.
3. book: a textbook, book chapter, how-to guide, or opinion piece. Broadly, any text without a clear contribution and/or scientific rigour.
4. platform: a scientific article where an agentic system is NOT being evaluated. This could be, for example, a text describing a codebase supporting workflows.

We'd like to see papers where an agentic system is evaluated in some capacity but with scientific rigour. These would be articles.

Prompt 1: System prompt for classifying the papers crawled in either articles, surveys, books, or other technical reports.

Cause	Fix	R
Scalability cost	Cost control	14.940
Evaluation limitations	Evaluation benchmark/datasets	14.906
Evaluation limitations	Measurement improvements	12.276
Verification	Verification	7.220
Prompt dependence	Prompt engineering	6.665
Model development	Model training	6.617
Input and data issues	Evaluation benchmark/datasets	6.570
Domain gaps	Retrieval	6.492
Workflow reliability	Dynamic workflow adaptation	5.804
Reasoning and planning	Planning	5.354
Human factor dependence	Dynamic workflow adaptation	4.906
Tool and API failures	Tool selection and grounding	4.593
Context limits	Memory and state management	4.528
Verification	Error self-recovery	3.812
Prompt dependence	Planning	3.318
Intrinsic problem complexity	Domain-specialised tooling and components	3.263
Tool and API failures	Domain-specialised tooling and components	3.176
Input and data issues	Retrieval	2.898
Domain gaps	Domain-specialised tooling and components	2.652
Workflow reliability	MAS architecture	2.592

Table 5: Residuals for the cause-fix contingency matrix from Figure 6. While some cause, fix pairs are expected, it is worth noting outliers such as (prompt dependence, planning).

Type	$H(\Pr[c\|t])$	$H(\Pr[f\|t])$	NMI	Prevalence (%)	Avg Rigour (%)	Mismatch
Stack-wide Sensitivity	0.939	0.918	0.271	38.8	44.8	0.214
Interaction Quality	0.913	0.915	0.358	26.7	51.4	0.130
Decision-making	0.913	0.885	0.282	35.2	47.1	0.185
Control Robustness	0.905	0.910	0.262	41.4	42.0	0.239
Underlying Capabilities	0.901	0.899	0.143	86.8	47.0	0.459
Planning/Workflow Robustness	0.876	0.900	0.174	64.1	44.1	0.358
Workflow-level Coverage	0.851	0.905	0.270	60.8	43.7	0.342
Environment	0.835	0.878	0.249	50.2	45.6	0.272
Content Quality	0.834	0.875	0.154	55.7	46.5	0.297
Governance	0.785	0.872	0.350	29.3	45.3	0.160
Knowledge Management	0.762	0.799	0.254	47.6	46.9	0.252
Runtime/Infra	0.728	0.886	0.203	71.1	46.4	0.381
Data	0.750	0.876	0.373	28.6	50.6	0.141
Multimodal	0.719	0.859	0.322	26.7	52.4	0.127
Measurement	0.684	0.770	0.177	77.7	50.2	0.386
Misc	0.958	0.968	0.685	—	—	—-
Safety and Security	0.657	0.701	0.263	35.9	50.5	0.177

Table 6: (*Left*): normalised entropies for causes and fixes with respect to types, and normalised mutual information between causes and fixes. (*Right*): rigour mismatch with respect to prevalence for every negative type, with *Misc* omitted.

I need you to help me sift through some papers. You'll be getting text from a scientific article, and I need you to label it in a certain way.
Your task will be to label it as follows:

1. statistical_tests: whether the paper contains any type of statistical analysis beyond mere numbers (e.g., Cohen's kappa, standard deviation, etc). 1 if it does, 0 otherwise.
2. open_llms: whether the paper use any open LLMs (Qwen, Llama, etc), or closed and API-based (GPT-5, Claude, etc), or both. 2 if open-source only, 1 if both, 0 if closed models only.
3. human_evaluators: whether the authors used human annotation in their evaluation process (1 if yes, 0 if not)
4. automatic_evaluators: whether the authors used automatic measurements (LLMs-as-judges) in their evaluation process (1 if yes, 0 if not)
5. calibrated_evaluators: if the authors used LLMs-as-judges, whether these were calibrated: this means either an inspection of the outputs, or a more formal work with statistical measures (e.g., a t-Test). 1 if yes, 0 if not.
6. metrics: which metrics did the authors use (accuracy, F1, etc)?

Prompt 2: First article classification prompt. It categorises the papers based on research features typical to LLM-centred research, such as statistical tests and type of evaluators used. Remark the skew towards LLMs-as-judges measurement–this is because metrics already captures other evaluations.

I need you to help me sift through some agentic workflow papers. You'll be getting text from a scientific article, and I need you to give me two labels.

These are:
1. is_theoretical: a paper with a primarily theoretical contribution explaining or modelling an aspect of agentic workflows (and maybe minor experiments)
2. has_theoretical_background: a paper with a primarily applied contribution (e.g., a results section) but including a theoretical framework for its work. The framework must be novel (e.g., 'proposed framework').

Typically papers which are primarily applied will not be theoretical, and there is a chance that the paper will be neither.
All labels are binary (they either have it or they don't).

Prompt 3: Second article classification prompt, focused solely on theoretical work. The split between classification prompts was based on number of fields, and then common themes. Our rationale was that it would simplify the problem for the annotators and ensure higher performance.

I need you to help me sift through some agentic workflow papers. You'll be getting text from a scientific article, and I need you to label it with two main areas.
The first is its domain (which field it belongs to?). The second relates to negative results: I want to see if there are any limitations/shortcomings discussed in that paper, and who they relate to.

1. primary_domain: what predominant domain/field belongs to (1-3 words: for example, 'cybersecurity', 'hci', 'AI+science', 'deep research', etc.)
2. secondary_domain: any other domain it covers (same as before, 1-3 words), or empty ("")
3. negative: whether it discusses limitations, errors, and/or any other negative results or shortcomings of agentic workflows. Explicitly marked future work also counts, so long as it is stated clearly that it remains an open problem. This must be 0 (it doesn't have any) or 1 (it discusses them)
. 4. whose: who do these negative results belong to. Use 'self' it the authors discuss negative results from their own work; 'other' if it relates to other works; and 'both' if they relate to both. This can be empty ("") if no negative results are discussed.

Prompt 4: System prompt for agentic workflow article classification. This prompt seeks to determine the domain and type of results provided (i.e., whether the authors provided open problems as either a solution or as future work).

I need you to help me sift through some papers. You'll be getting text from a scientific article related to agentic workflows, and I need you to label it in a certain way.
This paper contains a discussion of negative results (by that we mean: limitations, errors, shortcomings, explicitly-noted open problems) by {whom}.
Your task will be to label it and return a JSON as follows:

1. negative_types: a list of dicts. Each dict is of the form: {type: {causes: ..., fixes: ..., whys: ...} }, where:
- type is a str of 1-3 words each (e.g., 'timeouts', 'hallucinations', etc.) classifying each type of negative result discussed.
- causes must be a list of 1-3 words each of why–according to this work–that negative type occurs (e.g., 'llm unreliability', 'hardware constraints', etc.). It may be empty if no causes are discussed.
- fixes: a list of 1-3 words each of the suggested fixes, if any, by this work to the negative results (e.g., 'finetuning', 'routers'). This list can also be empty if no fixes are proposed for this type. These fixes also can range from their proposed methods, to future work.
- whys: a list of 1-3 words each of the reasons why this work suggests these fixes (e.g., 'reduce call number'). This list can also be empty if no fixes or whys are proposed for this specific fix.
5. summary: a SHORT (two or three sentences max) summary of the negative results discussed.

Prompt 5: System prompt for the negative result classification task. This prompt is comparatively free-form and requests both the types of negative results provided, their reasons, and any fixes provided.