BackTrend: Evaluating Scientific Weak-Signal Prediction via Backward Reconstruction

Anonymous ACL submission

Abstract

Scientific weak signals are early, low-visibility research directions that later become central to mature scientific topics, yet existing resources such as trend tracking, citation forecasting, and foresight reports rarely provide validated reference sets that link concrete early precursors to later paradigms. We introduce BackTrend, a retrospective benchmark in which, given a mature target topic and a temporal evidence constraint, systems must recover two types of precursors: *problem-space signals* (underrecognized research problems) and *solution-space signals* (emerging methods for known problems). BackTrend contains 253 mature target topics and 2,027 human-validated weak signals across 13 domains grouped into AI, Science, Healthcare, and Engineering. We evaluate frontier LLMs, RAG systems, and agentic research systems via set-level and signal-level semantic matching together with human annotation. Current systems generate plausible but misaligned precursors, with topic drift, granularity mismatch, and incomplete coverage; GPT-5.4 leads overall mainly through cross-domain robustness, additional retrieval or web search does not consistently help, and LLM-based judging only partially agrees with human assessment of weak-signal matching.

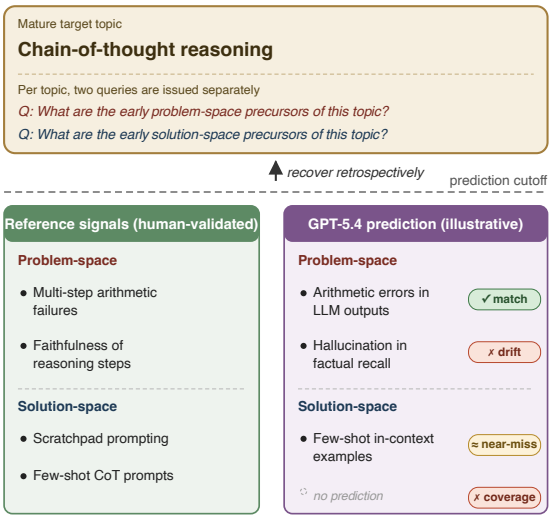

Figure 1: BackTrend task overview, illustrated with chain-of-thought reasoning. Given a mature target topic and the prediction cutoff (§3.1), systems retrospectively recover the early weak signals that later contributed to it, decomposed into *problem-space* (underrecognized research problems) and *solution-space* (emerging methods for known problems). The right column shows an illustrative GPT-5.4 prediction with the four match outcomes: exact match, topic drift, lexical near-miss, and coverage failure. Years are omitted; the benchmark uses an emergence window of 2023–2024.

1 Introduction

Identifying early indicators of transformative research, or weak signals, has long been a difficult goal across scientific disciplines (Ansoff, 1975; Porter et al., 2004). Detecting such signals is valuable for researchers, corporations, and policymakers because it informs long-term investment, research prioritization, and policy design (Kajikawa et al., 2008; Yoon and Kim, 2012; Lee et al., 2014; Ogawa and Kajikawa, 2015). As AI and machine learning accelerate the pace of scientific research, effective foresight depends on detecting low-visibility but potentially high-impact ideas within a large volume of background noise, and recent work continues to identify open issues in methods for detecting such emerging themes (Ren and Zhao, 2021; Liu et al., 2023). Across scientific domains, LLM-based research agents are increasingly deployed to search literature and synthesize emerging areas, yet are rarely tested on whether they can recover low-visibility precursors that later become central, leaving an open challenge for long-context reasoning and abstraction control.

We address this challenge with **BackTrend**[1], a

[1]Our code and data are available at https://anonymous.4open.science/r/BackTrend, and will be released publicly upon publication.

retrospective benchmark for scientific weak-signal prediction (Figure 1). Given a *mature target topic* and a temporally restricted evidence window, systems must recover early *problem-space signals* (underrecognized research problems) and *solution-space signals* (emerging methods for known problems) that later became important precursors of that topic. Because prospective evaluation would require waiting years for outcomes to mature, BackTrend uses backward reconstruction as a controlled proxy for scientific foresight. We construct the reference set with a deep research agent grounded in Semantic Scholar, then have PhD-level domain experts among the authors validate every candidate for topical relevance, temporal consistency, and reference support; at evaluation time, systems must recover the held-out signals using only historically available evidence from publications published before the prediction cutoff year.

Our results show that strong systems can produce plausible precursors but routinely miss the benchmark's intended structure. GPT-5.4 leads overall (46.9 human F1) on the strength of cross-domain robustness rather than dominance in any single slice; LLM judges consistently overestimate performance relative to human annotators, and the dominant errors are topic drift, granularity mismatch, lexical near-miss, and incomplete coverage. Surprisingly, neither deeper retrieval in RAG nor larger search budgets in DeepResearch reliably help, and BERTScore alone barely separates systems (84.2–86.9) despite a human F1 range from 14.2 to 46.9. Weak-signal prediction therefore tests semantic alignment, abstraction-level control, and coverage across heterogeneous scientific domains rather than fluency in generating scientific phrases.

Our main contributions are summarized below:

- We formulate *retrospective scientific weak-signal prediction*: recovering early problem-space and solution-space precursors of mature research topics under a temporal evidence constraint that simulates research foresight.
- We construct BackTrend, a benchmark of 253 mature target topics and 2,027 human-validated weak signals across 13 domains in AI, Science, Healthcare, and Engineering.
- We benchmark frontier LLMs, RAG systems, and DeepResearch agents on BackTrend, finding that current systems produce plausible but misaligned signals and that automatic evaluation only partially agrees with human judgment.

2 Related Work

Benchmark Construction for Emerging Topics and Weak Signals. Existing benchmarks for emerging-topic analysis mainly evaluate whether systems can detect and track salient events or trends over time, rather than identify early scientific ideas that later develop into dominant paradigms (Allan, 2002; Petrović et al., 2010; Deng et al., 2022). In scientific domains, benchmark construction has also focused on trajectory labeling or future impact prediction, such as classifying topics as rising or declining or forecasting citations (Prabhakaran et al., 2016; Moiseeva and Schütze, 2020; Ofer et al., 2024; Gu and Krenn, 2024; Ajith et al., 2026). Foresight and horizon-scanning reports offer another related resource, but the signals they curate are usually broad thematic areas rather than concrete technical precursors (Day and Schoemaker, 2005; Saritas, 2013; van Rij, 2010). BackTrend addresses these limitations by constructing benchmark instances retrospectively from mature research topics, defining weak signals as empirically grounded early precursors rather than transient events, broad themes, or short-term impact patterns. Each instance decomposes into problem-space and solution-space precursors and is evaluated under a prediction cutoff that withholds post-cutoff information from the system.

Weak Signal Detection and Emerging Trend Analysis. Weak-signal detection has long been studied in foresight, bibliometrics, and technology intelligence as the task of identifying faint early indicators of future change (Ansoff, 1975; Hiltunen, 2008; Holopainen and Toivonen, 2012). Early computational approaches relied on keyword frequencies, citation networks, and co-occurrence statistics, while later work adopted topic models and contextual embedding methods to better capture latent semantic structure and thematic evolution over time (Small, 1973; Yoon, 2012; Song et al., 2018; Blei et al., 2003; Rudolph and Blei, 2018; Yao et al., 2018; Grootendorst, 2022; Boutaleb et al., 2024; Ebadi et al., 2026). However, these methods are still typically evaluated through case studies or proxy trend metrics, because existing resources rarely provide verified examples of concrete early precursors to mature scientific themes. By furnishing such a reference set, BackTrend enables direct comparison between modern LLM-based research agents and traditional bibliometric or topic-modeling systems on the same precursor-recovery

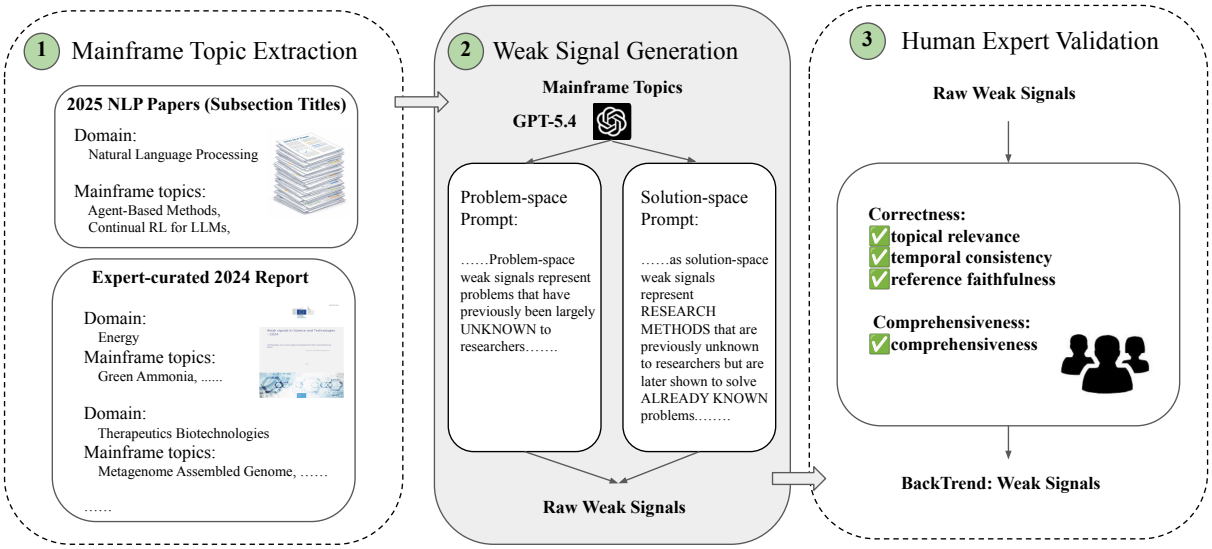

Figure 2: Overview of the BackTrend construction pipeline. We first compile mature target topics from a 2024 JRC report (O et al., 2025) and additional manually curated NLP topics. A DeepResearch agent then generates problem-space and solution-space weak signals grounded in Semantic Scholar-based web search. Finally, three PhD-level domain experts among the authors validate the candidates for relevance, temporal consistency, reference faithfulness, and coverage to produce the final BackTrend benchmark.

task under a unified evaluation protocol.

3 BackTrend Benchmark

BackTrend is a retrospective weak-signal prediction benchmark in which systems recover early problem-space and solution-space precursors of mature research topics. Below we define the task and the two weak-signal categories (§3.1), detail the LLM-assisted construction pipeline with expert validation (§3.2), and report per-domain statistics for the resulting benchmark (§3.3).

3.1 BackTrend Task

Mature Target Topics. Given a scientific domain, mature target topics serve as anchor concepts for weak-signal construction. Each mature target topic represents a research direction that has already achieved substantial visibility within its domain through sustained publication activity.

Problem-space and Solution-space Weak Signals. We define a weak signal as a concrete research direction that initially received limited attention but later grew significantly in prominence. We distinguish two types: (1) *problem-space weak signals* are underrecognized research problems or problem formulations that later become central to the mature target topic; (2) *solution-space weak signals* are early methods, techniques, or design principles that were not yet widely adopted but later became important solutions to already recognized problems. In short, problem-space signals surface new questions, while solution-space signals offer emerging answers to existing ones. Concretely, each weak signal in BackTrend consists of four fields: a name, a technical description, a rationale explaining why it qualifies as a weak signal, and a list of supporting reference papers (see §3.2).

Task Formulation. We formulate weak-signal discovery as the task of identifying historically grounded early research directions that later contribute to the emergence of a given mature target topic. Formally, given a domain-specific mature target topic M, a time window t, and a designated signal space (problem or solution), the model must retrieve or infer a set of candidate weak signals $s = \{s_1, \ldots, s_n\}$. A valid weak signal $s_i \in s$ must exhibit low visibility during its early stage while demonstrating substantial growth in prominence over subsequent years. The time window t in BackTrend is 2023–2024, and we refer to its upper bound (end of 2024) as the *prediction cutoff year*: at evaluation time, systems may only use information available up to this date, simulating a research-foresight setting in which M's eventual maturity is not yet known to the model.

3.2 Benchmark Construction

We next detail the BackTrend construction process, with an overview shown in Figure 2.

Mature Target Topic Extraction. To systematically construct this topic set, we draw on two complementary sources. First, we derive both the high-level scientific domains and a subset of specific topics from an authoritative 2024 report (O et al., 2025). It leverages large-scale data analysis and expert validation to identify weak signals in science and technology, providing well-curated summaries of both emerging and established research frontiers. Second, we manually curate 32 additional topics from the NLP and LLM domains to ensure adequate coverage of the field most central to our study. Together, these sources yield a benchmark spanning 13 broad scientific domains with diverse topical coverage.

Weak Signal Generation. For each mature target topic and its associated domain, we use the DeepResearch Agent (Claude-4.6-Opus with agentic web search) to generate weak signals along the problem-space and solution-space dimensions. We construct two structured prompts per topic, one for each space (Appendix Figure 3 and Figure 4), each instructing the model to act as an expert analyst in the corresponding domain. To ensure the factual grounding of each generated signal, we equip the model with a web search tool backed by the Semantic Scholar API. During generation, the model autonomously issues multiple search queries to retrieve and verify real academic papers, and is explicitly instructed to provide exact paper titles, publication years, and URLs as references. We additionally instruct the model to avoid generating generic techniques applicable to any domain (*e.g.,* "manifold optimization," "benchmark datasets") and purely application-oriented labels (*e.g.,* "for medical imaging," "for finance"). The model returns a structured output containing, for each weak signal, its name, a concise technical description of the signal, an explanation of why it qualifies as a weak signal for the given mature target topic, and a list of verified references. This process is applied to each of the 253 mature target topics across 13 scientific domains, yielding a total of 994 problem-space weak signals and 1,033 solution-space weak signals. We use Claude-4.6-Opus only for benchmark candidate generation; none of the evaluated systems in Section 4.1 has access to the final validated reference signals, ensuring strict separation between the construction model and the models under evaluation.

Expert Validation. Because the initial weak-signal candidates are generated by an LLM-assisted research agent, we treat the generation stage only as a *candidate proposal* step: a candidate is included in BackTrend only after passing expert verification, and the final benchmark should be understood as a human-validated reference set constructed with LLM assistance, rather than as raw model output. Seven domain experts among the authors at post-doctoral and graduate-student levels (Appendix Table 6) independently review each generated weak signal together with its description, rationale, and supporting references; signals are assigned such that each is reviewed by at least one expert whose primary research area matches the signal's high-level domain category (AI, Science, Healthcare, or Engineering), with the two cross-domain reviewers handling overlap and arbitration. Reviewers verify four properties: (1) *topical relevance*, *i.e.,* whether the weak signal is genuinely related to the mature target topic and correctly categorized as problem-space or solution-space; (2) *temporal consistency*, *i.e.,* whether the cited references correspond to early-stage ideas that preceded the later prominence of the mature target topic, with reviewers applying the operational guideline that a valid signal's related publication count should grow at least fivefold from the 2023–2024 emergence window to 2025 (empirically verified in Appendix A.2); (3) *reference faithfulness*, *i.e.,* whether the cited papers are real and faithfully support the interpretation of the proposed signal; and (4) *coverage comprehensiveness*, *i.e.,* whether the weak-signal set provides broad coverage of salient early research directions, ensuring that no major precursor theme is systematically overlooked, while acknowledging that the set is not intended to be exhaustive. To assess reliability, each of 50 randomly sampled signals was independently re-validated by two of five participating experts; the resulting pairwise comparisons yield Cohen's $\kappa = 0.746$ on the overall accept/reject decision, indicating substantial agreement. Reviewers revise or reject candidates that fail any of these checks; in cases of disagreement, the instance is further discussed among the reviewers and conflicts are resolved through consensus-based adjudication. After validation, 93 of the 2,027 weak signals were revised, with the remainder accepted as correct, in-

Domain	#Topics	#P-WS	#S-WS	#Search
AI				
NLP	32	141	156	19.1
AI / ML	25	100	111	19.8
Science				
Advanced Materials	36	133	149	15.7
Energy	18	78	73	16.3
Environment & Agri.	17	56	64	16.7
Quantum & Crypto.	28	103	112	16.9
Healthcare				
e-Health	15	55	58	17.6
Med. Imaging	9	36	33	16.3
Therapeutics & Biotech	19	76	72	15.3
Engineering				
Aerospace	11	48	47	17.4
Digital Twins	9	37	31	16.5
Info. & Comm. Tech.	20	77	75	15.7
Mobility & Transport	14	54	52	18.4
Total / Avg.	**253**	**994**	**1,033**	**16.9**

Table 1: Data statistics of BackTrend across domains. **#Topics**: number of mature target topics. **#P-WS / #S-WS**: number of problem- / solution-space weak signals. **#Search**: average web searches issued per topic during dataset construction.

dicating that the LLM-assisted pipeline produces highly reliable candidates while expert verification further ensures conceptual precision, citation correctness, and thematic breadth.

3.3 Data Statistics

BackTrend covers 253 mature target topics and 2,027 weak signals across 13 domains organized into four high-level categories: AI (Natural Language Processing; Artificial Intelligence & Machine Learning), Science (Advanced Materials & Manufacturing; Energy; Environment & Agriculture; Quantum & Cryptography), Healthcare (e-Health; Medical Imaging; Therapeutics & Biotechnologies), and Engineering (Aerospace; Digital Twins; Information & Communication Technologies; Mobility & Transport). Per-domain statistics are reported in Table 1. At evaluation time, systems are asked to recover the weak signals associated with each mature target topic under a temporal search cut-off, using only historically available information, simulating prospective foresight.

4 Experiment Setup

We next discuss the evaluated systems and our automated and human evaluation protocols.

4.1 Evaluated Systems

In our experiments, We evaluate three categories of weak-signal prediction systems. All systems operate under the prediction cutoff year defined in §3.1 (end of 2024), enforced through prompt instructions for parametric systems and through hard filtering of retrieval results for systems that can search external corpora at inference time.

Frontier LLMs. We include GPT-5.4, Qwen3.5-397B-A17B (Qwen Team, 2026), and DeepSeek-R1-0528 (DeepSeek-AI, 2025). Temporal consistency is enforced via explicit prompt instructions specifying the emergence time of weak signals.

Retrieval-Augmented LLMs (RAG). We evaluate Qwen3-8B and Qwen3-30B (Team, 2025) with retrieval augmentation, which are equipped with retrieval mechanisms to better ground predictions in evidence from historical scientific literature. For each mature target topic, we retrieve a set of candidate papers from Semantic Scholar using the topic as a query. The retrieved papers are encoded using the bge-base-en-v1.5 embedding model (Xiao et al., 2024) and ranked by cosine similarity. The top 30 most relevant papers are then selected and provided as external context to support weak-signal prediction. To ensure temporal consistency, retrieved papers are filtered on the client side to include only those published before the prediction cutoff year.

DeepResearch Systems. We evaluate two agentic LLM systems, DR-Tulu-8B (Shao et al., 2025) and Tongyi-DeepResearch-30B-A3B (Tongyi-DR-30B-A3B) (Team et al., 2026), which integrate multi-step reasoning with tool-augmented academic retrieval and evidence collection. Tongyi-DR-30B-A3B follows a ReAct-style framework, where the model iteratively performs reasoning and tool calls to query Semantic Scholar and refine its predictions. In contrast, DR-Tulu adopts a workflow-based agentic search pipeline that orchestrates model inference and retrieval through a modular multi-service architecture. In both systems, given a mature target topic, the model retrieves relevant papers and generates weak signals grounded in the collected academic evidence. To ensure temporal consistency, retrieval is restricted to papers published before the prediction cutoff year, preventing information leakage from future publications during the multi-step search process.

Model	AI			Science			Healthcare			Engineering			Overall		
	Set	Signal	Human	Set	Signal	Human	Set	Signal	Human	Set	Signal	Human	Set	Signal	Human
GPT-5.4	53.8	56.8	38.9	**63.1**	60.3	**44.3**	**71.3**	**69.1**	**69.4**	**68.6**	**62.9**	**64.4**	**63.6**	**61.6**	**46.9**
Qwen3.5-397B	57.1	**60.9**	**43.1**	51.9	51.8	39.2	52.5	49.9	20.0	55.3	48.5	31.3	53.9	52.8	38.2
Tongyi-DR-30B-A3B	**58.4**	56.9	42.0	56.9	55.1	28.1	57.6	54.9	49.5	58.9	58.0	45.3	57.8	56.1	35.9
DR-Tulu-8B	58.1	57.8	20.1	62.0	**61.7**	31.3	56.6	55.8	51.1	62.3	61.4	43.8	60.3	59.7	30.8
DeepSeek-R1-0528	55.3	60.1	18.6	53.6	52.1	23.7	58.2	57.6	14.3	57.1	55.8	39.9	55.5	55.6	24.1
Qwen3-30B (RAG)	37.0	41.1	4.2	56.7	58.3	21.3	54.0	56.5	50.0	53.4	53.0	35.2	51.1	53.0	19.7
Qwen3-8B (RAG)	31.6	37.2	5.1	55.4	58.6	17.4	51.1	51.0	45.8	49.7	51.3	11.1	48.1	50.9	14.2

Table 2: Main results on weak-signal prediction across four domain groups (AI, Science, Healthcare, and Engineering). We report set-level and signal-level LLM F1, along with human F1. Overall denotes the average across domains. Higher is better. The best score in each column is highlighted in bold, and the second best is underlined. Rows are ordered by Overall Human F1.

4.2 Evaluation Protocol

For each model, we provide a mature target topic M in a specific domain together with a predefined historical time window t (spanning the years 2023–2024), and ask the model to predict the weak signals that emerged during that period; the prediction prompt templates for the problem-space and solution-space settings are shown in Appendix Figure 5 and Figure 6. Let the predicted weak-signal set be $s = \{s_1, \ldots, s_m\}$ and the human-validated reference set be $g = \{g_1, \ldots, g_n\}$. Our main table reports F1 under two LLM-based evaluation settings and a human evaluation, all using an inclusive semantic-overlap criterion: two signals are considered matched if they substantially overlap in research question, core method, or described phenomenon, even if their wording differs. All LLM-based evaluations use Gemini-3-Flash as the judge, chosen from a third model family to remain independent of both the benchmark constructor (§3.2) and the evaluated systems (§4.1); the full set-level and signal-level judging prompts are shown in Figure 7 and Figure 8.

In **set-level LLM evaluation**, we provide the entire predicted set s and reference set g jointly to an LLM judge, which directly estimates set-level precision and recall. This measures whether the predicted set is semantically aligned with the reference set as a whole. In **signal-level LLM evaluation**, we evaluate coverage at the level of individual signals. For each predicted signal s_i, the judge determines whether it matches any reference signal; symmetrically, for each reference signal g_j, it determines whether any prediction covers it. Precision and recall are computed as the fraction of matched signals in each direction. This provides a finer-grained measure of whether individual weak signals are successfully discovered. Since LLM judgments are stochastic, both settings are repeated for $K = 5$ runs and scores are averaged.

For **BERTScore-based F1**, the binary match indicators in both set-level and signal-level evaluation are replaced by continuous BERTScore similarities; results and comparison against LLM and human evaluation are deferred to §5.3.

Finally, we report a **human-judged F1**. For each predicted signal s_i, the human evaluator is required to judge whether it semantically matches any reference signal in g; symmetrically, for each reference signal g_j, they judge whether it is covered by any prediction in s. Precision and recall are computed as the fraction of matched signals in each direction.

5 Experiments

We use BackTrend to evaluate frontier LLMs, RAG systems, and DeepResearch agents along three axes: overall accuracy across the four domain groups (§5.1), the failure modes that arise when predictions miss the reference set (§5.2), and additional analyses covering retrieval, search budgets, the choice of evaluation metric across BERTScore, LLM, and human judgment (§5.3).

5.1 Main Results

Table 2 shows that BackTrend is challenging even for strong LLM and DeepResearch systems. Although several models achieve moderate LLM-based F1 scores, human-judged F1 remains substantially lower across nearly all systems, indicating that many predictions are plausible but fail to match the benchmark's intended precursor framing. The central finding is therefore not simply that GPT-5.4 obtains the highest overall score, but

that BackTrend probes a capability distinct from generic scientific fluency: successful weak-signal prediction requires temporal grounding, semantic boundary control, abstraction-level calibration, and cross-domain coverage, rather than the ability to generate topically relevant scientific phrases.

Cross-domain robustness, more than single-domain strength, drives overall performance. GPT-5.4 obtains the highest overall score across the entire evaluated benchmark (63.6 set-level LLM F1, 61.6 signal-level LLM F1, 46.9 human F1), largely because it remains stable outside AI rather than by dominating any single slice. It ranks first on all three metrics in both Healthcare and Engineering, and also leads Science on set-level LLM F1 and human F1 while placing a close second on signal-level LLM F1. It does not, however, lead the AI slice, which is itself the most heterogeneous category: Tongyi-DR-30B-A3B obtains the best AI set-level LLM F1 (58.4), Qwen3.5-397B obtains the best AI signal-level LLM F1 (60.9) and best AI human F1 (43.1), and DR-Tulu-8B ranks a close second on AI set-level LLM F1 (58.1). This dispersion contrasts with Healthcare and Engineering, where top positions are more concentrated, suggesting that performance on BackTrend hinges on broad cross-domain consistency rather than single-slice dominance within any one category.

Automatic metrics are useful but incomplete substitutes for human judgment. Human-judged F1 is markedly lower than LLM-based F1 for nearly every system evaluated in our experiments, and the human ranking partially reshuffles the LLM-based ranking outside the top position. This gap suggests that automatic judges often reward broad topical overlap, whereas human annotators apply a stricter notion of benchmark equivalence: a generated signal may be scientifically plausible and lexically related to the target topic, yet still fail if it shifts the problem framing, uses a different abstraction level, or omits a central precursor dimension. We quantify this further in Section 5.3 by adding BERTScore as a third evaluation regime.

5.2 Error Analysis and Case Study

We analyze all 506 topic-direction pairs generated by GPT-5.4 against the benchmark weak signals, and identify the following four main error types:
(1) Topic drift. Topic drift refers to cases where the prediction stays near the target topic but shifts to an adjacent subproblem. For example, in Building Decarbonization (problem space), the benchmark focuses on electrification bottlenecks and retrofit trade-offs, whereas the model instead predicts challenges related to bio-based material supply.
(2) Granularity mismatch. Granularity mismatch refers to cases where the prediction is plausible but too narrow relative to the benchmark. In Truck–Drone (problem space), the benchmark emphasizes logistics-level issues such as rendezvous and coordination, whereas the model predicts engineering-level concerns such as ground-effect instability and rotor-wash interaction.
(3) Lexical near-miss. Lexical near-miss refers to cases where the prediction is topically close to the benchmark but semantically different. For example, in Quantum-Resistant Algorithms (problem space), the benchmark emphasizes migration and deployment risks, whereas the model predicts issues related to proof models and attack mechanisms.
(4) Coverage failure. Coverage failure refers to cases where the model recovers one slice of the benchmark but misses other central dimensions. For instance, in City Digital Twin (problem space), the benchmark includes governance, citizen inclusion, and organizational barriers, whereas the model primarily predicts technical fidelity issues.

These four types of errors indicate that the central challenge is *semantic alignment with the intended precursor structure* rather than surface fluency: models must match the right problem framing, abstraction level, and coverage scope. Additional representative cases across all four error types are provided in Appendix A.4.

	$k=10$		$k=30$		$k=50$	
Model	**Set**	**Signal**	**Set**	**Signal**	**Set**	**Signal**
Qwen3-30B (RAG)	37.4	48.8	**49.2**	51.5	46.4	**54.7**
Qwen3-8B (RAG)	**42.7**	**50.5**	40.4	43.1	41.5	44.4

Table 3: Effect of retrieval budget (k) on RAG performance. Results are reported as F1 (%). Larger models benefit from increased retrieval budget, whereas smaller models perform best at lower k. Signal-level evaluation consistently yields higher scores.

5.3 Analysis

We conduct three studies probing to better understand what drives BackTrend performance and how trustworthy our metrics are: (i) the retrieval budget in RAG, controlled by the number of retrieved documents; (ii) the number of web searches in the multi-step DeepResearch system; and (iii) the

Budget	Set F1	Signal F1
$b = 0$	46.7	45.7
$b = 1$	**51.0**	**53.4**
$b = 3$	48.8	43.4
$b = 8$	40.9	39.8

Table 4: Effect of web search budget (b) on Tongyi-DR-30B-A3B. Results are reported as F1 (%). Performance peaks at $b = 1$ and degrades with larger budgets in this setting, hinting that search quality may matter more than search volume. Signal-level evaluation is generally consistent with set-level trends.

Model	BS F1	LLM F1	Human F1
GPT-5.4	86.8	**61.6**	**46.9**
Qwen3.5-397B	85.5	52.8	38.2
Tongyi-DR-30B-A3B	86.0	56.1	35.9
DR-Tulu-8B	85.9	59.7	30.8
DeepSeek-R1-0528	**86.9**	55.6	24.1
Qwen3-30B (RAG)	**86.9**	53.0	19.7
Qwen3-8B (RAG)	86.0	50.9	14.2

Table 5: Overall F1 under three evaluation regimes: signal-level BERTScore (BS), signal-level LLM judge (LLM), and human annotation (Human). BERTScore is poorly discriminative; LLM-judge ranking partially reshuffles under human evaluation. Per-domain breakdown is in Appendix Table 7.

agreement between automatic (BERTScore, LLM judge) and human evaluation on the main results.

Retrieval budget in RAG. We ablate the retrieval budget in RAG by varying the number of retrieved documents. This study examines how performance changes with more or fewer retrieved documents, revealing whether gains come from additional evidence or more efficient use of a smaller set. Table 3 reports RAG performance under different retrieval budgets. For Qwen3-30B (RAG), increasing the number of retrieved documents from $k = 10$ to $k = 30$ improves performance, with signal-level F1 peaking at $k = 50$. In contrast, Qwen3-8B (RAG) achieves its highest performance at $k = 10$, with further increases in retrieval depth leading to a decline. Across all settings, signal-level F1 consistently exceeds set-level F1, suggesting that signal-level evaluation is more tolerant of partial semantic matches.

Web search rounds for DeepResearch. We ablate the web search budget in DeepResearch by varying the number of allowed search rounds (b). This study examines how performance changes as the model is given more opportunities to issue search queries. Table 4 reports the results. Increasing the search budget from $b = 0$ to $b = 1$ improves both set-level and signal-level F1. However, further increasing b leads to consistent performance degradation, with the lowest scores observed at $b = 8$. For Tongyi-DR-30B-A3B, a small search budget performs best in our setting; larger budgets appear to introduce less relevant evidence or distract the generation process, suggesting that search quality may matter more than search volume.

Agreement between automatic and human evaluation. We compare three evaluation regimes on the overall benchmark: signal-level BERTScore (BS), signal-level LLM judge (LLM), and human expert annotation (Human). Table 5 shows that BERTScore is poorly discriminative: nearly all of the evaluated systems fall within roughly two points of each other (84.2–86.9) despite human F1 ranging from 14.2 to 46.9. LLM judges spread the systems more (50.9–61.6) but still compress the differences relative to human evaluation, and the human ranking reshuffles the middle of the LLM ranking (*e.g.,* Qwen3.5-397B jumps from sixth on LLM F1 to second on Human F1). Embedding similarity therefore captures surface coherence rather than precursor matching, and LLM-based judging is useful as a scalable proxy but only partially aligns with expert annotation.

6 Conclusion

Our work shows that weak-signal prediction should be evaluated not by whether a model can generate plausible-sounding early ideas, but by whether it can recover the specific precursors that later proved important. Our results suggest that this remains difficult for current systems: many predictions are relevant on the surface, yet fail because they drift toward adjacent topics, operate at the wrong level of granularity, or capture only part of the benchmark's intended signal set. Strong performance depends on robustness across heterogeneous domains rather than single-slice dominance, adding more retrieval or web search is not always beneficial, and the gap between LLM-based evaluation and human judgment indicates that scalable automatic evaluation is useful but still incomplete as a substitute for careful expert assessment of weak-signal matching.

7 Limitations

BackTrend's 253 mature target topics are drawn from the 2024 JRC weak-signal report and a curated NLP/LLM supplement (§3.2), providing a controlled but bounded view of established research themes, so conclusions drawn from BackTrend should be read as evidence about how systems recover precursors of well-documented topics rather than a measure of free-form scientific foresight, and extending the topic pool to longer-tail, interdisciplinary, or pre-paradigmatic areas is a promising future direction. Reference signals are nominated by a single DeepResearch agent (Claude-4.6-Opus) before expert validation, which could introduce constructor-specific phrasing or abstraction biases; we mitigate this by grounding every nominated signal in verifiable Semantic Scholar references and by evaluating 5 of 7 systems from non-Claude model families (§3.2), and a direct sensitivity study with alternative constructor families is a natural next step. Following the practice of large-scale benchmark efforts (Hendrycks et al., 2021; Srivastava et al., 2023; Rein et al., 2023), all validators are co-authors, and we reduce single-reviewer bias by stratifying the panel across career stages and domain expertise, running independent reviews before adjudication, and reporting an inter-annotator agreement of Cohen's $\kappa = 0.746$ on a 50-signal overlap subset (§3.2); recruiting a broader pool of external domain experts, particularly for less-represented fields, is a useful future extension.

References

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. 2026. Prescience: A benchmark for forecasting scientific contributions. *Preprint*, arXiv:2602.20459.

James Allan. 2002. Topic detection and tracking: event-based information organization.

H. Igor Ansoff. 1975. Managing strategic surprise by response to weak signals. *California Management Review*, 18(2):21–33.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Allaa Boutaleb, Jerome Picault, and Guillaume Grosjean. 2024. Bertrend: Neural topic modeling for emerging trends detection. *Preprint*, arXiv:2411.05930.

George S. Day and Paul J. H. Schoemaker. 2005. Scanning the periphery. *Harvard Business Review*, 83(11):135–140, 142, 144–148.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, Xiang Chen, and Tianhua Zhou. 2022. Title2event: Benchmarking open event extraction with a large-scale chinese title dataset. *Preprint*, arXiv:2211.00869.

Ashkan Ebadi, Alain Auger, and Yvan Gauthier. 2026. Wisdom: An ai-powered framework for emerging research detection using weak signal analysis and advanced topic modelling. *Journal of Informetrics*, 20(1):101759.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Xuemei Gu and Mario Krenn. 2024. Impact4cast: Forecasting high-impact research topics via machine learning on evolving knowledge graphs. In *ICML 2024 AI for Science Workshop*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Elina Hiltunen. 2008. The future sign and its three dimensions. *Futures*, 40:247–260.

Mari Holopainen and Marja Toivonen. 2012. Weak signals: Ansoff today. *Futures*, 44(3):198–205. Special Issue: Weak Signals.

Yuya Kajikawa, Junta Yoshikawa, Yoshiyuki Takeda, and Katsumori Matsushima. 2008. Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, 75(6):771–782.

Sangjae Lee, Wanki Kim, Young Min Kim, Hyoung Yong Lee, and Kyong Joo Oh. 2014. The prioritization and verification of it emerging technologies using an analytic hierarchy process and cluster analysis. *Technological Forecasting and Social Change*, 87(C):292–304.

Zhenfeng Liu, Jian Feng, and Lorna Uden. 2023. Technology opportunity analysis using hierarchical semantic networks and dual link prediction. *Technovation*, 128:102872.

Alena Moiseeva and Hinrich Schütze. 2020. Trendnert: A benchmark for trend and downtrend detection in a scientific domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8512–8519.

Eulaerts O, Grabowska M, and Bergamini M. 2025. Weak signals in science and technologies - 2024. (KJ-01-25-060-EN-N (online)).

Dan Ofer, Hadasah Kaufman, and Michal Linial. 2024. What's next? forecasting scientific research trends. *Heliyon*, 10(1):e23781.

Takaya Ogawa and Yuya Kajikawa. 2015. Assessing the industrial opportunity of academic research with patent relatedness: A case study on polymer electrolyte fuel cells. *Technological Forecasting and Social Change*, 90:469–475.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, Los Angeles, California. Association for Computational Linguistics.

Alan L. Porter, W. Bradford Ashton, Guenter Clar, Joseph F. Coates, Kerstin Cuhls, Scott W. Cunningham, Ken Ducatel, Patrick van der Duin, Luke Georghiou, Theodore Gordon, Harold Linstone, Vincent Marchau, Gilda Massari, Ian Miles, Mary Mogee, Ahti Salo, Fabiana Scapolo, Ruud Smits, and Wil Thissen. 2004. Technology futures analysis: Toward integration of the field and new methods. *Technological Forecasting and Social Change*, 71(3):287–303.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

Haiying Ren and Yuhui Zhao. 2021. Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technovation*, 101:102196.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ozcan Saritas. 2013. *Systemic Foresight Methodology*, pages 83–117. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, and 1 others. 2025. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*.

Henry Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265 – 269.

Kisik Song, Kyuwoong Kim, and Sungjoo Lee. 2018. Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. *Technological Forecasting and Social Change*, 128(C):118–132.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, and 38 others. 2026. Tongyi deepresearch technical report. *Preprint*, arXiv:2510.24701.

Victor van Rij. 2010. Joint horizon scanning: identifying common strategic choices and questions for knowledge. *Science and Public Policy*, 37(1):7–18.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM 2018, page 673–681. ACM.

Janghyeok Yoon. 2012. Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications*, 39:12543–12550.

Janghyeok Yoon and Kwangsoo Kim. 2012. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 90:445–461.

A Appendix

A.1 Expert Information

In this section, we provide an overview of the experts that contribute to the validation process. For anonymity reasons, we do not list the experts' names. All experts are also authors of the paper, so we do not provide monetary compensation to the experts.

A.2 Empirical Validation of the Growth Criterion

To verify that expert temporal-consistency judgments align with measurable publication trends, we conducted a post-hoc check on a random sample of 50 expert-validated weak signals. For each signal, we queried the Semantic Scholar API using the signal name as a search string and computed its growth ratio between the 2023–2024 emergence window and 2025. We find that 47 of the 50 sampled signals (94%) exhibit at least a fivefold publication-count increase. The remaining three showed real but slower growth (typically 3–4× within the one-year follow-up window), reflecting the conservativeness of the fivefold threshold rather than a failure of expert validation. Notably, every sampled signal exhibited positive growth, which independently corroborates the reliability of the expert validation step.

A.3 BERTScore Results

BERTScore-based Evaluation Details. For completeness, we additionally evaluate weak-signal discovery using two BERTScore-based settings, which complement the LLM-based metrics reported in the main text. Let the predicted weak-signal set be $s = \{s_1, \ldots, s_m\}$ and the reference set be $g = \{g_1, \ldots, g_n\}$. Unlike the LLM-based settings, which rely on explicit semantic judgments, the BERTScore-based settings use contextual embedding similarity to quantify semantic overlap.

In *set-level BERTScore*, we first collapse each set into a single text sequence by concatenating all signals with a separator:

$$\mathrm{Concat}(s) = s_1 \oplus s_2 \oplus \cdots \oplus s_m,$$
$$\mathrm{Concat}(g) = g_1 \oplus g_2 \oplus \cdots \oplus g_n,$$

where $\oplus$ denotes concatenation. We then compute BERTScore between the aggregated prediction and the aggregated reference:

$$(P_{\text{set-bs}}, R_{\text{set-bs}}, F1_{\text{set-bs}}) = \mathrm{BERTScore}\big(\mathrm{Concat}(s), \mathrm{Concat}(g)\big).$$

This setting measures coarse-grained similarity between the two sets as wholes, but does not explicitly model one-to-one or one-to-many correspondence between individual weak signals.

In *signal-level BERTScore*, we instead compare weak signals individually. We first compute the full pairwise similarity matrix

$$S_{ij} = \mathrm{BERTScore}_{F1}(s_i, g_j), \quad i = 1, \ldots, m,\ j = 1, \ldots, n,$$

where each entry quantifies the semantic similarity between predicted signal s_i and reference signal g_j. We then aggregate this matrix directionally. Precision is defined as the average best-match similarity from each prediction to the reference set:

$$P_{\text{sig-bs}} = \frac{1}{m} \sum_{i=1}^{m} \max_{1 \le j \le n} S_{ij}.$$

Intuitively, this asks: for each predicted signal, how well does its closest reference counterpart match? Recall is defined symmetrically as the average best-match similarity from each reference signal to the prediction set:

$$R_{\text{sig-bs}} = \frac{1}{n} \sum_{j=1}^{n} \max_{1 \le i \le m} S_{ij}.$$

This asks: for each reference weak signal, how well is it covered by the best prediction? We then combine the two using the harmonic mean:

$$F1_{\text{sig-bs}} = \frac{2\, P_{\text{sig-bs}} R_{\text{sig-bs}}}{P_{\text{sig-bs}} + R_{\text{sig-bs}}}.$$

Compared with set-level BERTScore, this signal-level variant provides a finer-grained view of coverage by rewarding predictions that closely match individual reference weak signals rather than only the overall set semantics.

In summary, the four evaluation settings used in this work are: set-level BERTScore, signal-level BERTScore, set-level LLM judgment, and signal-level LLM judgment. The main text reports the two LLM-based F1 scores together with human-judged F1, while this appendix provides the BERTScore-based results and the full precision/recall breakdowns.

Expert	Level	Domain	Contribution
Expert 1	Postdoctoral researcher	AI, Science, Engineering	Ground-truth signal validation; model comparison
Expert 2	4th-year PhD student	AI	Ground-truth signal validation; model comparison
Expert 3	2nd-year PhD student	AI	Ground-truth signal validation; model comparison
Expert 4	1st-year PhD student	Science, Healthcare, Engineering	Ground-truth signal validation; model comparison
Expert 5	1st-year PhD student	Science	Ground-truth signal validation; model comparison
Expert 6	2nd-year MS student	Healthcare	Ground-truth signal validation
Expert 7	2nd-year MS student	Engineering	Ground-truth signal validation

Table 6: Experts involved in validation and evaluation.

	AI			Science			Healthcare			Engineering			Overall		
Model	**Set**	**Signal**	**Human**	**Set**	**Signal**	**Human**	**Set**	**Signal**	**Human**	**Set**	**Signal**	**Human**	**Set**	**Signal**	**Human**
GPT-5.4	**85.8**	86.2	38.9	**86.4**	86.9	**44.3**	86.7	86.8	**69.4**	86.9	87.3	**64.4**	**86.5**	86.8	**46.9**
Qwen3.5-397B	84.3	85.6	**43.1**	84.0	85.4	39.2	83.8	85.2	20.0	84.6	85.8	31.3	84.2	85.5	38.2
Tongyi-DR-30B-A3B	85.2	85.8	42.0	85.0	85.9	28.1	85.1	85.8	49.5	85.5	86.4	45.3	85.2	86.0	35.9
DR-Tulu	84.4	85.4	20.1	84.7	86.0	31.3	84.5	85.6	51.1	85.3	86.4	43.8	84.7	85.9	30.8
DeepSeek-R1-0528	**85.8**	**86.5**	18.6	86.3	86.9	23.7	**86.9**	**86.9**	14.3	87.1	87.5	39.9	**86.5**	**86.9**	24.1
Qwen3-30B (RAG)	85.5	86.4	4.2	**86.4**	**87.1**	21.3	86.2	86.7	50.0	**87.1**	**87.6**	35.2	86.3	**86.9**	19.7
Qwen3-8B (RAG)	83.6	85.4	5.1	84.7	86.2	17.4	84.4	85.7	45.8	85.1	86.6	11.1	84.5	86.0	14.2

Table 7: Appendix results on weak-signal prediction across four domain groups (AI, Science, Healthcare, and Engineering), using set-level and signal-level BERTScore F1 together with human F1. Overall denotes the average across domains. Higher is better. The best score in each column is highlighted in bold, and the second best is underlined. Rows are ordered by Overall Human F1.

A.4 Additional Error Cases

This appendix expands the discussion in Section 5.2 by providing concrete benchmark–prediction comparisons. The goal is to show that many low-scoring cases are not random failures, but systematic mismatches in framing, abstraction level, or semantic coverage. In particular, several examples are highly plausible on their own terms, which helps explain why lexical similarity can remain high even when semantic judge-based matching is low.

The examples in Table 8 follow the same four error types discussed in the main text. First, some cases exhibit *topic drift*, where the model stays within a reasonable neighboring area but shifts away from the benchmark's intended problem definition. Second, some cases show *granularity mismatch*, where the model predicts narrower mechanisms or engineering details while the benchmark is formulated at the level of research paradigms or system bottlenecks. Third, some examples are *lexical near-misses*: the predictions are clearly in the same field and share substantial vocabulary with the benchmark, but the actual research framing is different. Finally, some cases demonstrate *coverage failure*, where the model captures one dimension of the benchmark but omits other equally central dimensions.

Topic	Direction	Reference signal(s)	GPT-5.4 prediction(s)	Error interpretation
Building Decarbonization	Problem	Distribution-grid and service-capacity bottlenecks; retrofit-versus-redevelopment whole-life carbon tradeoffs; refrigerant-related emissions risk in heat-pump-led decarbonization	Soil carbon depletion from crop-residue harvesting; land-use competition from hemp/fiber crops; moisture and microbial risks in bio-based insulation	Topic drift to adjacent sustainability issues in bio-based materials; the predictions remain environmentally relevant, but miss the benchmark's core building-system bottlenecks.
Atmospheric Water Harvesting	Problem	Water-quality contamination in atmospheric-harvested water; outdoor reliability of sorbent-based AWH under real low-humidity cycling	Declining natural dew formation; uncaptured greenhouse humidity losses; fog-loss and transpiration-based atmospheric moisture signals	Topic drift from engineered AWH constraints to broader ecological and agricultural moisture phenomena.
Truck–Drone Systems	Problem	Truck as a mobile micro-fulfillment terminal; drone–truck rendezvous; urban access asymmetry; truck-assisted multi-drone coordination bottlenecks	Ground-effect instability; rotor-wash interaction with traffic and debris; dynamic payload shift; battery thermal stress	Granularity mismatch: predictions are plausible engineering concerns, but they do not match the benchmark's logistics-level abstraction.
City Digital Twins	Problem	Governance maturity; citizens as native entities; gap between city information models and true urban twins; social and organizational barriers; resilience-oriented requirements	Urban temporal desynchronization; invisible infrastructure coupling; citizen behavior drift; data provenance collapse; microclimate fidelity gaps	Coverage failure: predictions focus on technical fidelity while missing governance, citizen, and socio-organizational dimensions.
Quantum-Resistant Algorithms	Problem	Cryptographic agility as a prerequisite; migration-path visibility; side-channel fragility of PQC implementations; harvest-now-decrypt-later exposure	QROM fragility; hybrid quantum–classical lattice reduction threats; superposition decryption attacks; structural leakage in module-lattice cryptography	Lexical near-miss: semantically plausible PQC concerns, but framed around proof models and attack mechanics rather than migration and deployment risks.
Idea Mining from Internal Knowledge	Solution	GraphRAG over private corpora; hierarchical abstractive indexing; LLM-driven ontology / knowledge-graph construction; LLM-assisted thematic induction	Latent knowledge elicitation via contrastive decoding; self-refine / Reflexion loops; sparse autoencoder probing; activation steering; self-generated synthetic data	Topic drift toward internal reasoning and interpretability methods rather than corpus-grounded knowledge discovery over organizational documents.

Table 8: Representative benchmark–prediction comparisons for GPT-5.4 predictions. The examples illustrate that many low-scoring outputs are semantically plausible near-misses that diverge in framing, abstraction level, or benchmark coverage.

A.5 Prompts

In this section, we present comprehensive prompt examples that we use to obtain signals and conduct LLM-as-a-judge evaluations. We use NLP as our field, and the prompts can be easily adapted for other fields.

Construction Prompt for Problem-Space Signals

You are an expert analyst of {field} research. Your task is to identify early weak signals that later contributed to a specified mature target topic. We distinguish two categories of weak signals. The category you are asked to identify here is the "problem-space weak signal": an underrecognized research problem or problem formulation that was not yet widely recognized by the research community at the time it emerged, but that later became central to the mature target topic.

The complementary category is the "solution-space weak signal": an early research method, technique, or design principle that was not yet widely adopted but later became an important solution to an already-recognized problem. In short, problem-space signals surface new questions, while solution-space signals offer emerging answers to existing ones.

You will be given a mature target topic and asked to identify its problem-space weak signal predecessors. The question will have the following format:
What are the early problem-space weak signals in {field} that emerged between [2023–2024] and later contributed to the mature target topic [{mature_target_topic}]?

Regarding "weak signals" in general (not specific to either category):
Definition-wise: Weak signals are early research ideas that initially received limited attention but grew significantly in prominence throughout subsequent years, ultimately contributing to the current framing of the mature target topic [{mature_target_topic}].
Logistics-wise: A weak signal here is a concrete research direction or idea.

Avoid emitting topics that are:
- generic techniques that could apply anywhere (*e.g.,* "manifold optimization", "benchmark datasets"),
- purely application labels ("for medical imaging", "for finance", "for ..." in general)

You must use web search repeatedly to find and verify real academic papers that support each weak signal you identify. Do not rely solely on parametric memory: actively search the web for relevant publications to ensure all references are real, accurate, and verifiable.

Return only JSON in the following format:

```
{
  "weak_signals": [
    {
      "signal":
        "<name of the weak signal>",
      "what_it_was":
        "<describing what it was specifically>",
      "why_weak_signal":
        "<explaining why it was a problem-space weak signal
        for [{mature_target_topic}] specifically>",
      "references": [
        {
          "title":
            "<exact paper title>",
          "year":
            <4-digit year>,
          "url":
            "<paper URL>"
        }
      ]
    }
  ]
}
```

Rules for references:
- Each weak signal must include references.
- Use real papers only; do NOT invent titles, years, or URLs.
- The URL should be a direct paper link (*e.g.,* arxiv, doi) when possible.
- If a URL is unavailable, still provide exact title and year.

Figure 3: Construction prompt for problem-space signals.

Construction Prompt for Solution-Space Signals

You are an expert analyst of {field} research. Your task is to identify early weak signals that later contributed to a specified mature target topic. We distinguish two categories of weak signals. The complementary category to the one you are asked about here is the "problem-space weak signal": an underrecognized research problem or problem formulation that was not yet widely recognized by the research community at the time it emerged, but that later became central to the mature target topic.

The category you are asked to identify here is the "solution-space weak signal": an early research method, technique, or design principle that was not yet widely adopted but later became an important solution to an already-recognized problem. Unlike problem-space signals, the corresponding problems were already present and known to the research community; what was missing was an effective solution.

You will be given a mature target topic and asked to identify its solution-space weak signal predecessors. The question will have the following format:
What are the early solution-space weak signals in {field} that emerged between [2023–2024] and later contributed to the mature target topic [{mature_target_topic}]?

Regarding "weak signals" in general (not specific to either category):
Definition-wise: Weak signals are early research ideas that initially received limited attention but grew significantly in prominence throughout subsequent years, ultimately contributing to the current framing of the mature target topic [{mature_target_topic}].
Logistics-wise: A weak signal here is a concrete research direction or idea.

Avoid emitting topics that are:
- generic techniques that could apply anywhere (*e.g.,* "manifold optimization", "benchmark datasets"),
- purely application labels ("for medical imaging", "for finance", "for ..." in general)

You must use web search repeatedly to find and verify real academic papers that support each weak signal you identify. Do not rely solely on parametric memory: actively search the web for relevant publications to ensure all references are real, accurate, and verifiable.

Return only JSON in the following format:

```
{
  "weak_signals": [
    {
      "signal":
        "<name of the weak signal>",
      "what_it_was":
        "<describing what it was specifically>",
      "why_weak_signal":
        "<explaining why it was a solution-space weak signal
        for [{mature_target_topic}] specifically>",
      "references": [
        {
          "title":
            "<exact paper title>",
          "year":
            <4-digit year>,
          "url":
            "<paper URL>"
        }
      ]
    }
  ]
}
```

Rules for references:
- Each weak signal must include references.
- Use real papers only; do NOT invent titles, years, or URLs.
- The URL should be a direct paper link (*e.g.,* arxiv, doi) when possible.
- If a URL is unavailable, still provide exact title and year.

Going forward, the [{mature_target_topic}] handle in the question template above will be substituted one mature target topic at a time.

Figure 4: Construction prompt for solution-space signals.

Prediction Prompt for Problem-Space Signals

You are an expert analyst of frontier {domain} research. Your task is to identify early weak signals that later contributed to a specified mature target topic. We distinguish two categories of weak signals. The category you are asked to identify here is the "problem-space weak signal": an underrecognized research problem or problem formulation that was not yet widely recognized by the research community at the time it emerged, but that later became central to the mature target topic.

The complementary category is the "solution-space weak signal": an early research method, technique, or design principle that was not yet widely adopted but later became an important solution to an already-recognized problem. In short, problem-space signals surface new questions, while solution-space signals offer emerging answers to existing ones.

You will be given a mature target topic and asked to identify its problem-space weak signal predecessors. The question will have the following format:
What are the early problem-space weak signals in {domain} that emerged between [2023–2024] and later contributed to the mature target topic [{mature_target_topic}]?

Regarding "weak signals" in general (not specific to either category):
Definition-wise: Weak signals are early research ideas that initially received limited attention but grew significantly in prominence throughout subsequent years, ultimately contributing to the current framing of the mature target topic [{mature_target_topic}].
Logistics-wise: A weak signal here is a concrete research direction or idea.

Avoid emitting topics that are:
- generic techniques that could apply anywhere (*e.g.,* "manifold optimization", "benchmark datasets"),
- purely application labels ("for medical imaging", "for finance", "for ..." in general)

You must return ONLY valid JSON in exactly this format (no other text):

```
{"weak_signals": [
  {
    "signal": "<name of the weak signal>",
    "what_it_was": "<sentences describing what it was, including the year>",
    "why_weak_signal": "<sentences on why it was a problem-space weak signal for [{mature_target_topic}]>",
  }
]}
```

Rules:
- Return ONLY the JSON object above, no markdown fences, no explanation.
- Each weak signal must have all three fields: signal, what_it_was, why_weak_signal.
- Do not include generic techniques that could apply anywhere.

Figure 5: Prediction prompt template for problem-space weak signals.

Prediction Prompt for Solution-Space Signals

You are an expert analyst of frontier {domain} research. Your task is to identify early weak signals that later contributed to a specified mature target topic. We distinguish two categories of weak signals. The complementary category to the one you are asked about here is the "problem-space weak signal": an underrecognized research problem or problem formulation that was not yet widely recognized by the research community at the time it emerged, but that later became central to the mature target topic.

The category you are asked to identify here is the "solution-space weak signal": an early research method, technique, or design principle that was not yet widely adopted but later became an important solution to an already-recognized problem. Unlike problem-space signals, the corresponding problems were already present and known to the research community; what was missing was an effective solution.

You will be given a mature target topic and asked to identify its solution-space weak signal predecessors. The question will have the following format:
What are the early solution-space weak signals in {domain} that emerged between [2023–2024] and later contributed to the mature target topic [{mature_target_topic}]?

Regarding "weak signals" in general (not specific to either category):
Definition-wise: Weak signals are early research ideas that initially received limited attention but grew significantly in prominence throughout subsequent years, ultimately contributing to the current framing of the mature target topic [{mature_target_topic}].
Logistics-wise: A weak signal here is a concrete research direction or idea.

Avoid emitting topics that are:
- generic techniques that could apply anywhere (*e.g.,* "manifold optimization", "benchmark datasets"),
- purely application labels ("for medical imaging", "for finance", "for ..." in general)

You must return ONLY valid JSON in exactly this format (no other text):

```
{"weak_signals": [
  {
    "signal": "<name of the weak signal>",
    "what_it_was": "<sentences describing what it was, including the year>",
    "why_weak_signal": "<sentences on why it was a solution-space weak signal for [{mature_target_topic}]>",
  }
]}
```

Rules:
- Return ONLY the JSON object above, no markdown fences, no explanation.
- Each weak signal must have all three fields: signal, what_it_was, why_weak_signal.
- Do not include generic techniques that could apply anywhere.

Figure 6: Prediction prompt template for solution-space weak signals.

Set-Level LLM-as-a-Judge Prompt

System prompt.
You are a research direction evaluator. Be inclusive: if two research directions address substantially overlapping research questions or methods, consider them a match.

User prompt template.
I have two sets of research weak signals. Compare the predicted set against the reference set.

Rules:
- A match means two research directions substantially overlap in research question, core method, or described phenomenon.
- Different angles or terminology are acceptable if the overlap is substantial.
- Return set-level scores, not matched pairs.
- Precision = how well the predicted set is covered by the reference set.
- Recall = how well the reference set is covered by the predicted set.
- Both scores must be numbers between 0 and 1 inclusive.

Reference signals:
```
{gt_block}
```

Predicted signals:
```
{pred_block}
```

Return ONLY valid JSON:
```
{
  "precision": <float>,
  "recall": <float>
}
```

Figure 7: Set-level LLM-as-a-Judge prompt used for direct set-wise precision and recall estimation.

Signal-Level LLM-as-a-Judge Prompt

System prompt.
You are a research direction evaluator. Be inclusive: if two research directions address substantially overlapping research questions or methods, consider them a match.

User prompt template.
For each candidate research direction, decide if it substantially overlaps with any reference research direction.

Two research directions match if they address the same research question, use the same core method, or describe the same phenomenon, even if from different angles or with different terminology.

Reference research directions:
```
{ref_block}
```

Candidate research directions:
```
{cand_block}
```

Return ONLY JSON:
```
{"matches": [<int>, ...]}
```

with exactly {n_cand} elements.
1 = substantial overlap with any reference research direction, 0 = no overlap.

Figure 8: Signal-level LLM-as-a-Judge prompt used for pairwise matching evaluation.