ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Supplementary materials for

Zhipeng ZHU, Guo LI, Mark Xuefang ZHU, Shanshan KONG, Zhenyu WANG, Wenting XIONG, 2026. Evaluating large language models on the core cognitive capabilities expected of technology managers: a hierarchical benchmark. *ENGINEERING Information Technology & Electronic Engineering*, 27(10):260168.
https://doi.org/10.1631/ENG.ITEE.2026.0168

1 Evaluation protocol for LLM-as-a-Judge

Table S1 Task 2 auditing protocol via M-Prometheus-7B

Item	Description
Evaluation protocol	Guided by prompts, M-Prometheus-7B was employed at a temperature setting of 0.0 to provide reference-free direct evaluation of the alignment between each generated technical description and its actual TRL based on an expert-informed predefined rubric. The model scored the initial pool of 631 items on a 1–5 scale, yielding a score distribution (from 1 to 5) of 1, 26, 50, 543, and 11 items, respectively. By retaining only those items that achieved a score of 4 or higher, this auditing process ultimately yielded the final dataset of 554 items.
Rubric	1: Critical Failure: Contradicts NASA definitions* (e.g., wrong hardware fidelity for the TRL) or lacks evidence. 2: Insufficient: Vague. Mentions TRL but fails to define the specific "Relevant Environment" or "Hardware Fidelity" (e.g., Breadboard vs. Brassboard). 3: Borderline: Uses non-standard terminology or misses specific completion criteria like "scaling requirements" or "life-limiting factors". 4: Compliant: Clearly identifies correct Hardware Fidelity, Test Environment, and Integration Level. Evidence supports the TRL. 5: Exemplary: Perfectly aligned with NASA definitions. Explicitly details Hardware (e.g., "Engineering Unit"), Environment (e.g., "Thermal-Vac"), and verifies models.
Prompt template	**ROLE:** You are a NASA Independent Review Board expert conducting a strict Technology Readiness Assessment. **TASK:** Use the provided Rubric and NASA definitions to evaluate the quality of the generated Technical Description against the actual TRL. Return a JSON object with two fields: "Score" (integer 1-5) and "Feedback" (an evidence-backed justification citing specific gaps or alignments). **INPUT:** {Technical Description}; {TRL} **RUBRIC:** . . .
Evaluation example	**Technical Description:** An ammonia Loop Heat Pipe (LHP) prototype incorporating a passive Thermal Control Valve (TCV) was developed. This high-fidelity prototype was tested in a thermal vacuum chamber, a relevant environment. Testing successfully demonstrated a heat rejection conductance turn-down capability in excess of 50:1 and the system's ability to passively shut down under cold sink and low power conditions. The prototype also withstood nine freeze and thaw cycles without condenser deformation and was capable of restarting after exposure to a -120°C sink temperature. **TRL:** 6 **Score:** 5 **Feedback:** The technology description is exemplary and perfectly aligned with TRL 6 standards. It explicitly details: 1. Hardware Fidelity: "high-fidelity prototype" and "prototype" meet the TRL 6 requirement for a prototype or engineering unit demonstrating form, fit, and function. 2. Environment: "tested in a thermal vacuum chamber, a relevant environment" directly satisfies the TRL 6 relevant environment criterion. 3. Completion Criteria: The description verifies resilience to life-limiting mechanisms through specific tests: "heat rejection conductance turn-down capability in excess of 50:1", "passively shut down under cold sink and low power conditions", "withstood nine freeze and thaw cycles without condenser deformation", and "capable of restarting after exposure to a -120°C sink temperature". These tests demonstrate robustness against operational extremes and failure modes, fulfilling the TRL 6 requirement to verify technology resilience. The description uses standard terminology, provides concrete evidence, and aligns with the example structure in the standard.

*NASA TRL definitions from the TRA guide accompanied the scoring.

Table S2 Task 2 candidate LLM reasoning quality evaluation protocol via M-Prometheus-7B

Item	Description
Evaluation protocol	Guided by prompts, M-Prometheus-7B was employed at a temperature setting of 0.0 to provide ground-truth-anchored evaluation of the reasoning quality of each candidate LLM response against the provided technical description and its actual TRL on a standardized 1–5 scale. This evaluation employs an expert-informed four-dimensional rubric: correctness of the final conclusion relative to the ground truth, logical soundness of the derivation process, effective identification of key technical evidence, and completeness of coverage regarding the necessary TRL criteria*.
Rubric	1: The final conclusion fundamentally contradicts the ground truth. The derivation process is illogical or entirely absent. The model fails to identify any key technical evidence and demonstrates zero coverage of the necessary TRL criteria. 2: The conclusion is vague or deviates from the ground truth. The logical soundness of the derivation is weak. The model mentions technical concepts but fails in the effective identification of specific key technical evidence. Completeness of coverage regarding necessary TRL criteria is highly inadequate. 3: The conclusion is mostly correct relative to the ground truth, but the derivation process contains logical gaps or lacks rigorous deduction. The model identifies some technical evidence but misses certain critical elements. Coverage regarding the necessary TRL criteria is only partial or incomplete. 4: The final conclusion is correct relative to the ground truth. The derivation process is logically sound. The model demonstrates effective identification of key technical evidence and provides complete coverage regarding the necessary TRL criteria to support the assessment. 5: The final conclusion perfectly matches the ground truth. The derivation process is flawlessly logical and highly rigorous. The model exhibits exceptional and granular identification of key technical evidence, demonstrating exhaustive completeness of coverage regarding all necessary TRL criteria.
Prompt template	**ROLE:** You are an objective and highly analytical judge assistant. Your core philosophy is to evaluate logic strictly, fairly, and without score inflation. **TASK:** Use the provided Rubric and NASA definitions to evaluate the reasoning quality of the Candidate LLM Response against the provided Technical Description and TRL. Return a JSON object with two fields: "Score" (integer 1-5) and "Feedback" (an evidence-backed justification). **INPUT:** {Technical Description}; {TRL}; {Candidate LLM Response} **RUBRIC:** . . .

Continued on next page

Table S2 (Continued)

Item	Description
Evaluation example	**Technical Description:** A scaled, additively manufactured electric igniter for a hybrid rocket motor was developed and innovated upon. This breadboard-level component was validated through test firings on a rocket motor test stand, where it successfully demonstrated the capability for multiple (7) ignitions using the same unit. **TRL:** 4 **Candidate LLM Response:** The technology description is compliant with TRL 4 standards. It correctly identifies the hardware as a "breadboard-level component" (low fidelity, demonstrates function only), which aligns with NASA's "Breadboard" hardware fidelity requirement. The description specifies validation through "test firings on a rocket motor test stand", which corresponds to a "Laboratory Environment" as no environmental stress is mentioned. The integration level is appropriately described as a "component". Evidence of successful multiple ignitions (7) supports the validation claim. However, the description could be more explicit about the laboratory setting and avoid non-standard terms like "innovated upon" to achieve exemplary status. **Score:** 4 **Feedback:** The final conclusion that the description meets TRL 4 standards is correct relative to the ground truth. The derivation process is logically sound, successfully mapping the provided text to NASA's TRL definitions. The model demonstrates effective identification of key technical evidence by accurately extracting "breadboard-level component" (hardware fidelity), "rocket motor test stand" (laboratory environment), and the 7 successful ignitions (functional validation). It provides complete coverage regarding the necessary TRL criteria to support the assessment. It falls short of a perfect 5 because, as noted within the response itself, the explanation lacks the exceptional granularity and exhaustive completeness required to achieve exemplary status.

*NASA TRL criteria from the TRA guide accompanied the scoring.

Table S3 Task 3 candidate LLM reasoning quality evaluation protocol via M-Prometheus-7B

Item	Description
Evaluation protocol	Guided by prompts, M-Prometheus-7B was employed at a temperature setting of 0.0 to provide reference-free direct evaluation of the reasoning quality of each candidate LLM response against the provided question, background, resolution criteria, and retrieval date on a standardized 1–5 scale. This evaluation employs an expert-informed five-dimensional rubric: context fidelity, evidence extraction, logical combination, uncertainty consideration, and probability.
Rubric	**A: Context fidelity** 1 - Terrible: Severe detachment. Reasoning relies entirely on fabricated evidence or assumptions contradicting the provided context. 2 - Poor: Contains clear factual errors, misrepresentations of the source, or introduces unsupported claims that skew the analysis. 3 - Adequate: Basic compliance. Relies on the provided materials to formulate a correct baseline answer but may miss subtle nuances. Minor, non-critical extrapolations may be present. 4 - Good: High fidelity. All major claims are correctly sourced. Paraphrases effectively while maintaining strict factual accuracy. Does not explicitly critique the source material's limitations. 5 - Exceptional: Perfect fidelity. Every factual claim is explicitly anchored to specific source material. Demonstrates "extreme ownership" by identifying and flagging internal inconsistencies or ambiguities within the materials themselves. Zero ungrounded extrapolation. **B: Evidence extraction** 1 - Terrible: Evidence extracted is entirely irrelevant, trivial, or misses the core predictive task. 2 - Poor: Incomplete or superficial extraction. Fails to distinguish critical predictive signals from peripheral information. 3 - Adequate: Extracts surface-level, obvious evidence. Prioritization is often based on narrative sequence rather than predictive impact. 4 - Good: Comprehensive extraction of all critical explicit evidence. Demonstrates strong understanding of relative importance (signal vs. noise) but does not fully articulate structural missing information. 5 - Exceptional: Extracts explicit evidence and accurately infers latent variables/causal mechanisms. Actively seeks disconfirming evidence. Explicitly identifies what key missing information would decisively shift the prediction. **C: Logical combination** 1 - Terrible: No coherent reasoning structure. The argumentative logic is fundamentally flawed or self-contradictory. 2 - Poor: Disjointed listing of points with minimal synthesis. No clear or logical argumentative thread connecting the variables. 3 - Adequate: Linear listing and brief discussion of evidence. The "synthesis" is merely a summary of points. Weights are implied only by sequence or text emphasis. 4 - Good: Clear logical argument with an implicit weighting structure. Explains why evidence matters and how it combines qualitatively (e.g., "A strongly supports, but B tempers"). Coherent argumentative thread.

Continued on next page

Table S3 (Continued)

Item	Description
	5 - Exceptional: Transparent, step-by-step synthesis. Explains the interactive effects of evidence (e.g., how Risk A neutralizes Signal B). Explicitly defines the directional impact of each key variable on the outcome, forming a cohesive, logically rigorous narrative prior to assigning numbers.
	D: Uncertainty consideration
	1 - Terrible: Presents a strictly deterministic, single-path narrative with zero acknowledgment of alternative possibilities.
	2 - Poor: Token, superficial acknowledgment of uncertainty (e.g., a boilerplate disclaimer that "nothing is certain").
	3 - Adequate: Acknowledges generic uncertainties (e.g., "market conditions could change"). Treatment of risk is basic and checklist-like.
	4 - Good: Identifies and explores major known unknowns and standard counter-arguments. Discusses how specific uncertainties could alter the outcome.
	5 - Exceptional: Conducts a rigorous "pre-mortem." Constructs plausible, specific failure scenarios that are not mere negations of the main forecast. Explores second-order effects (e.g., "Event A triggers B, making C likely"). Distinguishes clearly between epistemic limits and aleatory randomness.
	E: Probability
	1 - Terrible: The probability is internally inconsistent with the text, completely unjustified, or mathematically impossible given the stated premises.
	2 - Poor: The probability appears plucked from thin air, with minimal linkage to the severity or weight of the evidence discussed.
	3 - Adequate: The probability is loosely connected to the text. The jump from qualitative discussion to the specific final number feels somewhat arbitrary.
	4 - Good: The probability is well-justified and aligned with the qualitative analysis. The link from argument to number is clear, though the mapping relies slightly on intuitive estimation rather than strict formal updating.
	5 - Exceptional: The final probability is the direct, unavoidable output of the stated logic. Provides a complete, falsifiable audit trail (e.g., pseudo-Bayesian updating from a stated prior). The justification firmly withstands a "why not +/- 10%?" challenge.
	Scoring philosophy and aggregation rule
	1. The "5" is Sacred: A score of 5 represents superforecaster-level analytical rigor. It must be earned through explicit causal reasoning, deep structural logic, and total mathematical transparency.
	2. "3" is the Competent Baseline: A score of 3 represents a solid, instruction-following response without major errors. It is the benchmark for standard capability.
	3. The Failure Floor: Any forecast whose internal logic or probability mapping is so fundamentally broken that it guarantees performance worse than random guessing—specifically, a predictive rationale that would mathematically yield a Brier Score > 0.25 in binary forecasting—must immediately receive a score of 1 in Dimension C and Dimension E.

Continued on next page

Table S3 (Continued)

Item	Description
	4. Aggregation rule: (1) The Overall Score is not a simple unconstrained average. Dimension A serves as an absolute epistemic baseline. (2) Veto Condition: If Dimension A is scored ≤ 2, the Overall Score is capped at 2, regardless of performance in other dimensions (a perfectly reasoned argument based on hallucinations is scientifically invalid). (3) Standard Aggregation: If Dimension A is ≥ 3, the Overall Score is calculated as an evenly weighted average: $Overall = (A + B + C + D + E)/5$. (4) Rounding Rule: To map the overall performance back to the qualitative 1-5 tier system, the standard aggregation score is rounded to the nearest whole integer.
Prompt template	**ROLE:** You are an objective and highly analytical judge assistant. Your core philosophy is to evaluate logic strictly, fairly, and without score inflation. **TASK:** Use the provided Rubric to evaluate the reasoning quality of the candidate LLM response against the provided Question, Background, Resolution Criteria, and Retrieval Date. Return a JSON object with two fields: "Score" (integer 1-5) and "Feedback" (an evidence-backed justification). **INPUT:** {Question}; {Background}; {Resolution Criteria}; {Retrieval Date}; {Candidate LLM Response} **RUBRIC:** . . .
Evaluation example	**Question:** Is the "Cheetah" coding model a diffusion model? **Background:** Based on the provided source data, there is no evidence that a coding model named "Cheetah" is a diffusion model. The sources reference multiple distinct entities with the "Cheetah" name, none of which are described as a diffusion-based coding model. The primary "Cheetah" referenced appears to be a large language model (LLM) focused on efficiency and inference optimization, not image generation. The briefing synthesizes the following: The most prominent "Cheetah" in the corpus is Cheetah-45B4, a 45-billion parameter language model discussed in an ICLR 2024 paper. Its focus is on algorithmic and hardware optimizations for efficient, high-throughput inference on server-side systems. Techniques like speculative decoding are relevant to its performance profile. Separately, the sources mention PixArt-α, which is explicitly identified as a Transformer-based text-to-image (T2I) diffusion model. The conflation likely arises from the generic codename "Cheetah" being used for different projects within the same conference proceedings. **Resolution Criteria:** This question resolves to YES if and only if, on or before the market's close date, a primary and authoritative source explicitly states or confirms that the "Cheetah" coding model is a diffusion model. **1. POSITIVE EVENT DEFINITION** A "positive event" is the public release of a statement from a primary and authoritative source. Acceptable sources include: • Official documentation: A technical report, research paper, model card, etc. • Official announcements: A press release, blog post, or news item. • Authoritative executive statement: A direct, on-the-record statement from a founder or lead researcher. **2. WHAT DOES NOT COUNT** The following will NOT be considered valid for a YES resolution: • Rumors, speculation, or analyses from third-party outlets.

Continued on next page

Table S3 (Continued)

Item	Description

• Inferences drawn from model behavior or code structure.

• Statements that describe the model using related but architecturally distinct terms (e.g., "autoregressive model").

• Statements that are ambiguous, unless explicitly equated with being a diffusion model.

• Leaked information.

3. EDGE CASES AND WORDING CONSIDERATIONS

• Model Name Specificity: Must refer specifically to the "Cheetah" coding model.

• Architectural Specificity: Must confirm it is a diffusion model.

• Timing: Must be timestamped on or before the market's closing date.

• Contradictory Information: Based on the final, non-retracted statement.

• "Coding Model" Definition: Interpreted as an AI model designed for generating/manipulating code.

Retrieval Date: 2025-10-12T00:09:11Z

Candidate LLM Response:

— *Evidence Analysis* —

Current State of Knowledge (as of Retrieval Date: 2025-10-12):

1. Identified "Cheetah" Models:

• Cheetah-45B4: A 45B LLM focused on efficient inference.

• Cheetah (HE optimization): Optimizations for Homomorphic Encryption.

• No evidence of a "Cheetah" coding model described as a diffusion model.

2. Architectural Evidence:

• All referenced "Cheetah" models are LLMs or optimization frameworks using speculative decoding.

3. Diffusion Model Context:

• PixArt-α is explicitly identified as a diffusion model in the text. The absence of similar terminology for any "Cheetah" entity strongly implies it is not a diffusion model.

— *Resolution Criteria Assessment* —

Critical Gaps:

1. No evidence of any "Cheetah" model being described as a diffusion model.

2. All available evidence points to LLM/autoregressive architectures.

— *Pre-mortem & Uncertainty Consideration* —

While the current evidence strongly dictates a NO resolution, epistemic humility requires a pre-mortem analysis. A YES resolution could theoretically occur under two specific failure scenarios:

1. Undiscovered Model: A separate, currently unreferenced AI lab is developing a proprietary diffusion-based coding model also named "Cheetah" in stealth mode.

2. Late Announcement: The creators of the existing "Cheetah" project announce a fundamentally new diffusion-based sub-module or pivot their architecture shortly before the market closes.

However, the likelihood of a diffusion architecture being used specifically for a *coding* model remains highly novel and structurally uncommon.

— *Probability Assessment* —

Continued on next page

Table S3 (Continued)

Item	Description

• Base Rate Considerations: Diffusion models are primarily used for image/video/audio generation, whereas autoregressive models dominate code generation.

• Evidence Weight: The combination of architectural mismatch and strong negative evidence (the text knows what a diffusion model is, e.g., PixArt-α, but doesn't apply it to Cheetah) heavily weights the outcome toward NO.

Given the complete absence of evidence and the domain implausibility of diffusion for code, the probability is extremely low. I will assign a small residual uncertainty to account for the pre-mortem scenarios (undiscovered model or late announcement).

Final Probability: 2%

Score: 5

Feedback:

1. Context fidelity: 5/5

This response demonstrates exceptional fidelity. Every factual claim is explicitly anchored:

- Correctly identifies Cheetah-45B4.
- Accurately describes the HE optimization framework.
- Properly references PixArt-α.

Demonstrates "extreme ownership" by explicitly acknowledging what is not present in the sources.

2. Evidence extraction: 5/5

Extracts evidence at multiple levels:

- Explicit evidence: All Cheetah references are LLM frameworks.
- Latent variables: Identifies architectural incompatibility.
- Causal mechanisms: Recognizes diffusion for code would be highly novel.
- Missing/Disconfirming evidence: Actively notes PixArt-α's description vs Cheetah's absence.

3. Logical combination: 4/5

Provides clear, logical reasoning with explicit weighting (Base rate, Evidence strength). It falls short of a perfect score because it doesn't provide formal Bayesian updating with specific likelihood ratios.

4. Uncertainty consideration: 5/5

Conducts an exemplary pre-mortem analysis:

- Specific failure scenarios (undiscovered model, late announcement).
- Quantified residual uncertainty (2%).
- Epistemic humility and second-order thinking.

5. Probability: 4/5

The 2% probability is well-justified (absence of evidence, domain implausibility). However, the mapping is not perfectly quantified. It represents informed judgment rather than precise mathematical calculation.

6. Overall evaluation

Continued on next page

Table S3 (Continued)

Item	Description
	This response exemplifies superforecaster-level reasoning. It demonstrates perfect source fidelity, extracts latent evidence structures, conducts rigorous pre-mortem analysis, and provides well-justified probabilities. The only limitation preventing perfect scores across all dimensions is the absence of fully quantified Bayesian updating with explicit likelihood ratios.

2 Task 3 background synthesis RAG pipeline and hallucination analysis

Table S4 Task 3 background synthesis RAG pipeline

Pipeline stage	Description
Resolution criteria inference	Resolution criteria for 550 items (questions) were generated using gemini-3-flash and subsequently manually verified. This model was selected for its high inference speed and precise instruction-following capabilities. Ultimately, this hybrid human-in-the-loop paradigm accelerated the criteria formulation while supporting the logical rigor and unambiguous clarity of the final criteria. The prompt template is as follows: **ROLE:** You are an expert forecasting adjudicator. **TASK:** You have a forecasting question that lacks its original adjudication rules. You need to define strict, unambiguous criteria for how this question should resolve to "YES", and output the result as a concise paragraph titled "Resolution Criteria". **INPUT:** {Question} **INSTRUCTIONS:** 1. Define what constitutes a "positive event" (e.g., official press release, product listing, regulatory filing). 2. Define what does NOT count (e.g., rumors, supply chain leaks, analyst predictions). 3. Consider the specific wording of the question to handle edge cases.
Search query generation	Based on the forecasting question and resolution criteria, we first generated six keyword-based Google search queries via a direct strategy. Second, we employed a reasoning strategy that decomposes the main question into sub-questions (e.g., competitor status, technical hurdles, official timelines) to generate six targeted queries. Finally, these outputs were merged and deduplicated to form a unique query set of up to 12 queries for subsequent retrieval. This dual-pathway approach was designed to improve coverage of the evidence space, capturing both explicit factual mentions and latent contextual signals.
Initial Snippet Retrieval	Based on the generated queries, we utilized the Serper API to retrieve Google News search snippets. We constrained the search timeframe to retrieve only snippets published prior to the retrieval date (25% lifecycle cutoff). To prevent token overflow, we retained up to the top 40 snippets per item based on the search engine's default relevance ranking, yielding approximately 19,500 snippets for subsequent LLM-based filtering. By acting as verifiable factual anchors, these curated snippets offer the critical advantages of enhanced reproducibility, reduced information noise, and high engineering scalability, thereby grounding the LLMs' reasoning and mitigating generative hallucination.
Snippet scoring & filtering	Using DeepSeek-V3.2, we performed snippet-level scoring and filtering on the initial pool of up to 40 retrieved snippets per item, retaining up to the top 15 most relevant ones, yielding approximately 5,200 filtered snippets. By filtering out low-relevance noise, this distillation maximizes the signal-to-noise ratio and mitigates LLM attention degradation, grounding the final synthesis in high-density evidence. The prompt template is as follows: **ROLE:** You are an evaluator checking the relevance of retrieved snippets for a forecasting question. **TASK:** Score the RELEVANCE of the Snippets to the Question on a scale of 1-5, conditioned on the provided Resolution Criteria and Retrieval Date. Strictly assign a score of 1 to any snippet whose source is published after the Retrieval Date. Return a JSON array of up to the top 15 snippet IDs that score greater than 1, sorted in descending order by their relevance scores. **INPUT:** {Question}; {Resolution Criteria}; {Snippets}; {Retrieval Date} **RUBRIC:** 1: Irrelevant, off-topic, or the snippet's source is published after the Retrieval Date. 2: Vague or peripheral relevance; mentions relevant keywords but misses the core prediction target. 3: Relevant entities or contexts are explicitly mentioned, but the evidentiary value for the event's status is weak or ambiguous.

Continued on next page

Table S4 (Continued)

Pipeline stage	Description
	4: High evidentiary value. The snippet provides clear and direct information regarding key entities, dates, or event status. 5: Perfect evidentiary value. The snippet provides definitive, explicit, and highly dense evidence directly answering the prediction target or event status within the snippet.
Background synthesis	We first concatenated up to 15 filtered snippets along with their respective source titles, and then used DeepSeek-V3.2 to integrate them into a unified background briefing for each of the 550 items, with approximately 3,000 snippets ultimately integrated. This synthesis distills fragmented evidence into a cohesive, high-density narrative, standardizing the evidentiary foundation for each item while mitigating downstream attention degradation. To bound generative hallucination during this synthesis, we applied the following prompt template: **ROLE:** You are an elite Technical Intelligence Analyst. **TASK:** Synthesize a high-density background briefing based on the provided Question and Concatenated Snippets (each accompanied by its respective source title). **INPUT:** {Question}; {Concatenated Snippets} **INSTRUCTIONS:** 1. Create a unified, comprehensive background briefing. Length must be proportionate to the provided evidence. Do not pad or invent text to meet a length quota. 2. For each snippet utilized, extract factual data (e.g., official dates, metrics). All specific claims MUST be derived from the provided input. 3. If the target entity is missing from the provided snippets, DO NOT fabricate its status. You must explicitly declare: "No external evidence found." 4. Output as a plain text narrative.
Background quality assessment	To assess the quality of the synthesized background briefings, we conducted a rigorous assessment of their relevance to the corresponding items. Forecasting items whose background briefings failed our predefined standards (i.e., scored below 4) were excluded, retaining 421 of the initial 550 items for subsequent manual review. The prompt template is as follows: **ROLE:** You are an elite Information Quality Auditor. **TASK:** Score the RELEVANCE of the Background Briefing to the Question on a scale of 1-5, conditioned on the provided Resolution Criteria. Return only the raw integer score. **INPUT:** {Question}; {Resolution Criteria}; {Background Briefing} **RUBRIC:** 1: The background briefing is completely off-topic, addresses unrelated entities, or ignores the question's core subject and timeline. 2: The background briefing mentions the correct general domain or one core entity but fails to connect it to the specific question. The timeline is incorrect or ignored, and the logical link to the forecast is weak. 3: The background briefing correctly identifies the core entities and general timeline but provides context that is only broadly applicable, lacking focus on the specific mechanism or condition central to the question. 4: The background briefing directly addresses the core entities and the correct timeline. It provides focused context on the specific technical, regulatory, or market factors that the decision will hinge on, with only minor omissions or one unclear link. 5: The background briefing is precisely tailored to the question. It explicitly sets up the entities, the decision point, and the key conflicting evidence or criteria that make the forecast non-obvious, creating a clear, logical foundation for the subsequent prediction.

Table S5 A representative example of Task 3 background hallucination analysis

Item	Description
Question	Will Google's next video-making AI be called Veo 3?
Background	Based on the provided source data, the officially announced and launched product is definitively named "Veo 3". The evidence is explicit and multi-sourced. Google's official Generative AI on Vertex AI documentation, dated February 27, 2024, lists "Veo 3" and "Veo 3.1" as distinct model versions. An official Google AI Developers Forum announcement on August 10, 2025, is titled "Veo 3 is now available in the Gemini API." API documentation from December 11, 2024, references "Veo 3 models." Product pages for Google AI plans, dated April 23, 2024, mention user access to "Veo 3.1" and "Veo 3." The model is consistently referenced as "Veo 3" in technical guides and launch articles from December 2024 and January 2025. The "Focus Gap" is resolved: the canonical name is "Veo 3," with "Veo 3.1" identified as a subsequent iteration featuring enhanced audio generation capabilities as of May 14, 2024. The base model "Veo" was initially announced on May 14, 2024. The State of the Art, per the latest data, is Veo 3.1, which includes native audio generation (sound effects, ambient noise, dialogue) alongside best-in-class video quality. The lineage is clear: Veo (announced May 2024) → Veo 2 (unveiled Dec 2024) → Veo 3 (launched by Aug 2025) → Veo 3.1 (feature update noted by May 2024, with references through 2025). Therefore, Google's next major video-making AI after the data cutoff would logically be a successor to Veo 3.1, but the sources contain no announcement or data regarding a model named "Veo 3" as a future product; it is documented as a current, launched model.
Resolution criteria	This question will resolve to YES if and only if: 1. Official Announcement: Google or its parent company Alphabet officially announces, launches, or publicly releases a new video-making AI model or product that is explicitly and directly named "Veo 3". This must occur via an official, verifiable channel such as: – A press release or blog post on an official Google or Alphabet domain (e.g., blog.google, ai.google). – An official announcement at a Google-hosted event (e.g., Google I/O, Google Cloud Next). – Official documentation, a product page, or a launch on an official Google platform (e.g., the Google AI Studio, Google Labs). – A filing (e.g., trademark application) from Google/Alphabet that explicitly lists "Veo 3" as the name for a video-generation AI product. 2. Product Association: The announced "Veo 3" must be clearly identified as a video-generation AI model or product. It must be a successor or new version in a lineage that includes "Veo" (e.g., following a product like "Veo" or "Veo 2"). What Does NOT Count (Resolves NO): – Rumors, leaks, insider reports, or analyst predictions. – Mentions in unofficial social media posts, forums, or news articles not originating from Google. – Trademark applications or domain registrations by entities other than Google/Alphabet. – References to internal codenames, project names, or developmental titles that are not the official public product name. – Announcements of a product with a similar but different name (e.g., "Veo-3", "Veo III", "VideoVeo 3", "Veo 3.0" unless the official branding presented is unequivocally "Veo 3"). – A product that is merely an update or feature addition to an existing "Veo" model without being branded and announced as a distinct new model named "Veo 3".
Retrieval date	2025-02-06T14:49:13Z
Resolution	YES (1.0)

Continued on next page

Table S5 (Continued)

Item	Description
Hallucination analysis	This case exemplifies fabrication accompanied by internal inconsistency. The model invents specific but fictitious details, such as non-existent official Google documentation dates and forum announcements. Furthermore, the model constructs a catastrophic temporal contradiction within its own narrative, asserting that "Veo 3" launched in August 2025 while its upgraded version, "Veo 3.1", was documented over a year earlier in May 2024. This logical breakdown underscores the necessity of post-synthesis quality control and filtering mechanisms in the RAG pipeline to eliminate self-generated factual and temporal inconsistencies.

3 Human–LLM agreement evaluation

Table S6 Task-specific scoring rubrics

Score	Task 2 Rubric	Task 3 Rubric
1	Critical Failure: Contradicts NASA definitions* (e.g., wrong hardware fidelity for the TRL) or lacks evidence.	Deficient: The background fails to provide a useful foundation. It is both off-topic and devoid of concrete information.
2	Insufficient: Vague. Mentions TRL but fails to define the specific "Relevant Environment" or "Hardware Fidelity" (e.g., Breadboard vs. Brassboard).	Weak: The background is only minimally connected to the question and provides little to no substantive detail. The link to the forecast is tenuous.
3	Borderline: Uses non-standard terminology or misses specific completion criteria like "scaling requirements" or "life-limiting factors".	Functional: The background is broadly applicable but lacks focus or specificity. It provides a basic, correct context but does not create a sharp foundation for the specific forecast.
4	Compliant: Clearly identifies correct Hardware Fidelity, Test Environment, and Integration Level. Evidence supports the TRL.	Strong: The background is well-targeted and informative. It directly addresses the question with focused context and concrete details, creating a solid foundation for analysis.
5	Exemplary: Perfectly aligned with NASA definitions. Explicitly details Hardware (e.g., "Engineering Unit"), Environment (e.g., "Thermal-Vac"), and verifies models.	Exemplary: The background is precisely tailored and richly detailed. It perfectly frames the forecast's dilemma by integrating key, specific evidence within the exact context of the question.

*NASA TRL definitions from the TRA guide accompanied the scoring.

Table S7 Representative Task 2 disagreement case

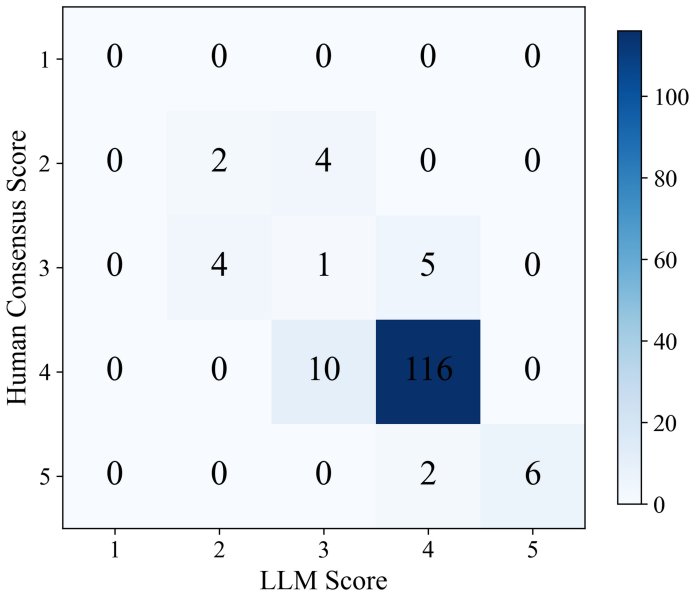
Item	Description
Technical description	A proof-of-concept HASEL soft actuator with embedded self-diagnosis capabilities was developed. This breadboard-level system utilized temporal impedance measurements of the actuator's driving power signal to monitor its operational state. Laboratory experiments were conducted to generate datasets, demonstrating the system's ability to infer information about the actuator's environment—such as detecting physical blockages and external perturbations—and to identify statistical indicators of impending failure. By intelligently altering the driving signal based on this monitoring, the project successfully demonstrated the functional principle of fault prediction and avoidance, which was shown to extend the actuator's operational lifetime.
TRL	4
Human consensus score and rationale	**Score:** 3 **Rationale:** Although the description identifies breadboard-level hardware and laboratory testing, it remains borderline because it does not explicitly provide specific completion criteria such as scaling requirements or life-limiting factors. This matches the rubric definition of Score 3.
LLM score and rationale	**Score:** 4 **Rationale:** The description is compliant because it clearly states hardware fidelity (breadboard-level system), test environment (laboratory experiments), and functional evidence of validation, which aligns with the rubric definition of Score 4.

Table S8 Representative Task 3 disagreement case

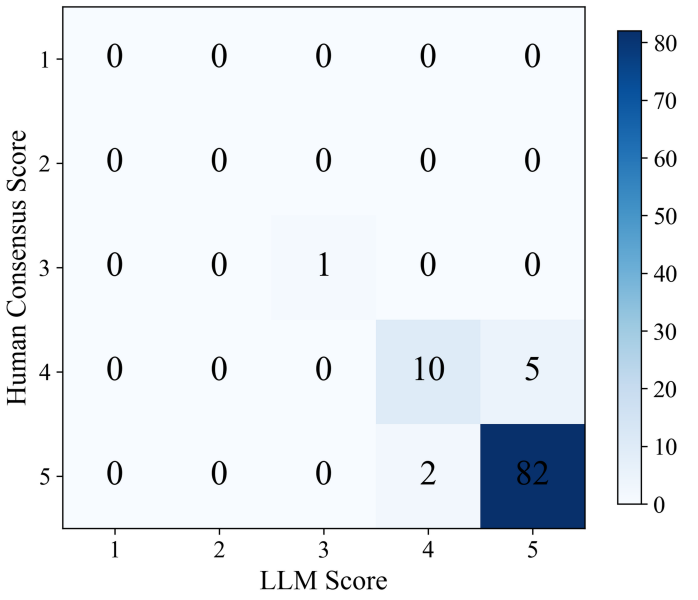
Item	Description
Question	Did Gemini 3 and GPT-5 both release within a 2-week span between August 1-14, 2025?
Background	Based on the provided source data, the release timelines for OpenAI's GPT-5 and Google's Gemini 3 are partially documented but with a critical gap. Multiple credible reports from July 24, 2025, indicate OpenAI planned a GPT-5 launch for August 2025, with one source specifying a launch date of August 7, 2025. This is corroborated by a reference to a livestream titled "Introducing GPT-5" dated August 7, 2025. For Google's Gemini, the data confirms a specific update: "Gemini 3 Flash is now live globally," with an associated date of August 21, 2025. This creates a direct conflict for the posed question. The evidence strongly suggests GPT-5 launched on or around August 7, while Gemini 3 Flash's global availability is tied to August 21. These dates are 14 days apart, placing them within a 2-week span. However, the core entity "Gemini 3" (distinct from "Gemini 3 Flash") is not explicitly mentioned in the provided sources. The State of the Art (SOTA) for Google's AI model family, as per this data, is the Gemini 3 Flash variant, released globally on August 21, 2025. The base rate for major AI model releases from leading labs like Google and OpenAI in the 2024-2025 period shows a pattern of incremental and flash-model-variant releases following flagship announcements, often within quarters of each other.
Resolution criteria	This question resolves to YES if and only if both conditions are met: (1) an official and verifiable Gemini 3 launch announcement by Google (or authorized channel); (2) an official and verifiable GPT-5 launch announcement by OpenAI (or authorized channel). For both releases, availability must fall within August 1–14, 2025 (inclusive), the first qualifying public event determines release date, and the release-date gap must be 14 days or fewer. Non-counting cases include rumors, leaks, previews without public availability, interim/non-successor names (e.g., GPT-4.5, Gemini 2.5), and private-only releases.
Resolution	NO (0.0)
Human consensus score and rationale	**Score:** 4 **Rationale:** The background is mostly relevant to the exact forecasting target and provides high-density evidence, including concrete dates, named entities, and source-linked facts. It gives a solid analytical foundation, but one minor omission remains in entity precision (Gemini 3 vs Gemini 3 Flash), so it fits Strong rather than Exemplary.
LLM score and rationale	**Score:** 5 **Rationale:** The background reaches perfect relevance by explicitly framing the decision point and the key conflicting evidence (Aug 7 GPT-5 signals vs Aug 21 Gemini 3 Flash availability, plus naming ambiguity). It also shows very high density with exact dates, comparative timing logic, and attributed sources, creating a complete and non-obvious forecast frame.

Table S9 Score distributions for Tasks 2 and 3

Score	Task 2 score distribution (%)		Task 3 score distribution (%)	
	Human consensus	**LLM**	**Human consensus**	**LLM**
1	0.0	0.0	0.0	0.0
2	4.0	4.0	0.0	0.0
3	6.7	10.0	1.0	1.0
4	84.0	82.0	15.0	12.0
5	5.3	4.0	84.0	87.0



Task 2 ($n = 150$)



Task 3 ($n = 100$)

Fig. S1 Human consensus vs. LLM confusion matrices for Tasks 2 and 3

4 Examples of tasks in TMBench

Task 1

SMC (China)

问题：为确保谈判成果准确转化为具有法律效力的文件，合同起草工作应与____同步进行。
- A. 谈判阶段
- B. 合同签订阶段
- C. 履约阶段
- D. 项目启动阶段

答案 : 谈判阶段

SMC (U.S.)

Question: In business planning, the SWOT analysis acronym stands for Strengths, Weaknesses, Opportunities, and ____.
- A. Threats
- B. Trends
- C. Targets
- D. Tactics

Answer : Threats

MMC (China)

问题：通过对科技人员实施科技成果产权激励并赋予其深度参与权，可以产生的主要积极效果包括：____。
- A. 增强科技人员对成果转化的主体意识，提高其参与转化的动力
- B. 从源头上提高科技成果的质量供给，使研发更贴近市场需求
- C. 减少科技人员的研发投入，降低创新成本
- D. 完全消除高校院所的“唯论文、唯职称、唯学历、唯奖项”倾向
- E. 使科技人员无需参与成果转化决策，专注于研发工作

答案 : A, B

MMC (U.S.)

Question: Central governments and local municipalities have implemented responses to address changes in innovation and entrepreneurship, including ____.
- A. a myriad of tax incentives
- B. regulatory and legal modifications
- C. the creation of local incubation and support activities
- D. increased tariffs on imports
- E. reduction in public education funding

Answer : A, B, C

JU (China)

问题：在获取谈判主体信息阶段，如果发现意向合作方存在重大信用瑕疵，建议没有必要与之进行谈判。

答案 : 正确

JU (U.S.)

Question: A license agreement drafted by a legal team to limit ambiguous terms and address real-world situations with specificity primarily benefits by clearly defining the scope of a subsequent royalty audit.

Answer : True

Task 2

Question: Evaluate the Technology Readiness Level (TRL) based on the following technical description.
Technical Description : "An ammonia Loop Heat Pipe (LHP) prototype incorporating a passive Thermal Control Valve (TCV) was developed. This high-fidelity prototype was tested in a thermal vacuum chamber, a relevant environment. Testing successfully demonstrated a heat rejection conductance turn-down capability in excess of 50:1 and the system's ability to passively shut down under cold sink and low power conditions. The prototype also withstood nine freeze and thaw cycles without condenser deformation and was capable of restarting after exposure to a -120°C sink temperature."

Answer : TRL= 6

Task 3

Question: Will GPT-5.2 exceed Claude Opus 4.5 on SWE-bench? The known information is as follows:
Background: Based on the provided data, the current state-of-the-art (SOTA) on the SWE-bench evaluation is held by Claude Opus 4.5, not GPT-5.2. While the query addresses a future performance outcome, the latest benchmark data indicates that the hypothesized scenario has already been surpassed by recent developments. A comprehensive overview of the situation is as follows: SWE-bench is a leading benchmark for evaluating large language models (LLMs) on real-world software engineering tasks, specifically resolving issues in open-source projects. Performance is measured by a pass rate, representing the percentage of issues a model correctly resolves. The landscape is highly competitive; according to the latest official leaderboard from November 18, 2025, Claude Opus 4.5 achieved a pass rate of 74.4% on the 'SWE-bench+' variant. This results in it leading its closest competitors: Gemini 3 Pro (74.2%) and GPT-5.2 (71.8%). Although earlier data from May 6, 2025, showed GPT-5.2 in the lead (71.80% vs. 70.60% for Claude 4.5 Sonnet), the updated metrics confirm that Claude Opus 4.5 has since improved and taken the top position.
Resolution criteria:
- Event Requirement: OpenAI, or a directly authorized representative, issues an official public announcement or publication stating that a model named "GPT-5.2" has been evaluated on the SWE-bench benchmark.
- Performance Requirement: The announced results demonstrate that GPT-5.2's performance outperforms Claude Opus 4.5 on the identical version of the SWE-bench benchmark.
- Source & Benchmark Clarity: The comparison must be based on the canonical, public SWE-bench. The performance data for Claude Opus 4.5 must originate from an official evaluation by Anthropic or a reputable, neutral third-party entity.

Retrieval date: 2025-12-08T18:36:15Z

Answer : NO

Fig. S2 Examples of different tasks in TMBench

5 Prompt templates for Task 1

SMC

- You are an expert in *TargetRegion*[*] technology transfer, specifically acting as a seasoned technology manager. You are tasked with answering single-select multiple-choice questions related to this domain.
- Before providing the final answer, you must provide a comprehensive analysis of the question and the available options, explaining the logic behind your choice.
- You MUST format your entire response strictly as a valid JSON object, enclosed within <task_output>tags.
- The JSON object must contain exactly two keys:
  - "analysis": A string containing your step-by-step reasoning.
  - "answer": A single string representing your final choice (e.g., "Option A").
- Example format:
  <task_output>
  {
  "analysis": "Here is the detailed reasoning explaining why the specific option is correct and why the others are incorrect...",
  "answer": "Option A"
  }
  </task_output>

[*] **Note:** The placeholder *TargetRegion* is instantiated as either United States or China, depending on the specific dataset from which the question originates.

MMC

- You are an expert in *TargetRegion*[*] technology transfer, specifically acting as a seasoned technology manager. You are tasked with answering multi-select multiple-choice questions related to this domain.
- Before providing the final answer, you must provide a comprehensive analysis of the question and the available options, explaining the logic behind your choice.
- You MUST format your entire response strictly as a valid JSON object, enclosed within <task_output>tags.
- The JSON object must contain exactly two keys:
  - "analysis": A string containing your step-by-step reasoning.
  - "answer": A JSON array of selected options, e.g. ["Option A", "Option C"].
- Example format:
  <task_output>
  {
  "analysis": "Here is the detailed reasoning explaining why the specific options are correct and why the others are incorrect...",

"answer": ["Option A", "Option C"]
}
</task_output>

* **Note:** The placeholder *TargetRegion* is instantiated as either United States or China, depending on the specific dataset from which the question originates.

JU

- You are an expert in *TargetRegion** technology transfer, specifically acting as a seasoned technology manager. You are tasked with answering true/false judgment questions related to this domain.
- Before providing the final answer, you must provide a comprehensive analysis of the statement, explaining the logic behind your judgment.
- You MUST format your entire response strictly as a valid JSON object, enclosed within <task_output>tags.
- The JSON object must contain exactly two keys:
  - "analysis": A string containing your step-by-step reasoning.
  - "answer": A single string representing your final judgment (i.e., "True" or "False").
- Example format:
  <task_output>
  {
  "analysis": "Here is the detailed reasoning explaining why the statement is true or false...",
  "answer": "True"
  }
  </task_output>

* **Note:** The placeholder *TargetRegion* is instantiated as either United States or China, depending on the specific dataset from which the question originates.

6 Prompt templates for Task 2

PA1

- SYSTEM ROLE: You are a Strict Technology Manager responsible for the rigorous assessment of technology portfolios. Your primary goal is to determine the accurate Technology Readiness Level (TRL) of a project to ensure no premature advancement risks are taken.
- OBJECTIVE: Evaluate the provided Technical Description exclusively based on the Official NASA TRL Definitions below. You must act as a gatekeeper: do not award a TRL level unless all criteria for that level (and previous levels) are explicitly met in the text.
- OFFICIAL NASA TRL DEFINITIONS: ...
- REQUIRED TERMINOLOGY CONSTRAINTS:
  - Breadboard (TRL 4/5): Low fidelity. Focuses on function only. Not form or fit.

  - Prototype (TRL 6+): High fidelity. Near-identical to final unit in form, fit, and function.
  - Relevant Environment: Simulates key stresses (e.g., vacuum, thermal).
  - Operational Environment: The actual real-world mission condition.
- EVALUATION GUIDELINES:
  - Evidence Only: Do not assume implied tests. If the text says "tested in lab," do not assume "relevant environment" unless explicitly described.
  - Hierarchy Rule: To achieve Level N, the technology must satisfy all requirements of Level N AND imply the completion of Levels 1 through $N-1$.
  - Conservative Judgment: If the description is ambiguous between two levels (e.g., between 4 and 5), assign the lower level to minimize risk.
- TASK INSTRUCTIONS:
  - Read the Technical Description.
  - Match the description against the terminology and definitions.
  - Determine the highest verified TRL.
  - Provide a concise analysis of the evidence.
  - Output the reasoning followed by the final TRL level in the specified format.
- FORMAT CONSTRAINT:
  - Output your final answer in this exact format:
  - Reasoning: [Provide a concise analysis of the evidence]
  - Final Answer: TRL [Number]
- TECHNICAL DESCRIPTION: ...

PA2

- SYSTEM ROLE: You are a Strict Technology Manager responsible for the rigorous assessment of technology portfolios. Your primary goal is to determine the accurate Technology Readiness Level (TRL) of a project to ensure no premature advancement risks are taken.
- OBJECTIVE: Evaluate the provided Technical Description exclusively based on the Official NASA TRL Definitions below. You must act as a gatekeeper: do not award a TRL level unless all criteria for that level (and previous levels) are explicitly met in the text.
- OFFICIAL NASA TRL DEFINITIONS: ...
- REQUIRED TERMINOLOGY CONSTRAINTS:
  - Breadboard (TRL 4/5): Low fidelity. Focuses on function only. Not form or fit.
  - Prototype (TRL 6+): High fidelity. Near-identical to final unit in form, fit, and function.
  - Relevant Environment: Simulates key stresses (e.g., vacuum, thermal).
  - Operational Environment: The actual real-world mission condition.
- EVALUATION GUIDELINES:

- Evidence Only: Do not assume implied tests. If the text says "tested in lab," do not assume "relevant environment" unless explicitly described.
- Hierarchy Rule: To achieve Level N, the technology must satisfy all requirements of Level N AND imply the completion of Levels 1 through $N-1$.
- Conservative Judgment: If the description is ambiguous between two levels (e.g., between 4 and 5), assign the lower level to minimize risk.

- **EXAMPLES:**
  - Example 1:
    - Technical Description: "A breadboard Er:YAG laser transmitter was built for direct generation of a 1.65 μm wavelength. This low-fidelity component was tested in a laboratory environment, where it demonstrated an output of greater than 500 μJ per pulse at a 10 kHz pulse repetition rate and a tunable operating wavelength at a methane absorption line. The technical performance results from these breadboard experiments established that the laser approach could meet NASA's methane lidar requirements."
    - Final Answer: TRL 4
  - Example 2:
    - Technical Description: "A dual-element PPKTP waveguide module was designed, fabricated, and packaged. This brassboard component was delivered and integrated with a partner's space-to-ground entanglement distribution test bed. In this relevant test environment, the module demonstrated a polarization entanglement fidelity of 0.95."
    - Final Answer: TRL 5
  - Example 3:
    - Technical Description: "A prototype thermal corer, an improved PVEx drill with integrated minichannels, was developed. It was characterized through in-house performance testing and was successfully demonstrated in a simulated lunar environment inside a thermal vacuum chamber. The testing in relevant conditions showed an ice recovery rate greater than 60% with the regolith environment maintained at -30°C and 1E-3 Torr."
    - Final Answer: TRL 6

- **TASK INSTRUCTIONS:**
  - Read the Technical Description.
  - Refer to the Examples above to understand the mapping between description features and TRL levels.
  - Determine the highest verified TRL based on the Definitions and Assessment Rules.
  - Provide a concise analysis of the evidence.
  - Output the reasoning followed by the final TRL level in the specified format.

- FORMAT CONSTRAINT:
  - Output your final answer in this exact format:

  - Reasoning: [Provide a concise analysis of the evidence]
  - Final Answer: TRL [Number]
- TECHNICAL DESCRIPTION: ...

PA3

- SYSTEM ROLE: You are a Strict Technology Manager responsible for the rigorous assessment of technology portfolios. Your primary goal is to determine the accurate Technology Readiness Level (TRL) of a project to ensure no premature advancement risks are taken.
- OBJECTIVE: Evaluate the provided Technical Description exclusively based on the Official NASA TRL Definitions below. You must act as a gatekeeper: do not award a TRL level unless all criteria for that level (and previous levels) are explicitly met in the text.
- OFFICIAL NASA TRL DEFINITIONS: ...
- REQUIRED TERMINOLOGY CONSTRAINTS:
  - Breadboard (TRL 4/5): Low fidelity. Focuses on function only. Not form or fit.
  - Prototype (TRL 6+): High fidelity. Near-identical to final unit in form, fit, and function.
  - Relevant Environment: Simulates key stresses (e.g., vacuum, thermal).
  - Operational Environment: The actual real-world mission condition.
- EVALUATION GUIDELINES:
  - Evidence Only: Do not assume implied tests. If the text says "tested in lab," do not assume "relevant environment" unless explicitly described.
  - Hierarchy Rule: To achieve Level N, the technology must satisfy all requirements of Level N AND imply the completion of Levels 1 through $N-1$.
  - Conservative Judgment: If the description is ambiguous between two levels (e.g., between 4 and 5), assign the lower level to minimize risk.
- **EXAMPLES:**
  - Example 1:
    - Technical Description: "A breadboard Er:YAG laser transmitter was built for direct generation of a 1.65 μm wavelength. This low-fidelity component was tested in a laboratory environment, where it demonstrated an output of greater than 500 μJ per pulse at a 10 kHz pulse repetition rate and a tunable operating wavelength at a methane absorption line. The technical performance results from these breadboard experiments established that the laser approach could meet NASA's methane lidar requirements."
    - Reasoning: The hardware is explicitly identified as a low-fidelity "breadboard" tested in a benign "laboratory environment", which aligns perfectly with TRL 4 criteria.
    - Final Answer: TRL 4
  - Example 2:

    - Technical Description: "A dual-element PPKTP waveguide module was designed, fabricated, and packaged. This brassboard component was delivered and integrated with a partner's space-to-ground entanglement distribution test bed. In this relevant test environment, the module demonstrated a polarization entanglement fidelity of 0.95."
    - Reasoning: The description identifies a brassboard component, corresponding to medium-fidelity hardware under the stated definitions, and reports its integration with a partner's space-to-ground entanglement distribution test bed and demonstrated polarization entanglement fidelity of 0.95 in an explicitly identified relevant test environment, which supports TRL 5, while the absence of a high-fidelity prototype required for TRL 6 precludes TRL 6 or higher under the Hierarchy Rule.
    - Final Answer: TRL 5
  - Example 3:
    - Technical Description: "A prototype thermal corer, an improved PVEx drill with integrated minichannels, was developed. It was characterized through in-house performance testing and was successfully demonstrated in a simulated lunar environment inside a thermal vacuum chamber. The testing in relevant conditions showed an ice recovery rate greater than 60% with the regolith environment maintained at $-30°$C and 1E-3 Torr."
    - Reasoning: The high-fidelity "prototype" system demonstrated in a "simulated lunar environment" (Relevant Environment) satisfies the definition of TRL 6.
    - Final Answer: TRL 6

- **TASK INSTRUCTIONS:**
  - Read the Technical Description.
  - Refer to the Examples above to understand the mapping between description features and TRL levels.
  - Determine the highest verified TRL based on the Definitions and Assessment Rules.
  - Provide a single sentence of reasoning explaining why it matches a specific TRL.
  - Output the reasoning followed by the final TRL level in the specified format.
- FORMAT CONSTRAINT:
  - Output your final answer in this exact format:
  - Reasoning: [Provide a concise analysis of the evidence]
  - Final Answer: TRL [Number]
- TECHNICAL DESCRIPTION: ...

7 Prompt templates for Task 3

PF1

- QUESTION: ...
- BACKGROUND: ...

- RESOLUTION CRITERIA: ...
- RETRIEVAL DATE: ...
- INSTRUCTION: Output your Reasoning Process, Probability and Confidence Assessment.

PF2

- QUESTION: ...
- BACKGROUND: ...
- RESOLUTION CRITERIA: ...
- RETRIEVAL DATE: ...
- **INSTRUCTIONS:**
  - Given the above question, rephrase and expand it to help you answer more effectively. Maintain all information in the original question.
  - Using your knowledge of the world and topic, as well as the information provided, provide a few reasons why the answer might be no. Rate the strength of each reason.
  - Using your knowledge of the world and topic, as well as the information provided, provide a few reasons why the answer might be yes. Rate the strength of each reason.
  - Aggregate your considerations. Think like a superforecaster (e.g., Nate Silver).
  - Output an initial probability (prediction) given steps 1-4.
  - Evaluate whether your calculated probability is excessively confident or not confident enough. Also, consider anything else that might affect the forecast that you did not previously consider (e.g., base rate of the event). Determine your confidence level based on the provided categories.
  - Output your Reasoning Process, Probability and Confidence Assessment.

PF3

- QUESTION: ...
- BACKGROUND: ...
- RESOLUTION CRITERIA: ...
- RETRIEVAL DATE: ...
- **INSTRUCTIONS:**
  - Given the above question, rephrase and expand it to help you answer more effectively. Maintain all information in the original question.
  - Analyze the provided Background text to provide a few reasons why the answer might be no.
    - Constraint: You may use your training for reasoning logic, but all evidence, specific data, and entity attributes must come exclusively from the provided Background.

    - Negative Constraint: Do not use external knowledge to import facts, reputations, or psychological traits (e.g., "impulsive," "firefighting") unless explicitly described in the text.
    - Uncertainty: Explicitly label ambiguous information or rumors found in the text as such; do not attempt to resolve them with external knowledge.
    - Rate the strength of each reason based on the evidence provided.
  - Analyze the provided Background text to provide a few reasons why the answer might be yes.
    - Constraint: Adhere to the same strict evidence constraints as mentioned above. Do not hallucinate external facts or industry averages.
    - Entity Verification: Verify all entity-claim links against the text. Do not default to the most famous entity associated with a topic (e.g., do not attribute a report to a major org if the text specifies a different one).
    - Rate the strength of each reason based on the evidence provided.
  - Aggregate your considerations. Think like a superforecaster (e.g., Nate Silver).
    - Source Integrity: Do not capitalize phrases to look like document titles (e.g., "The 2025 Reliability Report") unless they appear verbatim in the text.
    - Weigh the evidence found in the Background to form a balanced view.
  - Output an initial probability (prediction) given the preceding analysis.
  - Evaluate your confidence.
    - Base Rates: derive base rates only from historical data explicitly present in the Background. If no relevant historical data is present, state that no base rate is available. Do not invent "standard industry practices" or "comparable events."
    - Consider if the evidence in the text is sufficient or conflicting. Determine your confidence level based on the provided categories.
  - Output your Reasoning Process, Probability and Confidence Assessment.

8 Distribution of the 370-item subset

Table S10 Task distribution in the 370-item subset and characteristics of its Domain Knowledge QA items

Dimension	Category	n	%
Overall tasks ($N = 370$)			
Task category	Domain Knowledge QA	250	67.6
	Technology Readiness Assessment	75	20.3
	Technology Market Forecasting	45	12.2
Domain Knowledge QA items ($N = 250$)			
Domain	FK	94	37.6
	PR	79	31.6
	PP	77	30.8
Context	China	182	72.8
	U.S.	68	27.2
Question type	SMC	120	48.0
	MMC	77	30.8
	JU	53	21.2

9 Experimental results from TMBench ($T = 0.0$)

Table S11 LLM performance on three tasks

Model	Domain knowledge QA			Technology readiness assessment			Technology market forecasting			
	Accuracy	Score	F1	Accuracy	RS	MAE	BS	RS	ECE	AR
gpt-5-mini	0.778	1.839	0.913	0.747	3.655	0.353	0.239	3.893	0.167	1.350
gemini-3-flash	0.786	1.846	0.919	0.771	3.928	0.284	0.248	3.995	0.186	1.388
grok-4.1-fast	0.773	1.835	0.910	0.780	4.153	0.291	0.251	4.732	0.197	1.373
claude-h-4.5	0.761	1.828	0.905	0.789	3.600	0.251	0.247	3.857	0.189	1.306
mistral-small-3.1	0.759	1.832	0.905	0.782	3.917	0.268	0.246	3.996	0.166	1.340
llama-4-mav	0.758	1.815	0.909	0.763	3.845	0.298	0.242	4.043	0.153	1.372
DeepSeek-V3.2	0.816	1.885	0.934	0.784	3.641	0.267	0.229	4.378	0.141	1.422
qwen3-next-80b	0.758	1.826	0.906	0.762	3.687	0.294	0.251	4.861	0.197	1.452
kimi-k2-0905	0.765	1.826	0.911	0.775	3.859	0.278	0.236	3.485	0.160	1.403
Mean	**0.773**	**1.837**	**0.912**	**0.773**	**3.809**	**0.287**	**0.243**	**4.138**	**0.173**	**1.378**

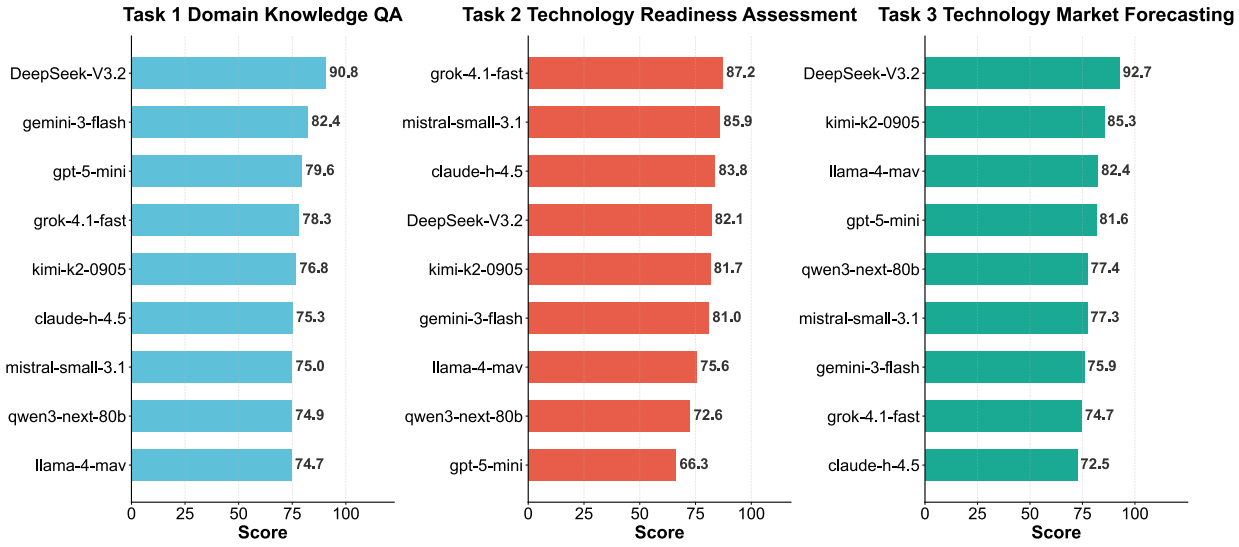


Fig. S3 LLM rankings across three tasks

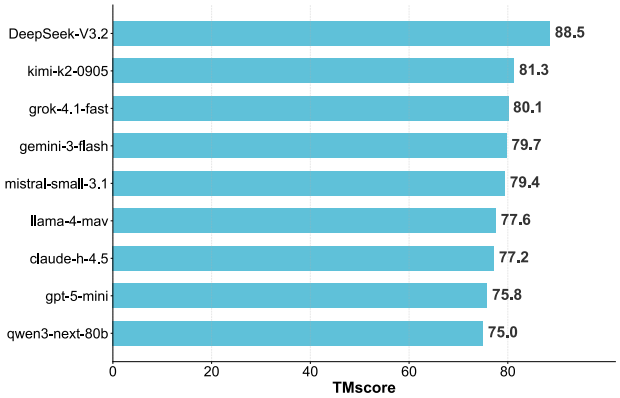


Fig. S4 LLM rankings based on TMscore

Table S12 Performance of LLMs across domains

Model	Accuracy			F1			HR		
	PR	FK	PP	PR	FK	PP	PR	FK	PP
gpt-5-mini	0.824	0.745	0.772	0.921	0.905	0.910	0.070	0.149	0.122
gemini-3-flash	0.828	0.749	0.790	0.921	0.912	0.921	0.075	0.145	0.113
grok-4.1-fast	0.814	0.745	0.767	0.911	0.905	0.912	0.089	0.152	0.118
claude-h-4.5	0.807	0.732	0.751	0.906	0.903	0.905	0.091	0.151	0.121
mistral-small-3.1	0.796	0.738	0.748	0.903	0.902	0.904	0.095	0.151	0.113
llama-4-mav	0.798	0.723	0.760	0.913	0.905	0.906	0.084	0.146	0.128
DeepSeek-V3.2	0.894	0.759	0.807	0.952	0.926	0.931	0.041	0.103	0.082
qwen3-next-80b	0.810	0.722	0.748	0.914	0.900	0.903	0.080	0.149	0.125
kimi-k2-0905	0.807	0.738	0.757	0.915	0.907	0.909	0.077	0.147	0.117
Mean	**0.820**	**0.739**	**0.767**	**0.917**	**0.907**	**0.911**	**0.078**	**0.144**	**0.115**
SD	**0.030**	**0.012**	**0.020**	**0.014**	**0.008**	**0.009**	**0.016**	**0.015**	**0.014**

Table S13 Performance of LLMs across contexts

Model	Accuracy		F1		HR	
	China	U.S.	China	U.S.	China	U.S.
gpt-5-mini	0.751	0.848	0.912	0.914	0.130	0.078
gemini-3-flash	0.749	0.883	0.913	0.934	0.131	0.067
grok-4.1-fast	0.755	0.823	0.914	0.900	0.130	0.099
claude-h-4.5	0.751	0.788	0.913	0.887	0.128	0.109
mistral-small-3.1	0.751	0.781	0.912	0.884	0.128	0.105
llama-4-mav	0.747	0.786	0.913	0.898	0.127	0.106
DeepSeek-V3.2	0.812	0.827	0.943	0.912	0.079	0.073
qwen3-next-80b	0.751	0.774	0.912	0.891	0.127	0.102
kimi-k2-0905	0.747	0.813	0.911	0.910	0.133	0.073
Mean	**0.757**	**0.814**	**0.916**	**0.903**	**0.124**	**0.090**
SD	**0.021**	**0.036**	**0.010**	**0.016**	**0.017**	**0.017**

Table S14 Performance of LLMs across question types

Model	Accuracy			Score			F1			HR		
	SMC	MMC	JU	SMC	MMC	JU	SMC	MMC	JU	SMC	MMC	JU
gpt-5-mini	0.873	0.564	0.869	0.909	3.966	0.913	0.909	0.914	0.914	0.057	0.257	0.049
gemini-3-flash	0.879	0.561	0.897	0.917	3.966	0.926	0.917	0.916	0.926	0.054	0.251	0.051
grok-4.1-fast	0.870	0.551	0.871	0.908	3.956	0.910	0.908	0.911	0.911	0.054	0.274	0.058
claude-h-4.5	0.849	0.543	0.871	0.893	3.960	0.905	0.895	0.914	0.905	0.069	0.247	0.069
mistral-small-3.1	0.833	0.552	0.886	0.882	3.984	0.913	0.884	0.915	0.915	0.072	0.242	0.064
llama-4-mav	0.840	0.533	0.891	0.887	3.912	0.926	0.897	0.912	0.926	0.068	0.260	0.044
DeepSeek-V3.2	0.894	0.618	0.920	0.932	4.063	0.936	0.933	0.937	0.936	0.039	0.162	0.044
qwen3-next-80b	0.833	0.546	0.887	0.881	3.965	0.918	0.890	0.914	0.918	0.074	0.238	0.056
kimi-k2-0905	0.854	0.534	0.893	0.900	3.933	0.922	0.901	0.914	0.922	0.060	0.248	0.054
Mean	**0.858**	**0.556**	**0.887**	**0.901**	**3.967**	**0.919**	**0.904**	**0.916**	**0.919**	**0.061**	**0.242**	**0.054**
SD	**0.022**	**0.026**	**0.016**	**0.017**	**0.042**	**0.010**	**0.015**	**0.008**	**0.009**	**0.011**	**0.032**	**0.009**

Table S15 Performance of LLMs on technology readiness assessment

Model	Accuracy			RS			MAE		
	PA1	PA2	PA3	PA1	PA2	PA3	PA1	PA2	PA3
gpt-5-mini	0.746	0.758	0.737	3.715	3.776	3.475	0.339	0.336	0.383
gemini-3-flash	0.774	0.776	0.764	4.173	3.942	3.668	0.280	0.267	0.305
grok-4.1-fast	0.792	0.771	0.776	4.327	4.265	3.868	0.262	0.296	0.314
claude-h-4.5	0.798	0.785	0.784	3.663	3.643	3.494	0.240	0.251	0.263
mistral-small-3.1	0.771	0.774	0.800	3.955	3.950	3.847	0.280	0.271	0.253
llama-4-mav	0.758	0.765	0.767	3.939	3.850	3.747	0.305	0.294	0.296
DeepSeek-V3.2	0.769	0.780	0.803	3.747	3.746	3.430	0.287	0.269	0.244
qwen3-next-80b	0.765	0.764	0.758	3.755	3.753	3.552	0.287	0.300	0.294
kimi-k2-0905	0.785	0.753	0.787	3.930	3.966	3.682	0.267	0.301	0.265
Mean	**0.773**	**0.770**	**0.775**	**3.912**	**3.877**	**3.640**	**0.283**	**0.287**	**0.291**
SD	**0.016**	**0.010**	**0.021**	**0.222**	**0.182**	**0.162**	**0.028**	**0.025**	**0.042**

Table S16 Performance of LLMs on technology market forecasting

Model	BS			RS			ECE			AR		
	PF1	PF2	PF3	PF1	PF2	PF3	PF1	PF2	PF3	PF1	PF2	PF3
gpt-5-mini	0.247	0.235	0.235	3.763	3.961	3.955	0.182	0.168	0.150	1.345	1.376	1.328
gemini-3-flash	0.252	0.251	0.240	3.763	4.120	4.102	0.189	0.203	0.167	1.365	1.390	1.410
grok-4.1-fast	0.258	0.254	0.241	4.692	4.749	4.754	0.201	0.198	0.190	1.392	1.374	1.353
claude-h-4.5	0.253	0.249	0.241	3.778	3.892	3.901	0.193	0.191	0.183	1.254	1.334	1.331
mistral-small-3.1	0.259	0.240	0.237	3.997	3.991	4.000	0.198	0.146	0.155	1.370	1.305	1.346
llama-4-mav	0.260	0.236	0.230	4.012	4.015	4.102	0.196	0.133	0.130	1.363	1.360	1.392
DeepSeek-V3.2	0.240	0.228	0.220	4.299	4.329	4.506	0.157	0.148	0.117	1.402	1.398	1.466
qwen3-next-80b	0.265	0.245	0.242	4.644	4.964	4.976	0.215	0.198	0.179	1.445	1.459	1.452
kimi-k2-0905	0.249	0.237	0.221	3.302	3.590	3.563	0.194	0.153	0.131	1.452	1.365	1.391
Mean	**0.254**	**0.242**	**0.234**	**4.028**	**4.179**	**4.207**	**0.192**	**0.171**	**0.156**	**1.376**	**1.373**	**1.385**
SD	**0.008**	**0.009**	**0.009**	**0.451**	**0.434**	**0.450**	**0.016**	**0.027**	**0.026**	**0.059**	**0.043**	**0.051**

10 Performance of LLMs vs. human baseline and primary-metric tests

Table S17 Performance of LLM baselines across three tasks

Model	Task 1 Accuracy	Task 2 Accuracy	Task 3 BS
gpt-5-mini	0.784	0.720	0.235
gemini-3-flash	0.856	0.773	0.246
grok-4.1-fast	0.780	0.787	0.270
claude-h-4.5	0.708	0.800	0.243
mistral-small-3.1	0.716	0.773	0.253
llama-4-mav	0.756	0.760	0.255
DeepSeek-V3.2	0.800	0.787	0.236
qwen3-next-80b	0.792	0.733	0.257
kimi-k2-0905	0.804	0.787	0.254
Mean	**0.777**	**0.769**	**0.250**
SD	**0.046**	**0.027**	**0.011**

Table S18 Performance of human baselines across three tasks

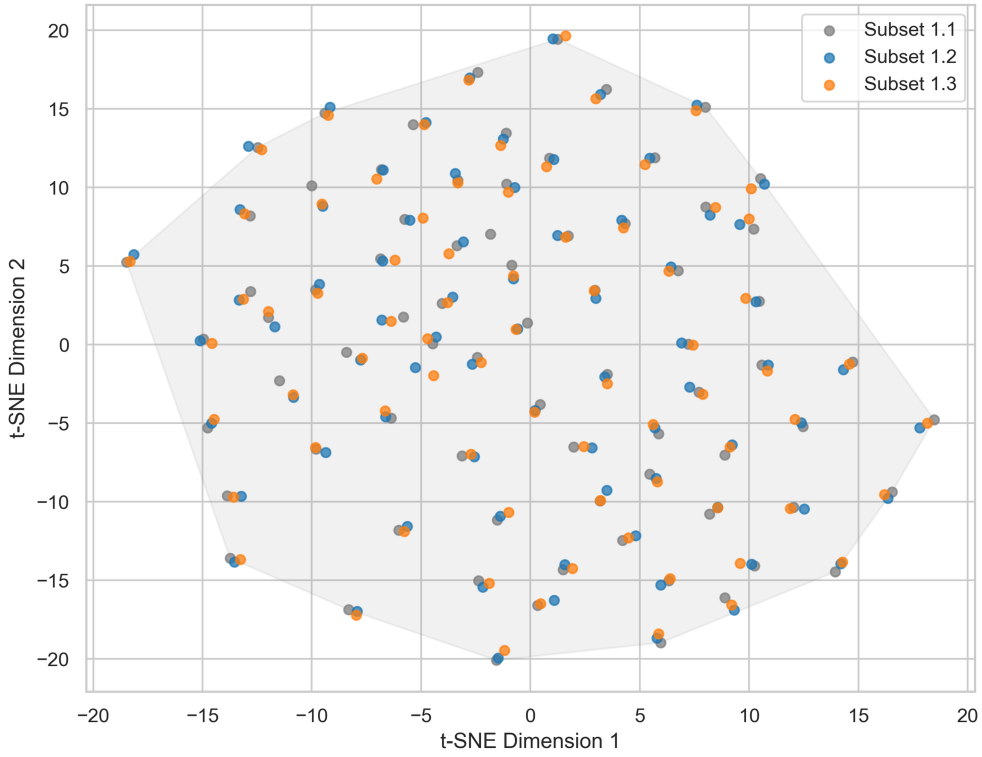
Human	Task 1 Accuracy	Task 2 Accuracy	Task 3 BS
Expert 1	0.836	0.760	0.244
Expert 2	0.832	0.733	0.248
Expert 3	0.800	0.720	0.263
Expert 4	0.796	0.707	0.264
Expert 5	0.776	0.707	0.274
Mean	**0.808**	**0.725**	**0.259**
SD	**0.025**	**0.022**	**0.012**

Table S19 Statistical tests on primary task metrics

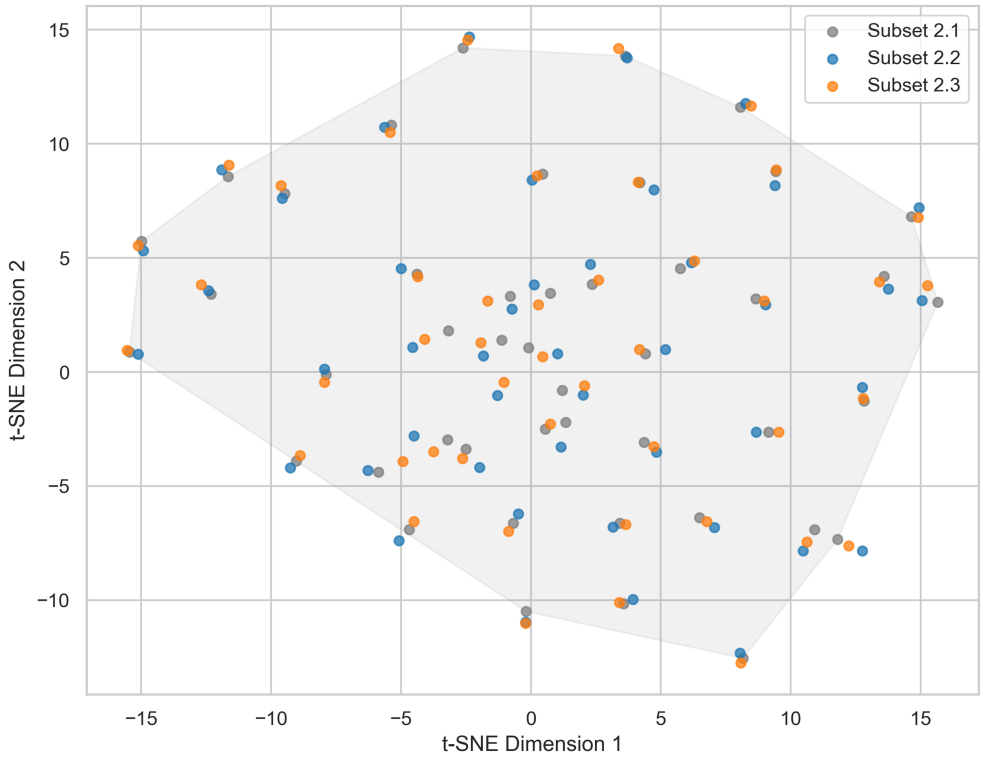
Comparison	Outcome	Test	Effect	Holm p
Model-ranking comparisons				
gemini-3-flash vs DeepSeek-V3.2	Task 1 Accuracy	McNemar	+0.056	< 0.001
grok-4.1-fast vs mistral-small-3.1	Task 2 Accuracy	McNemar	-0.002	0.892
DeepSeek-V3.2 vs llama-4-mav	Task 3 BS	Wilcoxon signed-rank	+0.010	0.034
Human-baseline comparisons				
LLM mean vs human mean	Task 1 Accuracy	Wilcoxon signed-rank	-0.031	0.065
LLM mean vs human mean	Task 2 Accuracy	Wilcoxon signed-rank	+0.044	1.000
LLM mean vs human mean	Task 3 BS	Wilcoxon signed-rank	+0.009	1.000
Prompt-strategy comparisons				
PA3 vs PA1	Task 2 Accuracy	Wilcoxon signed-rank	+0.015	0.006
PA3 vs PA2	Task 2 Accuracy	Wilcoxon signed-rank	+0.016	0.006
PF2 vs PF1	Task 3 BS	Wilcoxon signed-rank	+0.010	0.037
PF3 vs PF1	Task 3 BS	Wilcoxon signed-rank	+0.017	< 0.001
PF3 vs PF2	Task 3 BS	Wilcoxon signed-rank	+0.007	0.038

Model-ranking comparisons use the first- and second-ranked models in each task-level composite ranking shown in Fig. 1 in the main text. LLM mean and human mean denote averages across the nine evaluated LLMs and five experts, respectively. Effect is computed from paired item-level data in the direction shown in the Comparison column: for Accuracy, first comparator minus second comparator; for BS, second comparator's BS minus first comparator's BS, so positive values indicate lower BS for the first comparator. All tests were two-sided. McNemar tests used continuity correction; Wilcoxon signed-rank tests used Pratt zeros. Holm correction was applied separately within the model-ranking, human-baseline, Task 2 prompt, and Task 3 prompt families. Values are rounded to three decimals; values below 0.001 are shown as < 0.001. Task-level composite scores and TMscore were not tested.

11 Robustness test of the item synthesis pipeline

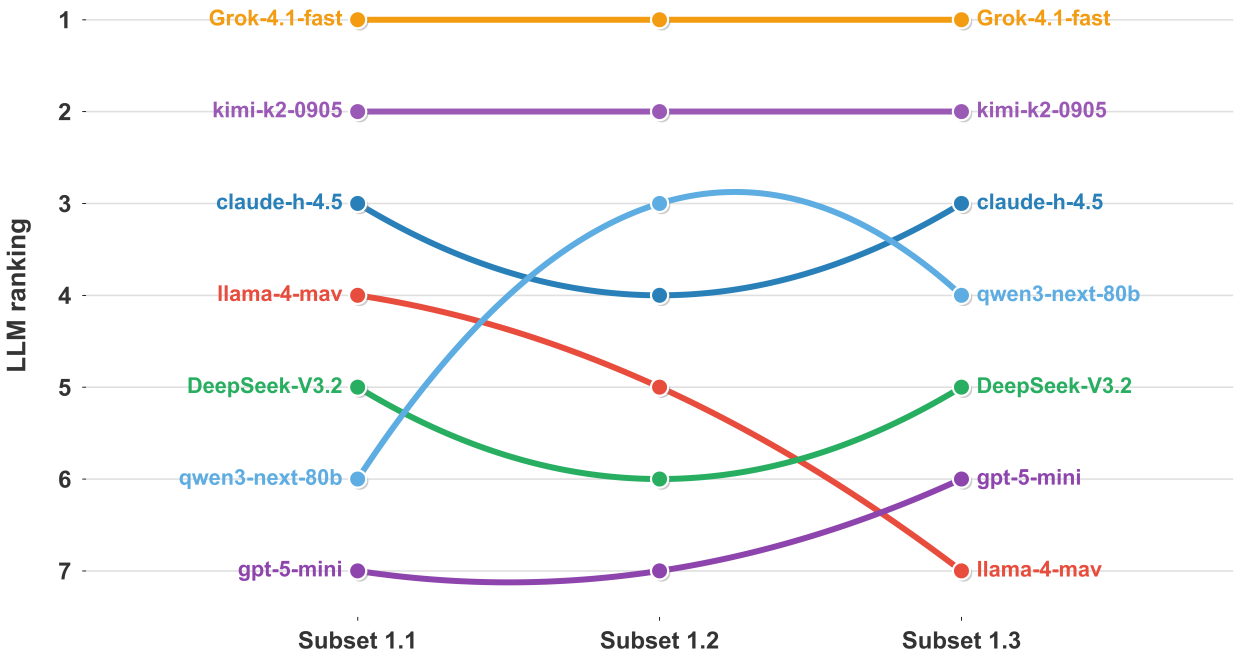


Task 2 ($n = 75$)

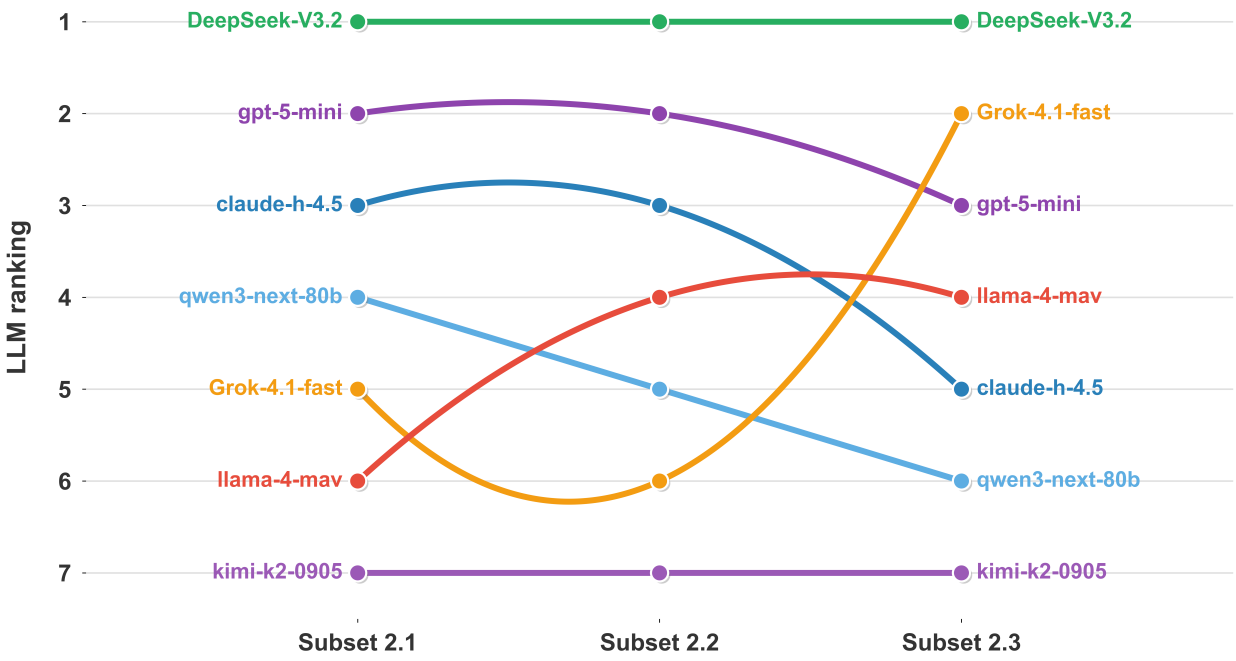


Task 3 ($n = 45$)

Fig. S5 Semantic embedding distributions of subsets for Tasks 2 and 3



Task 2 ($n = 75$)



Task 3 ($n = 45$)

Fig. S6 Stability of downstream model rankings across baseline and variant subsets for Tasks 2 and 3

12 Human–LLM agreement evaluation for Reasoning Score

The agreement evaluation used the same representative subsets as the human baseline evaluation. At a temperature setting of 0.7, we sampled model-generated reasoning outputs that covered all nine evaluated models within each task and included the corresponding prompting strategies, yielding 75 Task 2 responses and 45 Task 3 responses. The 120 sampled reasoning outputs were then independently scored by three experts for reasoning quality. Table S20 reports the corresponding score distributions.

Table S20 Reasoning Score distributions for Tasks 2 and 3

Score	Task 2 score distribution (%)		Task 3 score distribution (%)	
	Human consensus	M-Prometheus-7B	Human consensus	M-Prometheus-7B
1	0.0	1.3	0.0	0.0
2	6.7	5.3	2.2	0.0
3	14.7	2.7	15.6	8.9
4	64.0	78.7	57.8	53.3
5	14.7	12.0	24.4	37.8