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

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0168

Evaluating large language models on the core cognitive capabilities expected of technology managers: a hierarchical benchmark

Zhipeng ZHU, Guo LI, Mark Xuefang ZHU✉, Shanshan KONG, Zhenyu WANG, Wenting XIONG

School of Information Management, Nanjing University, Nanjing 210023, China

Abstract: Technology managers are instrumental in facilitating technology transfer. While large language models (LLMs) have demonstrated potential in this domain, evaluating them on the core cognitive capabilities expected of technology managers remains challenging due to the lack of domain-specific benchmarks. To address this challenge, we construct TMBench, a domain-specific benchmark for evaluating the cognitive capabilities expected of technology managers. It comprises a collection of 3474 questions spanning diverse scenarios, covering three tasks and five question types. We integrate typical metrics with LLM-as-a-Judge using both fixed and entropy-based weights to calculate the corresponding dimensional and composite scores. Leveraging this mechanism, we evaluate nine representative models under varying prompt configurations and temperature settings to investigate their capabilities, characterize their error patterns, and benchmark them against human experts. Our results show heterogeneous capabilities, with no single model ranking first across all three capability levels. For higher-order capabilities, LLMs show more favorable point estimates than the human baseline, but the differences are not statistically significant. Chain-of-thought-based prompting is associated with prompt-sensitive capability profiles, with more favorable open-weight performance in the two higher-order tasks. Error analysis identifies knowledge base, retrieval, and application deficits as the most prevalent error category. Building on these findings, this study provides a reference for model selection, prompt engineering, and targeted optimization when applying LLMs in technology transfer.

Key words: Technology transfer; Technology manager; Benchmark; Large language model (LLM)

1 Introduction

Technology transfer denotes the translation of technological innovations into practical applications, typically involving the movement of these innovations from research and development (R&D) organizations to recipient entities (Rogers et al., 2001). As a key element for economic development across all levels of industry (Morrissey and Almonacid, 2005), technology transfer validates academic research results in a real setting and improves industrial development and business processes. However, achieving these outcomes demands intensive collaboration between industry and academia throughout the research lifecycle (Gorschek et al., 2006). To meet this demand, technology managers operate at the interface of science and business to bridge the gap between resource requirements and availability (Lin et al., 2022). Such efforts facilitate the advancement of discoveries for the public good (Sheft, 2008). As vital champions and intermediaries, technology managers' core functions encompass matching supply with demand by aligning the divergent motivations of researchers and industry, navigating intellectual property (IP) and contractual barriers (Harrison, 2006), synthesizing technical and business intelligence (Rifkin et al., 1999), and fostering new ventures to facilitate commercialization (Association of University Technology Managers, 2024). However, navigating these functions presents challenges: reconciling cultural disparities between academia and industry, managing uncertainty during market entry (MacMillan and McGrath, 2004), and cultivating multidisciplinary skills (Cetindamar et al., 2016). Large language models (LLMs) address these challenges through the simulation of authentic industry scenarios (Genç, 2026), predictive forecasting (Dai et al., 2025), and cross-domain knowledge synthesis (Guo SK et al., 2025).

Although LLMs offer the potential to augment knowledge-

✉ Mark Xuefang ZHU, xfzhu@nju.edu.cn
Mark Xuefang ZHU, https://orcid.org/0000-0002-8244-5999
CLC number: TP18; G311

Received: June 2, 2026; Revision accepted: Aug. 9, 2026;
Crosschecked: Aug. 21, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

intensive tasks (Kampelopoulos et al., 2025), deploying them in high-stakes professional domains necessitates rigorous validation (Pham and Vo, 2024). Therefore, establishing standardized and domain-specific evaluation benchmarks is essential to ensure model reliability and proficiency in these domains (Chang YP et al., 2024). Such benchmarks provide practitioners with systematic tools to assess model performance (Zografos and Moussiades, 2025), thereby providing guidance for selecting suitable backbone models in real-world applications (Li JW et al., 2025). Existing benchmarks in technology transfer cover patent attorney qualifying examinations (Khera et al., 2025), patent claim generation (Jiang et al., 2025), and market trend prediction (Tak and Pele, 2025). However, existing benchmarks remain fragmented for evaluating LLMs in technology manager roles, highlighting the need for a dedicated evaluation system. The primary challenge lies in the context-specific formulation of the evaluation frameworks, data curation, and task designs.

In this study, we introduce TMBench, a domain-specific benchmark specifically designed to evaluate LLMs on the core cognitive capabilities expected of technology managers. It encompasses 3474 questions spanning three tasks and five question types, curated from professional textbooks and specialized platforms. These questions are constructed around domain-specific scenarios and have undergone a rigorous expert annotation process, human–LLM agreement evaluation, and robustness testing for model-dependent synthesis. By mapping tasks to hierarchical cognitive levels, TMBench characterizes task-specific capability profiles and cognitive alignment patterns of LLMs in this domain. Furthermore, it provides a reference for the responsible integration of artificial intelligence (AI) into professional workflows in technology transfer. To support a multidimensional evaluation, we integrate typical metrics with LLM-as-a-Judge using both fixed and entropy-based weights to calculate scores. Finally, we conduct an in-depth performance evaluation of nine representative LLMs under various prompting strategies and temperature settings. We also compare their performance against human experts on a distributionally representative 370-item subset. Our results show heterogeneous capabilities, with no single model ranking first across all three capability levels. For the two higher-order capabilities, LLMs are broadly comparable to the human baseline, showing higher task 2 Accuracy and lower task 3 Brier score (BS), but neither comparison supports a reliable LLM–human difference after correction. Furthermore, chain-of-thought (CoT)-based prompting strategies are associated with prompt-sensitive capability profiles in tasks 2 and 3, and open-weight models show more favorable performance under these prompts. Error analysis indicates that knowledge base, retrieval, and application deficits are the most prevalent category among 5967 retained high-frequency error instances (48.5%). Consequently, these findings support tailored model selection, prompt engineering, and targeted optimization for knowledge-boundary, contextual, and reasoning failures.

Our major contributions can be outlined as follows:

1. We propose a hierarchical evaluation framework combining Bloom's taxonomy with the core functions of technology managers to assess LLM capabilities across three progressive levels and benchmark their performance against human experts.

2. We construct TMBench, providing a robust instrument that establishes a standardized benchmarking pipeline and integrates typical metrics with LLM-as-a-Judge using both fixed and entropy-based weights to calculate scores.

3. We conduct extensive experiments to uncover insights into prompt sensitivity, temperature variability, and error patterns, thereby offering guidance for optimizing LLMs for technology transfer.

2 Background

2.1 Related works

Driven by the rapid advancement of LLM capabilities (Center for AI Safety et al., 2026), domain-specific benchmarks are undergoing a profound transformation from basic question answering to the verification of complex business logic (Busch and Leopold, 2024). This transformation prioritizes scientific rigor and alignment with business requirements (Liang et al., 2026), moving from single-dimensional metrics to systematic frameworks rooted in cognitive modeling (Chang CC et al., 2024). By leveraging Bloom's taxonomy, researchers have established hierarchical evaluation frameworks that progress from foundational knowledge memorization and understanding to higher-order capabilities (Sun et al., 2024). This structural shift redirects the assessment focus from mere fact retrieval to the comprehensive application of knowledge within complex scenarios (Xu LQ et al., 2025).

The operationalization of these hierarchical frameworks relies on data curation and task designs that emphasize authority and practicality (Nie et al., 2025). This is increasingly achieved by integrating authoritative sources, such as educational textbooks (Wang MS et al., 2026), with unstructured data (BenYoash et al., 2025). This curation strategy underpins hierarchical task designs that mirror cognitive progression: starting with knowledge-centric tasks for foundational domain mastery (Zhang et al., 2025) and progressing to material-oriented tasks for professional application, notably represented by material evaluation (Zhao XK et al., 2025) and forecasting (Schoenegger et al., 2024). For evaluation tasks, standard-adherence evaluation aims to replicate professional workflows. These tasks include detecting brand deviations in bid submissions (Liang et al., 2026) and checking the logical consistency of legal or financial texts (Saxena et al., 2025). Meanwhile, for forecasting tasks, the LLM-as-a-Prophet paradigm is integrated with retrieval-augmented generation (RAG) to evaluate LLM causal reasoning and strategic decision-making potential in uncertain environments, such as scientific discovery and financial markets (Halawi et al., 2024; Yang et al., 2025).

Beyond task designs, the experimental setups of benchmarks have transitioned toward a robustness-oriented paradigm characterized by multidimensional variable control and extensive parameter sensitivity analysis. Adhering to this paradigm, researchers have implemented variable partitioning across linguistic contexts, prompting strategies, and sampling temperatures to scrutinize LLM performance under varying conditions. First, proficiency in high-resource languages does not inherently translate to lower-resource or typologically

distinct linguistic systems (Reese and Smirnova, 2024). Second, reasoning strategies like CoT bolster complex problem solving via intermediate reasoning steps (Chadli et al., 2026). Third, lower temperature values lead to more deterministic and repetitive responses, whereas higher values encourage greater creativity and randomness (Khera et al., 2025).

The evaluation methodologies are also evolving. While typical metrics, including Accuracy and F1-score (Pei et al., 2025; Wang MS et al., 2026), remain standard for benchmarks, LLM-based evaluation methods, known as LLM-as-a-Judge (Zheng et al., 2023), have gradually become mainstream in natural language generation evaluation (Gao et al., 2025). These methods use advanced LLMs (e.g., Prometheus) for pairwise evaluation or pointwise scoring based on established evaluation criteria (Fan et al., 2025). By providing detailed and comprehensive judgments, they help mitigate the limitations of typical evaluation methods (Li YJ and Sun, 2026). Indeed, LLM-as-a-Judge can approximate human evaluations in some tasks (Wang RQ et al., 2025).

2.2 Research gaps

Domain-specific benchmarks have advanced across evaluation frameworks, data curation, task designs, experimental setups, and evaluation methodologies. However, their application to highly specialized and interdisciplinary domains reveals critical limitations, particularly within evaluation frameworks, data curation, and task designs. First, current taxonomies often remain too abstract or fragmented to capture cross-functional professional knowledge. Second, empirical data from real-world business scenarios remain scarce for these domains. Third, current task designs favor single-domain skills over the multidimensional capabilities essential for practical management. These limitations are particularly pronounced in technology transfer. While existing studies have explored the benchmarking of IP knowledge (Wang QY et al., 2025), automated technology readiness level (TRL) assessments (Betancourt et al., 2025), and general forecasting (Bastani et al., 2025), they remain fragmented and fail to evaluate the interdisciplinary capabilities demanded by technology managers. Crucially, technology managers require integrated core capabilities to execute tasks spanning IP management, technology readiness assessment (TRA), and market forecasting (COGR, 2021). These capabilities call for a unified benchmark that aligns authentic business data and real-world tasks with the professional functions of technology managers.

To address this need, we present TMBench, which adapts Bloom's taxonomy to stratify the core cognitive capabilities of technology managers across three hierarchical levels. These levels are operationalized as three distinct core tasks grounded in datasets derived from real-world business scenarios. Collectively, these tasks evaluate the multidimensional cognitive capabilities essential for technology managers. Furthermore, we manipulate variables across distinct linguistic contexts, progressive prompting strategies, and diverse sampling temperatures. By integrating typical metrics with LLM-as-a-Judge and using both fixed and entropy-based weights to calculate scores, TMBench yields granular insights into LLM capabilities across different tasks and experimental environments.

3 Methodology

3.1 Research framework

This study constructs TMBench to evaluate LLMs on the core cognitive capabilities expected of technology managers. First, we structure LLM capabilities into three hierarchical layers. By aligning these layers with the functions of technology managers, we operationalize them into three tasks. Second, we integrate task-specific sources encompassing training textbooks, NASA TechPort projects, and Manifold forecasting questions. Data for tasks 2 and 3 are retrieved programmatically via application programming interfaces (APIs). All collected data are refined by professional technology managers. Third, we implement task-specific pipelines that combine advanced LLMs for automated generation and auditing with expert manual review to support data fidelity. Within these pipelines, the raw textual corpus for task 1 is structured into question answering (QA) pairs via the Easy Dataset framework. Raw project records for task 2 are synthesized into TRL descriptions compliant with NASA guidelines. Backgrounds for task 3 are constructed through a web-search-based RAG approach with temporal constraints designed to reduce the risk of post-cutoff retrieval leakage. Ultimately, task 1 is finalized into 2586 questions through automated screening and expert cross-validation, while task 2 (554 questions) and task 3 (334 questions) are validated via human–LLM agreement evaluations. Subsequently, these tasks are used to evaluate the performance of nine representative LLMs across various temperature and prompt configurations.

3.2 Capability modeling and operationalization

While technology managers require cognitive, interactive, and execution-oriented capabilities in technology transfer (Lin et al., 2022), we prioritize the cognitive dimension because it forms the objectively measurable foundation necessary for practical execution. For instance, multi-stakeholder coordination, IP management, and commercialization practices presuppose a mastery of cross-domain knowledge and analytical evaluation (Kim JS and Kim, 2022). Hence, standardizing the evaluation of these cognitive prerequisites lays the necessary groundwork for future benchmarks. In practice, the cognitive demands of technology transfer follow a progressive complexity curve (Rifkin et al., 1999): from foundational multidimensional comprehension, through rigorous technology and maturity evaluation, to strategic commercialization foresight (MacMillan and McGrath, 2004; Gorschek et al., 2006; Association of University Technology Managers, 2024). To render these cognitive capabilities quantifiable, we align this cognitive progression with Bloom's taxonomy (Krathwohl, 2002), drawing on capability modeling from research (Yu et al., 2024; Badenes-Olmedo et al., 2026). Consequently, TMBench integrates a hierarchical evaluation framework comprising three core tiers: knowledge memorization and understanding, procedural application and analytical reasoning, and complex synthesis and predictive inference.

We operationalize these core capabilities by mapping them onto three tasks. First, technology managers rely on knowledge memorization and understanding to navigate cross-

domain information. Successful technology transfer requires mastering core technologies, IP law, and business principles (Kim JS and Kim, 2022). This multidisciplinary scope necessitates the retention and comprehension of domain-specific facts (Cetindamar et al., 2016). To evaluate LLM capabilities in this foundational tier, we design task 1—domain knowledge QA, formatted as single-select multiple-choice (SMC), multi-select multiple-choice (MMC), and true/false judgment (JU) questions. Second, technology managers employ procedural application and analytical reasoning to analyze developmental stages and assess technology readiness. Bridging the gap between laboratory R&D and commercial realization requires evaluating prototypes for deployment (Lane, 1999). To evaluate LLM capabilities in this analytical tier, we design task 2—TRA via scenario-based assessment questions. Third, technology managers exercise complex synthesis and predictive inference to forecast market trends. Technical breakthroughs must be aligned with unmet market demands (Lane, 1999). This requires integrating multi-source information to forecast market outcomes under uncertainty (Rifkin et al., 1999). To evaluate LLM capabilities in this predictive tier, we design task 3—technology market forecasting via scenario-based forecasting questions. Collectively, these tasks establish a structured cognitive evaluation pipeline: verifying multidisciplinary knowledge, diagnosing technical maturity, and forecasting probabilistic market trajectories. Consequently, performance across this pipeline provides a reliable proxy for evaluating LLM core cognitive capabilities in this domain.

3.3 Data collection

TMBench integrates task-specific data. For task 1, we curate a corpus from eight contemporary training textbooks for technology managers, including five published in China (Ni et al., 2024; CASTEM and NCSTE, 2025; Hunan University and NCSTE, 2025; Liu QY et al., 2025; MAIIS and NCSTE, 2025) and three in the U.S. (Schaufeld, 2015; COGR, 2021; Association of University Technology Managers, 2024), all of which are parsed into unified plain text. For task 2, we source completed technology projects from NASA TechPort (NASA, 2026). This dataset includes the final TRLs, project descriptions, and closed-out content. Task 3 leverages Manifold (Manifold, 2026), the world's largest social prediction platform, to retrieve forecasting questions, crowd-aggregated probabilities, and verified outcomes, selecting only cases created and resolved in 2025. Following established research (Halawi et al., 2024), we simulate early forecasting by setting a 25% lifecycle cutoff. After generating queries via direct questioning and logical decomposition, we use Google News's time-based search to filter post-cutoff results, reducing the risk of post-cutoff retrieval leakage and supporting the temporal integrity of historical scenarios. To ensure a systematic and reproducible collection process, the data for tasks 2 and 3 are retrieved programmatically via their respective public APIs. Finally, all collected data are curated and adapted by professional technology managers to support technical rigor and IP-standard compliance.

3.4 Data processing

For task 1, we use the Easy Dataset framework (Miao et al., 2025) to systematically process the collected textual corpus via hybrid chunking and intelligent text cleaning. This pipeline initially extracts 19 207 QA pairs, which are standardized to ensure uniform option formats and semantic independence. Notably, the Chinese questions are retained in their original language to prevent translation-induced terminology distortion. This approach mitigates English-centric bias, ensuring an authentic assessment of the localized knowledge retrieval and multilingual reasoning capabilities of contemporary LLMs (Li SW et al., 2025). Building on this standardized bilingual foundation, we implement a rigorous, multistage validation process to support data fidelity. First, automated scripts are used to eliminate formatting errors and logical anomalies. Second, consistent with recent research (Li KY et al., 2025; Tao and Xu, 2026), Gemini 3 Flash and DeepSeek-V3.2 serve as automated judges to verify answer solvability and facilitate topic classification. We select them for their broad internal knowledge and high-throughput capabilities (Georgiev et al., 2024; Liu AX et al., 2025). This dual-model approach enables the rapid verification of large-scale data while mitigating single-model biases. This phase, combined with question type adaptation, refines the dataset to 5719 items. Finally, expert cross-validation is conducted to manually filter out ambiguous or invalid entries, culminating in a curated dataset of 2586 items.

For task 2, data processing commences with 19 160 raw project records. Filtering incomplete entries yields 6073 projects. This subset is then subjected to a two-stage framework to ensure technical accuracy. During the generation phase, DeepSeek-V3.2 is used to synthesize TRL descriptions based on project descriptions and closed-out content, excluding software-based simulations and forward-looking statements in favor of verified hardware metrics and completed physical testing. This model is selected for its advanced logical reasoning and ability to follow complex, user-defined constraints (Liu AX et al., 2025), as well as its proven utility in recent studies (Xu K and Lian, 2026). This process yields 631 candidates. In the subsequent auditing phase, M-Prometheus-7B (Pombal et al., 2025) functions as an independent evaluator (see Table S1 in the supplementary materials for the auditing protocol) and assesses the descriptions against the TRA guide (NASA, 2020). Only entries demonstrating strict compliance (score $\geq$ 4 out of 5) with the official definitions of hardware fidelity, test environment, and integration level are preserved. The final dataset comprises 554 high-quality items.

For task 3, from an initial pool of 13 163 raw records, we refine the dataset to 550 entries by filtering out duplicates as well as ambiguous, subjective questions. After inferring explicit resolution criteria and generating targeted search queries, we use a DeepSeek-V3.2-driven, web-search-based RAG pipeline (Vu et al., 2024) to retrieve the initial snippets and generate time-bound background briefings for each entry (see Table S4 in the supplementary materials for the complete pipeline). The pipeline first enforces strict time-based constraints during retrieval to reduce temporal noise and the risk of post-cutoff retrieval leakage, and then ranks the remaining search snippets by their evidentiary significance to retain the top 15. By extracting key metrics and addressing information gaps with the latest available pre-cutoff data and base rates, the model synthesizes these materials into dense, temporally constrained contexts. A subsequent background quality assessment

excludes items with substandard context relevance, retaining 421 of them. Finally, after a manual review to remove vague and hallucinated items (see Table S5 in the supplementary materials for a representative example), we finalize a high-quality dataset comprising 334 items.

To further validate the quality of the datasets for tasks 2 and 3, we randomly select subsets (150 and 100 items, respectively) and evaluate them from human and LLM perspectives. First, three experts score the items on an ascending 5-point ordinal scale using task-specific rubrics. For task 2, the rubric enforces the same NASA TRL definitions used in the prior auditing phase. For task 3, the rubric assesses predictive background quality based on targeted relevance, information density, and evidentiary support. This evaluation demonstrates substantial reliability among the human experts, yielding Fleiss' kappa values (Deroy et al., 2026) of 0.75 for task 2 and 0.71 for task 3. Second, we establish a human consensus benchmark by accepting the initial unanimous scores and resolving any disparities through discussion. Third, we employ DeepSeek-V3.2 to independently score the same items, serving as the LLM benchmark. Finally, we compute quadratic weighted kappa (QWK) scores (DeVore, 2025) to measure the alignment between human consensus and LLM evaluations. The resulting QWK values of 0.70 and 0.77 indicate substantial human–LLM agreement. Further details of the human–LLM agreement evaluation are provided in Section 3 in the supplementary materials.

3.5 Data statistics

TMBench comprises three tasks and 3474 detailed questions (sample items are provided in Section 4 in the supplementary materials). Task 1 comprises 2586 questions divided into three aspects. Among them, the domain is relatively balanced across 37.9% fundamental knowledge (FK), 31.1% professional practice (PP), and 31.0% policies and regulations (PR); the Chinese context accounts for 72.3% compared with 27.7% for the U.S.; the question type is distributed as 48.3% SMC, 30.4% MMC, and 21.3% JU. Task 2 contains 554 questions, and task 3 comprises 334 questions.

Task 1 spans three domains and covers 19 topics, as detailed in Table 1. The total number of questions for each topic ranges from 57 to 331. The top three topics are the IP rights system and legal basis, industrial economics and corporate management, and technology transfer market and service agencies. The distribution of the most prevalent topics indicates that task 1 is fundamentally grounded in institutional IP frameworks and market-driven technology transfer mechanisms. The average distribution of SMC, MMC, and JU per topic approximates ratios of 9:7:4 in China and 7:2:3 in the U.S. Despite the divergence in MMC proportions, the Chinese and U.S. contexts maintain structural consistency through the larger share of SMC questions and a combined choice-based question coverage above 75%.

3.6 Evaluation protocols

Task 1 is assessed using a specialized experimental pipeline designed to ensure the robustness of the evaluation. For each test instance, the candidate choices are shuffled across multiple inference trials to mitigate positional bias. Each model is required to complete three repeated trials per question with CoT reasoning enabled. Furthermore, two localized system

Table 1 Distribution of domain knowledge QA

Domain	Topic	Number of questions						Total
		China			U.S.			
		SMC	MMC	JU	SMC	MMC	JU	
PR	Legal systems for technology transfer	87	38	25	11	2	7	170
	Fiscal and tax incentives	46	8	15	4	0	1	74
	Assets supervision and compliance	34	8	23	12	3	8	88
	IP judicial protection and dispute resolution	20	2	10	19	0	8	59
	IP rights system and legal basis	93	42	19	108	33	36	331
	Income distribution and equity incentives	36	11	18	11	1	3	80
FK	Technology transfer market and service agencies	53	101	21	13	10	15	213
	Key industries and frontier technology	62	84	16	9	4	6	181
	Sci-tech finance and capital markets	42	38	13	23	12	7	135
	Innovation management theory	48	32	27	12	1	5	125
	Talent development in technology transfer	16	25	19	17	6	12	95
	Industrial economics and corporate management	95	81	19	24	6	7	232
PP	Market analysis and marketing	9	21	19	11	4	4	68
	Technology valuation and maturity assessment	38	76	23	14	4	2	157
	Business planning and financing	25	30	29	14	3	4	105
	Negotiation strategies and communication	36	18	21	19	2	8	104
	Technology contract drafting	17	19	25	43	12	24	140
	Patent application and operation	50	20	19	51	20	12	172
	Technology trading process and registration	22	8	18	5	1	3	57
Total		829	662	379	420	124	172	2586

PR: policies and regulations; FK: fundamental knowledge; PP: professional practice

prompts are developed for the Chinese and U.S. contexts to align the model's persona with regional knowledge. Subsequently, output parsing is facilitated by regular expressions and heuristic rules to systematically extract the final answers. Detailed system prompt configurations are provided in Section 5 in the supplementary materials.

In task 2, the models are benchmarked on their capability to analyze the provided technical description materials. Each model is tasked with generating a specific TRL score (ranging from 1 to 9) with detailed reasoning. The evaluation leverages three distinct prompt variations, all anchored by a unified instructional core that mandates each model to adopt the persona of a strict technology manager. This shared foundation strictly enforces the official NASA TRL definitions alongside a suite of conservative assessment rules. Notably, a hierarchy rule is implemented to ensure that a TRL level is awarded only when all criteria for that level and its predecessors are explicitly satisfied. Building on these shared constraints, three prompts are designed to assess the model's sensitivity to varying instructional depths. PA1 serves as a zero-shot baseline. It requires the models to assess technical descriptions based solely on definitions and constraints without auxiliary examples. PA2 serves as a few-shot prompt. It integrates three representative examples to guide the mapping of technical features to the TRL levels. PA3 serves as a few-shot CoT prompt. It incorporates explicit reasoning within examples and requires a single-sentence evidence analysis prior to the final determination. Full configurations for PA1–PA3 are available in Section 6 in the supplementary materials.

In task 3, model performance is assessed through the analysis of standardized inputs, including the question, background, resolution criteria, and retrieval date. Each model is required to generate a structured output comprising three distinct components: a rigorous reasoning process, a probability score strictly bounded within $[0, 1]$, and a qualitative confidence assessment anchored to a predefined 10-point scale (ranging from almost no chance to almost certain) (Park et al., 2026). This process is guided by three distinct prompting strategies designed to test various reasoning frameworks. PF1 serves as a zero-shot baseline for direct prediction. It requires models to generate assessments without prescribed intermediate steps. PF2 serves as a zero-shot CoT. It implements a superforecaster CoT approach, which guides models through a structured seven-step workflow to rephrase the question, evaluate conflicting arguments, and aggregate considerations based on expert forecasting principles. PF3 serves as an evidence-grounded (EG) zero-shot CoT. It is built upon the structural framework of PF2 but enforces strict source integrity by mandating that all evidence be derived exclusively from the provided background. Full configurations for PF1–PF3 are available in Section 7 in the supplementary materials.

4 Experiments

4.1 Setup

We design a multifaceted evaluation framework to benchmark LLM capabilities. LLM performance is evaluated across three distinct tasks, each necessitating tailored prompting strategies and specialized testing procedures to accommodate their unique requirements for output content and formatting standards. Since temperature can influence the diversity of model-generated responses (Betancourt et al., 2025), each task is evaluated under two settings: 0.0 and 0.7. We select a temperature setting of 0.7 as the primary configuration because it strikes a balance between stability and creativity (Jung et al., 2025), facilitates multiple different ways of thinking that are essential for complex tasks (Wang XZ et al., 2023), and aligns with standard practices in prior research (Yao et al., 2023). Moreover, since ChatGPT is widely assumed to operate at a default temperature of 0.7 (Peters and Chin-Yee, 2025), this setting mirrors the real-world scenarios encountered by technology managers. Consequently, for their multifaceted tasks, this temperature strikes an effective tradeoff: It enables necessary analytical flexibility while mitigating the risk of factual hallucinations inherent in extreme temperatures.

To implement this framework, the entire pipeline is developed using Python 3.11. The evaluated LLMs are accessed via APIs, while M-Prometheus-7B is deployed as an LLM-as-a-Judge on a cloud graphics processing unit (GPU) server equipped with an RTX 4090 24 GB graphics card. All API interactions and data processing are performed locally on a laptop with an Ultra 9 285H central processing unit (CPU), RTX 5060 8 GB graphics card, and 32 GB of random access memory (RAM). All experiments were performed between January 20 and February 22, 2026.

4.2 Models

To ensure a comprehensive evaluation, we select nine high-performing models that represent a diverse spectrum of architectures, organizations, and access paradigms. These models are categorized into two primary groups: proprietary and open-weight models.

For the proprietary models, we evaluate four state-of-the-art efficiency-oriented variants designed for high-throughput reasoning: GPT-5 mini (OpenAI, 2026) and Gemini 3 Flash (Google DeepMind, 2025), which represent the latest advancements in lightweight frontier models; Grok 4.1 Fast, noted for its extended context handling (xAI, 2025); Claude Haiku 4.5 (Claude-H-4.5), the high-efficiency iteration of the Claude 4.5 family (Anthropic, 2026).

For the open-weight models, our selection covers broadly adopted architectures from U.S., Chinese, and French developers. We assess Llama 4 Maverick (Llama-4-Mav), a sparse mixture-of-experts (MoE) model with 17B active parameters (Meta AI, 2025). Together with the four proprietary models mentioned above, it completes our selection of U.S. models. To ensure linguistic and cultural diversity, we also include three leading Chinese models: DeepSeek-V3.2, a high-compute variant optimized for reasoning (DeepSeek AI, 2025); Qwen3-Next-80B-A3B-Instruct (Qwen3-Next-80B), an instruction-tuned model with 3B active parameters tailored for RAG workflows (Alibaba Group, 2025); Kimi-K2-Instruct-0905 (Kimi-K2-0905), a 1T-parameter MoE architecture (Moonshot AI, 2025). Finally, we include the French model Mistral Small 3.1, a 24B-parameter model noted for its latency optimization (Mistral AI, 2025).

4.3 Metrics

In task 1, model performance is quantified using a multi-dimensional framework. For SMC and JU, a binary scoring system is employed (1.0 for a correct response and 0.0 otherwise), whereas MMC questions use tiered logic, awarding 5.0 points for full matches and 3.0 points for proper subsets, with zero points assigned if any incorrect selections are present. Within this schema, Accuracy adheres to a strict correctness criterion, requiring a perfect match across all three shuffled iterations to be deemed successful. Concurrently, the average score (Score) reflects the arithmetic mean of the points awarded per question. To provide a granular assessment of selection quality, the F1-score (F1) measures the harmonic mean of the option-level precision and recall. Finally, the hard rate (HR) quantifies persistent knowledge gaps or robust errors, defined as the percentage of questions where the model fails to provide a correct answer in all three shuffled trials.

In task 2, model performance is quantified using three primary metrics: Accuracy, mean absolute error (MAE), and an LLM-as-a-Judge reasoning score (RS). Accuracy denotes the proportion of assessments where the predicted TRL aligns perfectly with the expert-validated ground truth. MAE (Eq. (1)) quantifies the average deviation between the predicted TRL ($\hat{y}_i$) and the ground truth (y_i) across N instances, thereby characterizing the severity of assessment discrepancies. Finally, RS measures the logical rigor and structural coherence of the model's justifications, indicating whether the derived TRL is supported by evidence-based reasoning. This metric is implemented via M-Prometheus-7B (Pombal et al., 2025), an open-source LLM-as-a-Judge framework that evaluates reasoning quality on a standardized 1–5 scale (see Table S2 in the supplementary materials for the reasoning evaluation). This assessment uses a custom rubric that systematically examines four critical dimensions: 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. By integrating these factors, the framework provides a reproducible and granular measure of how well the model supports its decisions.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{1}$$

In task 3, model performance is evaluated using four complementary metrics: BS, expected calibration error (ECE), average return (AR), and RS. To quantify accuracy, we employ BS (Eq. (2)), a proper scoring rule measuring the mean-squared difference between the predicted probabilities p_i and actual outcomes o_i over N instances. Although models generate qualitative confidence assessments, we use ECE (Eq. (3)) to assess the calibration of p_i against o_i, where N total samples are partitioned into $M = 10$ equal-width bins (B_m) of size $|B_m|$. Thus, we provide a mathematical measure of calibration that sidesteps the semantic ambiguity inherent in linguistic confidence labels. Beyond statistical accuracy, AR (Eq. (4)) measures practical utility relative to the crowd forecast as the mean gross payoff from fixed one-unit stakes over N trials. For trial i, the model bets on YES if $p_i > q_i$ and on NO otherwise, where $q_i \in (0, 1)$ is the crowd probability. Finally, we introduce RS to assess the quality of the rationale underlying the numerical forecast. Facilitated by M-Prometheus-7B, this metric evaluates the logical quality of the forecasting process across five dimensions: context fidelity, evidence extraction, logical combination, uncertainty consideration, and probability mapping (see Table S3 in the supplementary materials for the reasoning evaluation).

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{2}$$

$$\mathrm{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\frac{1}{|B_m|}\sum_{i\in B_m} o_i - \frac{1}{|B_m|}\sum_{i\in B_m} p_i\right|, \tag{3}$$

$$\mathrm{AR} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\dfrac{o_i}{q_i}, & p_i > q_i,\\[2ex] \dfrac{1-o_i}{1-q_i}, & p_i \le q_i.\end{cases} \tag{4}$$

Based on the evaluation of individual metrics, we compute the scores for each task and the aggregate total score across all tasks. The weighting strategy prioritizes core indicators at 50%, with the entropy method determining the distribution of the remaining weights. This composite weighting scheme is adopted because fundamental metrics like Accuracy and BS are widely used core indicators in LLM evaluation (Tofeeq et al., 2025; Yang et al., 2025), whereas entropy weights support data-driven weight allocation for the secondary metrics. Specifically, Accuracy serves as the core metric (50%) for tasks 1 and 2, with the former supplemented by F1 (29.6%) and Score (20.4%) and the latter complemented by RS (35.1%) and MAE (14.9%). For task 3, BS is designated as the core indicator (50%), with the remaining weights allocated to AR (14.6%), ECE (19.8%), and RS (15.6%). These weighted values are then synthesized to derive the final performance scores.

In addition to standard metrics, we conduct an in-depth analysis of high-frequency errors, which are defined as questions where at least five of the nine evaluated models fail. Specifically, failure is characterized as providing an incorrect answer in tasks 1 and 2 or performing worse than random guessing (BS $>$ 0.25) in task 3. This process identifies 710 high-frequency-error questions, which are subsequently examined at the level of model-question error instances, with each instance representing one evaluated model's failure on one retained question.

4.4 Human baselines

Since it is impractical for humans to complete the full TMBench (Gignac and Ilić, 2025), we establish human baselines by evaluating experts on three distributionally representative subsets totaling 370 items, above the typical 10% sampling ratio of prior research (Xu LQ et al., 2025). To support the representativeness of these subsets, we employ constraint-based sampling via Monte Carlo simulations. Specifically, we use aggregate model performance profiles (evaluated at a temperature of 0.7 to reflect operational usage) as statistical proxies to verify distributional equivalence. These profiles are measured by Accuracy for tasks 1 and 2 and BS for task 3. Accuracy and BS are selected as core metrics because they serve as intuitive outcome measures aligned with the decision-centric

priorities of technology managers. To ensure statistical representativeness, candidate subsets are required to satisfy distributional alignment thresholds against TMBench: Spearman correlation (≥ 0.95) for ranking fidelity (Zhao ZH et al., 2025) and MAE (≤ 0.04) for performance deviation. Specifically, for task 1 ($n = 250$), we employ stratified sampling across domains, contexts, and question types (detailed distribution in Section 8 in the supplementary materials) to generate a robust subset (Spearman = 0.97, MAE = 0.01). For task 2 ($n = 75$), items are directly sampled using the same thresholds (Spearman = 0.97, MAE = 0.01). For task 3 ($n = 45$), we apply identical thresholds to yield a subset whose performance distribution aligns with the full set (Spearman = 0.96, MAE = 0.01). Consequently, comparative results from these subsets provide a distributionally aligned basis for comparing human and LLM performance across TMBench.

The experts who participate in the human-baseline comparison are independent of those involved in data-quality validation, thereby avoiding prior item exposure. We employ convenience sampling to recruit five experts—certified technology managers with aviation backgrounds and over five years of professional experience. To ensure a fair comparison, following recent guidelines (Wei et al., 2025), we enforce equivalence in experimental setups between the experts and LLMs. First, we ensure task and tool equivalence by presenting them with instructions identical to those provided to the LLMs and requiring them to rely strictly on their own knowledge. Second, we maintain consistent scoring methods and rubrics for all shared measures, comparing performance exclusively on the 370-item subset. Specifically, the aggregated human baseline is computed as the experts' average Accuracy for tasks 1 and 2, and their average BS for task 3. This comparison serves to benchmark LLMs versus humans in the three targeted capabilities. Notably, we do not require written reasoning, as articulating cognitive processes imposes an excessive cognitive load and is time-prohibitive.

4.5 Statistical analysis

For statistical analysis, formal inference is restricted to predefined primary item-level metrics: Accuracy for tasks 1 and 2 and BS for task 3. Paired McNemar tests are used only for binary top-versus-runner-up Accuracy comparisons. Wilcoxon signed-rank tests with Pratt zeros are used for human-baseline comparisons, prompt-strategy comparisons, and BS comparisons. All tests are two-sided. The item is the unit of analysis: Paired binary correctness indicators are used for top-versus-runner-up Accuracy comparisons, paired item-level aggregate Accuracy values are used for human-baseline comparisons, and paired item-level BS values are used for task 3 comparisons. Planned comparison families are defined a priori as model-ranking comparisons, human-baseline comparisons, task 2 prompt comparisons, and task 3 prompt comparisons; Holm correction is applied separately within each family. Task-level composite scores and TMscore are treated as summary indices derived from the predefined TMBench scoring rule. The planned tests are summarized in Table S19 in the supplementary materials.

5 Main results

In this section, we present the experimental results from TMBench under a temperature setting of 0.7, while those at 0.0 are detailed in Section 9 in the supplementary materials. Based on these primary findings, we derive insights regarding the overall performance and rankings of the models.

5.1 Task-wise LLM performance

To provide a comprehensive evaluation of model capabilities, Table 2 presents the task-wise performance of the nine evaluated models. Task 1 suggests that the knowledge memorization and understanding capability reflects a foundational grasp of core domain knowledge, but remains constrained by a gap between option-level and question-level performance. Specifically, an overall Accuracy of 0.778 is consistent with this foundational grasp, while the higher mean F1 (0.921) points to a bottleneck: The models effectively discriminate between options but frequently fail to produce exactly matched MMC answer sets. This pattern is also reflected in the mean Score of 1.838, indicating prevalent partial-credit outcomes where models conservatively select a subset of correct options to avoid distractors. Accordingly, translating knowledge memorization and understanding into answer completeness appears to require more rigorous reasoning.

Task 2 indicates that the procedural application and

Table 2 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	0.787	1.809	0.923	0.751	3.649	0.342	0.237	3.801	0.162	1.361
Gemini 3 Flash	0.866	1.919	0.951	0.774	3.889	0.285	0.254	4.020	0.197	1.367
Grok 4.1 Fast	0.764	1.830	0.908	0.780	4.182	0.288	0.251	4.730	0.195	1.389
Claude-H-4.5	0.706	1.780	0.899	0.780	3.578	0.267	0.245	3.844	0.177	1.324
Mistral Small 3.1	0.721	1.759	0.900	0.782	3.922	0.265	0.240	3.997	0.163	1.348
Llama-4-Mav	0.763	1.827	0.919	0.771	3.850	0.293	0.242	4.121	0.148	1.375
DeepSeek-V3.2	0.811	1.876	0.934	0.789	3.626	0.265	0.232	4.343	0.148	1.394
Qwen3-Next-80B	0.785	1.859	0.921	0.770	4.106	0.291	0.250	4.879	0.203	1.420
Kimi-K2-0905	0.798	1.882	0.933	0.773	3.868	0.279	0.239	3.427	0.162	1.352
Mean	0.778	1.838	0.921	0.774	3.852	0.286	0.243	4.129	0.173	1.370

analytical reasoning capability shows foundational competency but also a reliability gap in performance outcomes and reasoning processes. At the outcome level, an overall Accuracy of 0.774 suggests a functional yet inconsistent grasp of technology readiness assessment. Simultaneously, the mean MAE of 0.286 is above the theoretical minimum of 0.226 and corresponds to an average absolute deviation of 1.265 levels per error. This pattern indicates level-skipping deviations and a deficiency in fine-grained assessment. At the process level, the mean RS of 3.852 suggests generally competent reasoning quality, but it remains below a consistently rigorous reasoning level. Furthermore, higher Accuracy does not necessarily align with a higher RS. For instance, while DeepSeek-V3.2 achieves the highest Accuracy and is tied to the lowest MAE, Grok 4.1 Fast attains the highest RS. Accordingly, enhancing this specific capability requires a dual refinement of outcome reliability and reasoning logic.

Task 3 indicates that the complex synthesis and predictive inference capability has slightly more favorable mean outcomes than the non-informative and market consensus benchmarks, but the margins remain insufficient for professional forecasting. Specifically, the mean BS of 0.243 remains slightly below the 0.250 non-informative baseline, while the mean AR of 1.370 is above the 1.000 break-even threshold. However, the mean ECE of 0.173 is above 0.100, indicating substantial miscalibration (Yang et al., 2025). In contrast, the mean RS of 4.129 suggests relatively strong reasoning-process scores, but this does not offset the calibration and decision-utility limitations. Accordingly, improving this specific capability requires concurrent gains in predictive accuracy, calibration, and practical utility while preserving reasoning quality.

5.2 LLM ranking by task

Fig. 1 reports score-based capability profiles across the three TMBench tasks. The three tasks operationalize the distinct cognitive capabilities required of technology managers. As shown in Table S19 in the supplementary materials, at the primary-metric level, Gemini 3 Flash has higher task 1 Accuracy than DeepSeek-V3.2 (Accuracy gain = 0.056, Holm-adjusted $p < 0.001$). In task 2, the score-based top comparison between Grok 4.1 Fast and Mistral Small 3.1 is not supported at the Accuracy level (Accuracy difference = −0.002, Holm-adjusted $p = 0.892$). In task 3, DeepSeek-V3.2 achieves lower BS than Llama-4-Mav (BS reduction = 0.010, Holm-adjusted $p = 0.034$). Overall, the score-based leading models vary across the three core capabilities, so technology managers should tailor model selection to task characteristics.

5.3 LLM ranking by TMscore

To evaluate the overall model profiles, Fig. 2 ranks all nine evaluated models based on TMscore, a composite summary index derived from the average of the three task-level composite scores. Designed to balance expansive factual coverage with high-density reasoning without over-rewarding memorization, this equal-weighting approach (Srivastava et al., 2023) yields a TMscore range of 73.9–87.9. The descending TMscore order is DeepSeek-V3.2, Kimi-K2-0905, Gemini 3 Flash, Qwen3-Next-80B, Llama-4-Mav, Grok 4.1 Fast, Mistral Small 3.1, GPT-5 mini, and Claude-H-4.5.

5.4 Comparison with human baselines

Compared with the human baselines, LLMs have lower task 1 Accuracy (0.777 vs. 0.808; Accuracy difference = −0.031, Holm-adjusted $p = 0.065$). For the two higher-order tasks, LLMs show more favorable results: Task 2 Accuracy is higher (0.769 vs. 0.725; Accuracy gain = 0.044, Holm-adjusted $p = 1.000$), and task 3 BS is lower (0.250 vs. 0.259; BS reduction = 0.009, Holm-adjusted $p = 1.000$). These tests do not support reliable LLM–human differences after Holm correction. Across the five experts, baseline variability is limited but non-negligible: Task 1 Accuracy has a standard deviation (SD) of 0.025 and ranges from 0.776 to 0.836; Task 2 Accuracy has an SD of 0.022 and ranges from 0.707 to 0.760; Task 3 BS has an SD of 0.012 and ranges from 0.244 to 0.274. The full

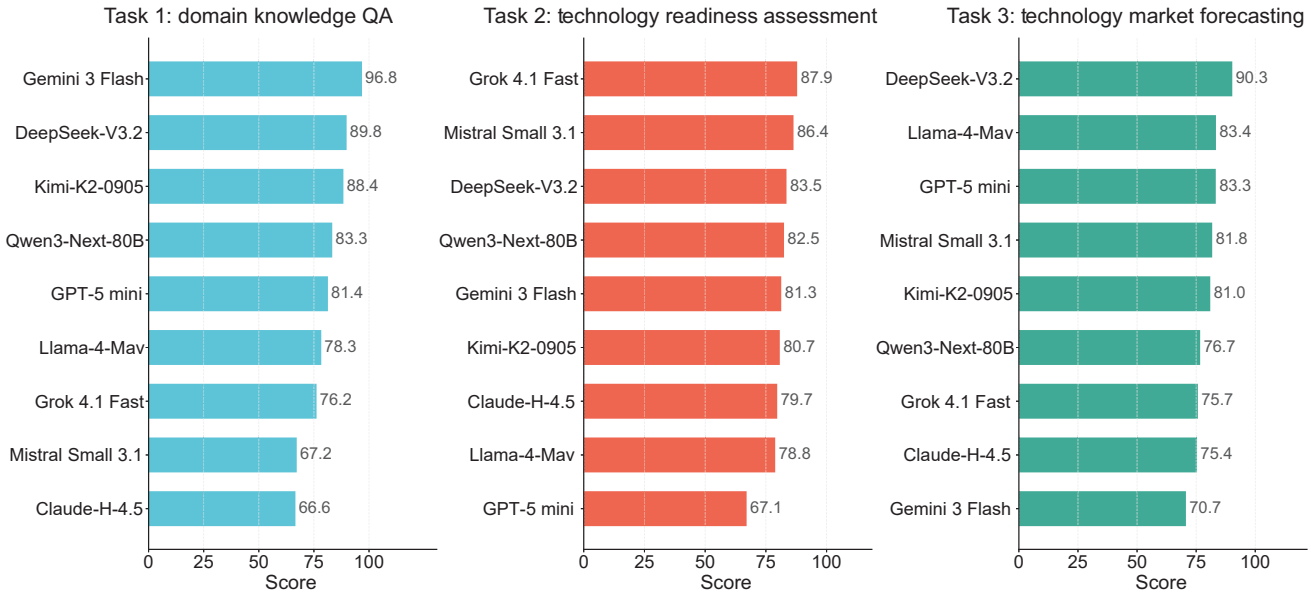


Fig. 1 LLM rankings across three tasks

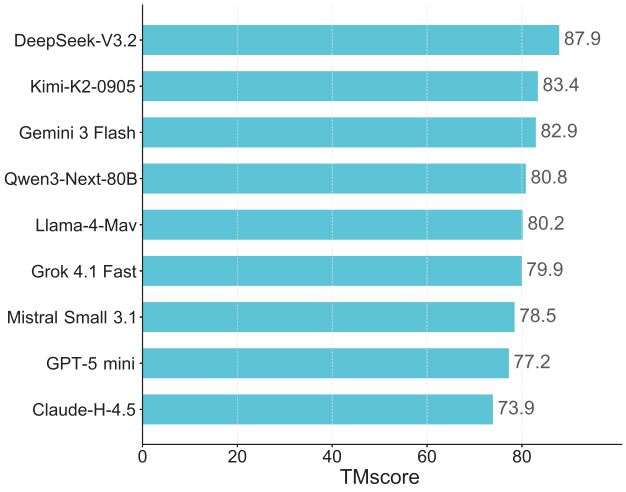


Fig. 2 LLM rankings based on TMscore

set of planned primary-metric tests is reported in Table S19 in the supplementary materials.

LLMs show lower levels of knowledge memorization and understanding than humans, which may reflect inherent static and cognitive constraints, particularly limitations in domain-specific knowledge. The factual knowledge embedded within them is restricted by temporal boundaries, rendering their fixed memory susceptible to becoming outdated (Huang et al., 2025). In vertical domains, LLMs exhibit deficits in metacognition and a lack of common sense reasoning (Kim J et al., 2025). They frequently fail to grasp cultural nuances or navigate ambiguous scenarios that permit multiple valid answers (Chiu et al., 2025).

For the two higher-order capabilities, LLMs show higher task 2 Accuracy and lower task 3 BS than the human baseline, but neither comparison supports a reliable LLM–human difference after correction. In procedural application and analytical reasoning, this pattern is consistent with LLMs' capacity for systematic execution. On complex synthesis and predictive inference, the slightly lower BS value indicates comparable performance on the representative subset rather than a definitive advantage over human experts.

5.5 Robustness test

To address potential biases introduced by our model-dependent synthesis pipeline, we perform a robustness test to evaluate the consistency of the item generation for tasks 2 and 3. Specifically, we leverage the identical representative subsets ($n = 120$) previously established during the human baseline evaluation. These subsets are selected to approximate TMBench's full-set model-performance profiles (Spearman ≥ 0.95). To synthesize new variant items based on the original stems of these 120 items, we use two alternative models with architectures distinct from the original DeepSeek-V3.2: the proprietary GPT-5 mini and the open-weight Qwen3-Next-80B. For task 2, we define the original 75 items generated by DeepSeek-V3.2 as subset 1.1, which serves as the item baseline. Following the same processing pipeline, we instruct GPT-5 mini and Qwen3-Next-80B to generate the corresponding 75 items, designated subsets 1.2 and 1.3. Applying identical protocols, we use subsets 1.2 and 1.3 to conduct downstream evaluation across seven evaluated models, excluding Gemini 3 Flash and Mistral Small 3.1 due to persistent API instability. Similarly, for task 3, we define the original 45 items generated by DeepSeek-V3.2 as subset 2.1. We apply the same generative and evaluation protocols to construct and evaluate subset 2.2 (via GPT-5 mini) and subset 2.3 (via Qwen3-Next-80B).

We evaluate the robustness of the item synthesis pipeline across two primary dimensions. First, to verify semantic equivalence, we extract text embeddings for all six subsets using OpenAI's text-embedding-ada-002. The embeddings demonstrate that the variant subsets (1.2 and 1.3; 2.2 and 2.3) exhibit high semantic alignment with their respective baselines (1.1 and 2.1, respectively). Second, to evaluate the stability of downstream model rankings, we calculate the Spearman rank correlation between the seven evaluated models' performance rankings on the baselines and those on their corresponding variant subsets. The analysis reveals moderate-to-high pairwise ranking consistency, with correlation coefficients ranging from approximately 0.61 to 0.89. Further details of the robustness test are provided in Section 11 in the supplementary materials.

As a complementary robustness check of RS, we assess evaluator agreement on 120 expert-scored model-generated reasoning outputs sampled from the representative subsets, comprising 75 task 2 responses and 45 task 3 responses. The three experts achieve substantial inter-rater agreement, with Fleiss' kappa values of 0.70 and 0.70, respectively. M-Prometheus-7B also shows substantial agreement with the human consensus benchmark, yielding QWK scores of 0.69 and 0.74, respectively. The sampling procedure and score distributions are reported in Section 12 in the supplementary materials.

6 Analysis and discussion

Based on the main results, we explore the models' task-specific capabilities under different conditions, examine temperature sensitivity, and conduct an error analysis.

6.1 LLM performance on domain knowledge QA

By evaluating the models' performance on domain knowledge QA, we identify variations in the models' knowledge memorization and understanding capability along the dimensions of domain, context, and question type.

6.1.1 LLM performance across domains

Table 3 presents a comparative overview of the models' performance across the three domains. This knowledge memorization and understanding capability exhibits a tiered distribution across the three domains, declining from PR through PP to FK. Specifically, PR yields the highest Accuracy and F1 alongside the lowest HR. These results indicate more reliable model performance within policies and regulations. PP occupies an intermediate position between PR and FK across metrics, reflecting moderate complexity in professional practice. FK shows the lowest Accuracy and the highest HR. This suggests that fundamental knowledge constitutes the most challenging domain for the models. These capability disparities may reflect domain knowledge-structure differences. PR is characterized by highly standardized and widely distributed content (Zhao XK et al., 2025). PP centers on the application of

Table 3 Performance of LLMs across domains

Model	Accuracy			F1			HR		
	PR	FK	PP	PR	FK	PP	PR	FK	PP
GPT-5 mini	0.797	0.769	0.798	0.910	0.923	0.927	0.074	0.131	0.101
Gemini 3 Flash	0.919	0.830	0.858	0.961	0.945	0.951	0.039	0.090	0.071
Grok 4.1 Fast	0.804	0.726	0.771	0.909	0.901	0.911	0.084	0.160	0.115
Claude-H-4.5	0.777	0.665	0.687	0.905	0.898	0.893	0.072	0.135	0.132
Mistral Small 3.1	0.762	0.692	0.715	0.891	0.905	0.897	0.091	0.128	0.128
Llama-4-Mav	0.819	0.721	0.760	0.927	0.914	0.916	0.061	0.138	0.108
DeepSeek-V3.2	0.885	0.765	0.792	0.949	0.928	0.930	0.044	0.111	0.078
Qwen3-Next-80B	0.883	0.727	0.757	0.947	0.910	0.912	0.052	0.144	0.117
Kimi-K2-0905	0.877	0.754	0.773	0.953	0.925	0.926	0.031	0.103	0.077
Mean	0.836	0.739	0.768	0.928	0.917	0.918	0.061	0.127	0.103
SD	0.056	0.048	0.049	0.025	0.015	0.018	0.021	0.022	0.023

knowledge within real-world scenarios (Khera et al., 2025). FK consists of vast and fragmented information, making long-tail knowledge difficult to retrieve (Toroghi et al., 2025). Hence, LLMs require retrieval augmentation for fragmented facts and contextual reasoning for professional applications.

PR serves as the primary capability differentiator among LLMs. Regarding SD, PR shows the greatest variability in Accuracy and F1. This indicates inter-model heterogeneity in handling policies and regulations. Conversely, FK exhibits the most homogeneous performance in Accuracy and F1. This indicates more uniform cross-model performance and lower discriminative value. This may reflect the fact that PR involves rigorous logical chains that are solvable but require a specific capability threshold (Servantez et al., 2024), while FK focuses on factual retrieval where long-tail knowledge poses a universal challenge (Huang et al., 2025). Hence, PR should be considered in future benchmarks to distinguish top-tier capabilities.

Open-weight models have higher PR Accuracy but show domain dependence, whereas proprietary models have higher FK and PP Accuracy with greater transfer resilience. Specifically, the five open-weight models have 0.021 higher PR Accuracy than the four proprietary models. However, the proprietary models have Accuracy advantages of 0.019 and 0.016 in PP and FK, respectively. This divergence is underscored by the disparity in their stability during domain transitions: From PR to FK, the open-weight models experience a drop of 0.113 in Accuracy, while the proprietary models remain more stable with a decline of 0.077. This contrast in stability is consistent with greater proprietary-model resilience to domain shifts and may reflect stronger performance on complex reasoning (Chowa et al., 2026). Thus, technology managers should consider these domain-specific profiles when selecting the appropriate model type.

6.1.2 LLM performance across contexts

Table 4 displays a comparative overview of the models' performance across the two contexts. LLMs exhibit higher Accuracy and reduced HR for knowledge memorization and understanding within the U.S. context. However, this contextual pattern may be partially confounded by the models' inherent English-language proficiency. No matter whether this gap stems from linguistic bias or conceptual alignment, their performance in Chinese tasks warrants optimization. This is because LLMs are trained on datasets drawn largely from English-centric high-resource languages and Western perspectives (Binegde and Zhang, 2026). Notably, Qwen3-Next-80B is the only model with marginally higher Accuracy in the Chinese context (0.786 vs. 0.782), but its HR is also higher (0.109 vs. 0.102). Thus, the model does not show a consistent Chinese-context advantage across both metrics.

The Chinese context shows stronger discriminatory power among LLMs. Specifically, it exhibits greater variability in Accuracy, F1, and HR. This suggests that the performance gap between the top-tier and lower-tier models is more pronounced in this context. This dispersion is consistent with disparate levels of model optimization for the specific linguistic and cultural nuances of the Chinese context. Conversely, the U.S. context produces more consistent performance. It reveals a more compressed performance profile. Hence, the Chinese context should be considered in future benchmarks to distinguish top-tier capabilities.

Chinese models have higher Accuracy than their U.S. counterparts in the Chinese context and achieve near-parity in the U.S. context. Specifically, the three Chinese models lead the five U.S. models by 0.031 in Accuracy within the Chinese context and trail by 0.005 in the U.S. context. Furthermore, the Chinese models demonstrate greater cross-contextual stability. When transitioning from the U.S. to Chinese contexts, they experience a decline of 0.018 in Accuracy, whereas the U.S. models have a drop of 0.054. This contrast may reflect the integration of culturally diverse datasets during pretraining (Decoupes et al., 2025). Hence, technology managers should remain context-aware when selecting models for localized applications.

6.1.3 LLM performance across question types

Table 5 provides a comparative overview of the models' performance across the three question types. LLMs show stronger performance on lower-demand knowledge judgments than on tasks requiring integrated inference. This is evidenced by the performance gradient. Specifically, JU yields the highest Accuracy of 0.886, which is close to 0.900. This suggests that when the logical demand remains minimal, retrieved knowledge may be sufficient to support accurate responses. With the increased demand for selection logic, SMC occupies an intermediate position at an Accuracy of 0.859. This reflects a moderate strain on the models' reasoning capacity. Conversely,

Table 4 Performance of LLMs across contexts

Model	Accuracy		F1		HR	
	China	U.S.	China	U.S.	China	U.S.
GPT-5 mini	0.764	0.845	0.925	0.916	0.114	0.078
Gemini 3 Flash	0.860	0.883	0.958	0.934	0.069	0.066
Grok 4.1 Fast	0.745	0.813	0.912	0.895	0.127	0.110
Claude-H-4.5	0.685	0.763	0.907	0.880	0.119	0.102
Mistral Small 3.1	0.693	0.793	0.903	0.893	0.126	0.094
Llama-4-Mav	0.758	0.777	0.926	0.901	0.111	0.088
DeepSeek-V3.2	0.802	0.834	0.942	0.914	0.082	0.075
Qwen3-Next-80B	0.786	0.782	0.931	0.894	0.109	0.102
Kimi-K2-0905	0.791	0.817	0.942	0.910	0.070	0.080
Mean	0.765	0.812	0.927	0.904	0.103	0.088
SD	0.054	0.038	0.018	0.016	0.023	0.015

Table 5 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	0.845	0.631	0.877	0.892	3.894	0.914	0.892	0.938	0.914	0.066	0.201	0.053
Gemini 3 Flash	0.929	0.719	0.935	0.950	4.137	0.951	0.950	0.954	0.951	0.031	0.151	0.033
Grok 4.1 Fast	0.860	0.534	0.875	0.903	3.946	0.915	0.903	0.908	0.915	0.062	0.268	0.051
Claude-H-4.5	0.816	0.501	0.751	0.884	3.859	0.846	0.885	0.918	0.852	0.061	0.225	0.078
Mistral Small 3.1	0.794	0.468	0.915	0.858	3.776	0.923	0.858	0.916	0.924	0.082	0.206	0.067
Llama-4-Mav	0.845	0.537	0.900	0.898	3.929	0.935	0.908	0.923	0.937	0.056	0.228	0.040
DeepSeek-V3.2	0.896	0.608	0.906	0.931	4.034	0.940	0.931	0.937	0.940	0.038	0.178	0.034
Qwen3-Next-80B	0.868	0.567	0.906	0.905	4.028	0.928	0.914	0.926	0.928	0.061	0.219	0.053
Kimi-K2-0905	0.882	0.588	0.907	0.926	4.063	0.940	0.927	0.937	0.940	0.040	0.150	0.036
Mean	0.859	0.573	0.886	0.905	3.963	0.921	0.908	0.929	0.922	0.055	0.203	0.049
SD	0.041	0.075	0.054	0.027	0.113	0.031	0.027	0.014	0.029	0.016	0.038	0.016

MMC yields the lowest Accuracy (0.573) while leading in F1. This contrast re-emphasizes the gap between option-level and question-level performance. While the models appear able to recognize knowledge fragments, they show weaker integration of these fragments in complex decision-making. Furthermore, the gap between the MMC Score (3.963) and the theoretical maximum of 5.000 is consistent with a conservative strategy: LLMs prioritize core correct options to secure partial credit while omitting ambiguous marginal candidates (Rahman et al., 2025). A plausible source of this capability limitation is an inferential lag where logical synthesis fails to keep pace with knowledge retrieval. Consequently, enhancing LLMs in this setting may require moving beyond data augmentation and focusing on the integration of structured reasoning frameworks.

Open-weight models have higher JU Accuracy in binary knowledge judgments, whereas proprietary models have higher MMC Accuracy in multifaceted reasoning. Specifically, the five open-weight models have 0.047 higher JU Accuracy than the four proprietary models. The proprietary models secure a 0.043 margin in MMC Accuracy. This pattern is also reflected in the transition from JU to MMC: The open-weight models experience a 0.353 decline in Accuracy, whereas the proprietary models have a drop of 0.263. These differences may relate to the fact that open-weight models benefit from effective pretraining and fine-tuning processes, which improve zero-shot generalization and downstream performance (Naveed et al., 2025). These strategies may support more stable binary knowledge judgments. Meanwhile, proprietary models may be better positioned for complex reasoning (Guo X et al., 2025). This enables them to navigate complex logical dependencies and filter out competitive distractors. Consequently, technology managers should consider open-weight models for binary knowledge judgments while considering proprietary models for multifaceted reasoning.

6.2 LLM performance on technology readiness assessment

To identify variations in the models' procedural application and analytical reasoning capability, Table 6 provides a comparative overview of their performance in TRA across the three prompting strategies. At the primary-metric level, PA3 achieves higher task 2 Accuracy than PA1 (Accuracy gain = 0.015, Holm-adjusted $p = 0.006$) and PA2 (Accuracy gain = 0.016, Holm-adjusted $p = 0.006$). These results support prompt-sensitive differences in procedural application and analytical reasoning. However, Table 6 also shows that PA3 has the lowest mean RS, indicating an outcome-accuracy gain without a corresponding improvement in reasoning quality. This pattern suggests that explicit CoT examples can help align final TRL labels with task definitions while still imposing reasoning-format constraints. Hence, technology managers should consider few-shot CoT for TRA while monitoring reasoning quality.

The task 2 prompting results reveal an Accuracy–RS

Table 6 Performance of LLMs on TRA

Model	Accuracy			RS			MAE		
	PA1	PA2	PA3	PA1	PA2	PA3	PA1	PA2	PA3
GPT-5 mini	0.744	0.758	0.751	3.755	3.715	3.478	0.343	0.325	0.357
Gemini 3 Flash	0.765	0.783	0.773	4.117	3.913	3.637	0.289	0.269	0.296
Grok 4.1 Fast	0.785	0.778	0.778	4.309	4.298	3.940	0.278	0.289	0.298
Claude-H-4.5	0.796	0.756	0.787	3.717	3.583	3.433	0.249	0.287	0.264
Mistral Small 3.1	0.771	0.778	0.798	3.959	3.951	3.856	0.274	0.269	0.253
Llama-4-Mav	0.760	0.765	0.787	3.957	3.845	3.747	0.316	0.289	0.274
DeepSeek-V3.2	0.780	0.769	0.818	3.751	3.688	3.439	0.269	0.291	0.235
Qwen3-Next-80B	0.747	0.776	0.785	4.262	4.188	3.868	0.321	0.292	0.258
Kimi-K2-0905	0.780	0.755	0.783	3.951	3.917	3.735	0.271	0.296	0.271
Mean	0.770	0.769	0.784	3.975	3.900	3.681	0.290	0.290	0.278
SD	0.017	0.011	0.018	0.218	0.231	0.195	0.030	0.016	0.036

tradeoff rather than a uniform benefit from greater prompt complexity. PA2's Accuracy remains comparable to that of PA1, and its RS declines, indicating that black-box few-shot examples do not translate into improved outcome accuracy and are associated with weaker reasoning quality. Moreover, the improvement brought by few-shot examples is usually unstable (Li HT et al., 2024). The inability of PA2 to produce higher Accuracy than PA1 may stem from its input–output examples, which provide numerical anchors without intermediate reasoning. PA3 then improves Accuracy but further reduces the mean RS from 3.975 in PA1 to 3.681, suggesting that explicit CoT examples support final-label accuracy while the restrictive single-sentence reasoning format constrains reasoning quality. Hence, prompt engineering should balance outcome accuracy with reasoning reliability rather than treating greater prompt complexity as uniformly beneficial in technology transfer.

Few-shot CoT prompting is associated with higher open-weight model Accuracy in procedural application and analytical reasoning. Specifically, the four proprietary models have 0.005 higher Accuracy than the five open-weight models under PA1, whereas the open-weight models secure a 0.022 margin under PA3. This indicates a shift from a proprietary-model advantage in the zero-shot baseline to an open-weight advantage under few-shot CoT prompting. The disparity in prompting sensitivity is further evidenced by the transition from PA1 to PA3: While the proprietary models maintain a relatively stagnant performance baseline, the open-weight models exhibit a 0.027 improvement. Such divergence suggests different characteristics of the two model categories. Proprietary models often demonstrate greater robustness to prompt variations, as long as the essential task information is conveyed (Ting et al., 2025). In contrast, the performance gap for the open-weight models may be reduced through advanced prompt engineering (Khera et al., 2025). Hence, technology managers should consider open-weight models with advanced CoT prompting strategies for complex reasoning tasks when prompt-engineering resources are available.

6.3 LLM performance on technology market forecasting

To identify variations in the models' complex synthesis and predictive inference capability, Table 7 provides a comparative overview of their performance in technology market forecasting across the three prompting strategies. At the primary-metric level, PF3 achieves lower task 3 BS than PF1 (BS reduction = 0.017, Holm-adjusted $p < 0.001$) and PF2 (BS reduction = 0.007, Holm-adjusted $p = 0.038$). PF2 also has lower BS than PF1 (BS reduction = 0.010, Holm-adjusted $p = 0.037$). These results support prompt-sensitive differences in complex synthesis and predictive inference for CoT-based forecasting prompts.

PF2's lower BS may stem from its structured CoT workflow, which asks models to consider both yes and no arguments, aggregate them from a superforecaster perspective, and review calibration-relevant factors such as base rates (Schoenegger et al., 2024). These steps offer a plausible explanation for the BS improvement from PF1 to PF2, but the evidence supports a task-specific prompt effect rather than a general preference for more elaborate prompting. Hence, technology managers should consider integrating dialectical analysis and rigorous self-evaluation into their prompt engineering practices.

The EG strategy can strengthen CoT prompting through strict grounding in the provided background. Specifically, PF3 achieves lower BS than PF2 with a 0.007 decrease after Holm correction; ECE is also lower with a 0.011 decrease, alongside an increase of 0.037 in RS and a decrease of 0.004 in AR. These patterns indicate higher predictive accuracy and higher RS for EG zero-shot CoT prompting in complex synthesis and predictive inference compared with zero-shot CoT, although AR does not increase. Moreover, PF3 exhibits lower dispersion than PF2, as evidenced by reduced SD values across BS, RS, ECE, and AR. This profile supports the value of evidence grounding within the CoT framework, while leaving practical-utility gains task-dependent.

Mechanistically, PF3 narrows the evidence space by requiring all reasoning, entities, and base rates to come from the background text and by flagging ambiguity rather than resolving it through external assumptions. This source-strict design can reduce informational noise and hallucination risk, but its practical value depends on whether the supplied materials are complete and accurate. Hence, technology managers should prioritize complete and accurate source materials when applying EG prompting to complex synthesis and predictive inference tasks.

EG CoT prompting is associated with lower open-weight model BS in complex synthesis and predictive inference. Specifically, while the four proprietary models have 0.003 lower BS

Table 7 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	0.243	0.236	0.233	3.644	3.853	3.907	0.172	0.159	0.154	1.364	1.396	1.322
Gemini 3 Flash	0.254	0.259	0.249	3.814	4.144	4.102	0.189	0.221	0.182	1.360	1.376	1.366
Grok 4.1 Fast	0.259	0.249	0.244	4.683	4.772	4.734	0.208	0.190	0.186	1.409	1.392	1.367
Claude-H-4.5	0.246	0.257	0.233	3.734	3.940	3.859	0.182	0.191	0.157	1.298	1.303	1.370
Mistral Small 3.1	0.256	0.233	0.231	3.997	3.994	4.000	0.197	0.141	0.151	1.364	1.332	1.347
Llama-4-Mav	0.256	0.237	0.232	4.054	4.096	4.213	0.177	0.139	0.127	1.377	1.363	1.386
DeepSeek-V3.2	0.237	0.227	0.232	4.195	4.323	4.512	0.156	0.137	0.153	1.371	1.418	1.392
Qwen3-Next-80B	0.261	0.253	0.237	4.763	4.922	4.952	0.218	0.210	0.181	1.417	1.438	1.407
Kimi-K2-0905	0.259	0.231	0.227	3.341	3.422	3.518	0.202	0.141	0.143	1.366	1.332	1.359
Mean	0.252	0.242	0.235	4.025	4.163	4.200	0.189	0.170	0.159	1.370	1.372	1.368
SD	0.008	0.012	0.007	0.468	0.461	0.456	0.019	0.033	0.020	0.034	0.044	0.025

than the five open-weight models under PF1, the open-weight models have lower BS under PF2 and PF3 by 0.014 and 0.008, respectively. This shift suggests that while the proprietary models hold an initial advantage in the zero-shot baseline, the open-weight models perform more favorably under CoT-based forecasting prompts. This pattern is accompanied by different prompting-sensitivity profiles during stage-wise transitions. From PF1 to PF2, while the proprietary models remain virtually unchanged, the open-weight models exhibit a 0.018 improvement. From PF2 to PF3, the open-weight models record a further 0.004 improvement, compared with a 0.011 gain for their proprietary counterparts. This divergence suggests that the open-weight models are responsive to CoT prompting (Ting et al., 2025). Hence, achieving stronger performance with open-weight models may depend on sustained CoT prompting refinement.

6.4 Temperature variability analysis

To examine the impact of temperature on the models' capabilities, Fig. 3 illustrates the score differences as the temperature increases from 0.0 to 0.7 across aggregate and task-specific levels. These temperature settings are employed to characterize convergent and divergent thinking, respectively. Temperature effects are mixed at the aggregate level. Specifically, the TMscore differences range from −3.3 to 5.9, with five of the nine models showing a positive net change. This indicates that the higher temperature setting is beneficial for some models but not uniformly so. This is because higher temperatures increase randomness and exploration, which can help LLMs discover better solutions (Renze, 2024). Accordingly, the core functions of technology managers may require model-specific temperature management.

Higher-order capabilities exhibit lower volatility between convergent and divergent thinking. In particular, subtask scores exhibit highly disparate sensitivities to the temperature shift. $S_{\text{Knowledge}}$ spans a broad range from −8.6 to 14.5, with four models exhibiting absolute fluctuations above 8 points. S_{Assess} ranges from −4.1 to 9.8, with only one model showing a 9.8-point gain. In contrast, S_{Forecast} displays greater stability, with fluctuations confined to a smaller range between −5.2 and 4.5. Here, $S_{\text{Knowledge}}$ denotes the composite score for task 1 (domain knowledge QA), which serves as an operational measure of knowledge memorization and understanding, S_{Assess} denotes the composite score for task 2 (TRA), which serves as an operational measure of procedural application and analytical reasoning, and S_{Forecast} denotes the composite score for task 3 (technology market forecasting), which serves as an operational measure of complex synthesis and predictive inference. These disparities suggest that the knowledge memorization and understanding capability is highly sensitive to temperature-induced shifts, whereas the complex synthesis and predictive inference capability demonstrates greater stability. This disparity in sensitivity may be related to two factors. On one hand, as LLMs are primarily trained on general-domain corpora, they often lack the specialized expertise required to address domain-knowledge tasks (Huang et al., 2025). This deficiency can increase factual-hallucination risk and temperature sensitivity in knowledge memorization and understanding. On the other hand, the influence of temperature appears partly stabilized when LLMs employ prompting strategies. Specifically, the

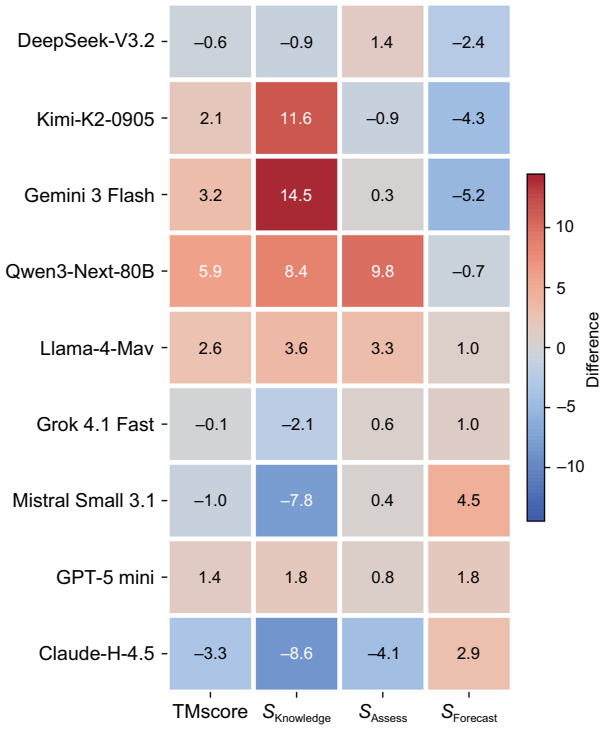


Fig. 3 Comparison of score variations across aggregate and task-specific levels. Each difference is calculated as the score at temperature 0.7 minus that at temperature 0.0

adoption of expert persona framing, CoT reasoning, and structured output specification appears to stabilize performance across varying temperature settings (Ting et al., 2025). These findings re-emphasize the importance of systematic prompting strategies.

Temperature sensitivity is model-specific, with only a modest group-level difference between open-weight and proprietary models. Specifically, the open-weight models yield a higher mean absolute TMscore variation (2.4) than the proprietary models (2.0), while absolute TMscore variations above 1.0 occur in three of the five open-weight models and three of the four proprietary models. These findings support model-specific temperature management rather than a strong model-category generalization.

6.5 Error analysis

To identify high-impact improvement areas, we analyze 5967 model–question error instances derived from frequently failed questions for which at least five of the nine evaluated models fail according to the task-specific criterion. As illustrated in Table 8, these errors are classified into three primary categories and nine subcategories. The distribution and nature of these errors highlight the models' specific cognitive and knowledge-based hurdles in TMBench.

Knowledge base, retrieval, and application deficits constitute the largest category, accounting for 48.5% of the analyzed error instances. First, domain knowledge deficiency accounts for 37.3% of the analyzed error instances. This failure typically

Table 8 Primary error categories among 5967 retained high-frequency error instances

Category	Subcategory	Description	Percentage (%)
Knowledge base, retrieval, and application deficits	Domain knowledge deficiency	The models' internal knowledge base is insufficient, outdated, or misapplied regarding the specific domain. They substitute precise, context-specific facts with generalized or obsolete information. This leads to the generation of plausible but factually incorrect answers that fail to meet specific query requirements.	37.3
	External knowledge hallucination	The models fabricate information or cite non-existent data from outside the provided context. Their reasoning is corrupted by erroneous external injections rather than logical failures. This leads to the misrepresentation of facts or contradictions with specific, authoritative sources.	9.0
	Domain knowledge misapplication	The models incorrectly allow general or real-world knowledge to override the specific domain constraints. They fail to confine their reasoning to the provided rules. They substitute appropriate logic with inapplicable external assumptions. This leads to answers that contradict explicit evidence.	2.2
Contextual processing and attentional impairments	Contextual association failure	The models identify individual facts correctly but fail to synthesize them into a coherent understanding. They miss logical or causal connections between elements. This results in a fragmented analysis rather than a holistic narrative derived from the information provided.	19.5
	Key constraint omission	The models fail to prioritize critical, non-negotiable constraints essential for a correct conclusion. They treat binding rules as indecisive and underweight them relative to circumstantial evidence. This renders the output invalid despite partial logical consistency.	4.9
	Multi-source information conflict	The models fail to correctly resolve conflicts between contradictory evidence from multiple sources. They lack the ability to apply logical principles, such as source reliability or temporal decay. They fail to adjudicate competing claims effectively. This leads to an inappropriate synthesis of conflicting information.	2.6
Cognitive reasoning and inference fallacies	Inferential leap	The models exhibit unjustified logical leaps from premises to conclusions and often rely on flawed or simplistic inferential steps. They fail to support inferences with evidence. This results in plausible surface-level answers that violate the specific logical or contextual constraints of complex tasks.	9.4
	Evidence synthesis fallacy	The models misjudge or incorrectly combine disparate pieces of evidence and fail to integrate them into a coherent conclusion. While processing individual data points correctly, they fail to synthesize them into a logically consistent whole that aligns with the task requirements.	7.7
	Logical inconsistency	The reasoning process contains internal contradictions or applies flawed logical rules (e.g., misapplying linear extrapolation). The output contradicts the stated premises or ignores constraints due to faulty reasoning patterns rather than a lack of factual knowledge.	7.4

occurs when the models lack the necessary domain-specific terminology or facts, such as relying on outdated 2015 legal provisions instead of 2021 amendments. Second, external knowledge hallucination (9.0%) is characterized by the introduction of factually incorrect information not present in the input, such as an imaginary government approval date. Third, domain knowledge misapplication (2.2%) reflects the incorrect application of valid knowledge to an inappropriate context. This is exemplified by a failure to correctly identify TRL 5 because the models do not recognize that specific pressure operational modes simulate the relevant environment required by technical standards. Collectively, these failures indicate a knowledge-boundary conflict (Longpre et al., 2021). This illustrates the inherent difficulty in upholding rigorous, context-specific constraints.

Contextual processing and attentional impairments account for 27.0% of the analyzed error instances. This points to difficulties in correctly parsing context and maintaining attention to explicit constraints. First, contextual association failure (19.5%) leads this category. In these instances, the models fail to link scattered contextual clues, such as the relationship between high debt levels and the valuation of capital-intensive firms. Second, key constraint omission (4.9%) represents instances of models directly overlooking explicit conditions. For example, the models fail to assign a TRL 5 rating by deeming preliminary environmental testing insufficient. In these cases, the models overlook that a brassboard-level prototype tested in such a manner explicitly satisfies the technical definition. Third, multi-source information conflict (2.6%) highlights the models' failure to resolve contradictions. This is exemplified when the models incorrectly apply a linear time-decay factor to conflicting signals regarding an executive order, instead of recognizing the exponential decay necessitated by a lack of progress in critical windows. These errors suggest a breakdown in relational reasoning.

Cognitive reasoning and inference fallacies represent 24.5% of the analyzed error instances. This points to limitations in the logical derivation process. Specifically, the models often fail to construct a valid and complete reasoning chain. First, inferential leaps (9.4%) involve unsupported logical jumps, such as incorrectly inferring that the strength, weakness, opportunity, and threat (SWOT) factors should be grouped by their source rather than by their relationship to external opportunities or threats. Second, evidence synthesis fallacy (7.7%) occurs when the models incorrectly weight or integrate multiple pieces of evidence. This failure often involves overemphasizing the term "proof-of-concept" while failing to synthesize evidence of component integration and validation in a laboratory environment during TRA. Third, logical inconsistency (7.4%) manifests as internal contradictions or conclusions that conflict with premises. A typical example involves conflating the stage for preventative patent research with the stage of actual infringement risk realization. Such failures can arise when LLMs can perform each reasoning step individually but fail to integrate these steps (Song et al., 2026). These findings suggest that even with accurate information, the models' internal reasoning patterns may fail to align with the specific logical or contextual constraints of complex tasks (Altermatt et al., 2025).

Building on these findings, we propose three recommendations to enhance domain-specific LLM capabilities. First, future benchmarking should prioritize specialized datasets incorporating high-frequency errors and deceptive cases, with coverage weighted toward knowledge-boundary, contextual, and reasoning failures according to their empirical frequency. Second, developers should employ targeted optimization, such as integrating RAG to bridge knowledge gaps while using context-focused supervision and reasoning checks to address attentional and inferential failures. Finally, technology managers should establish rigorous verification protocols for high-stakes applications to ensure evidence-grounded outputs and safeguard decision-making reliability.

7 Conclusions

We have introduced TMBench, a hierarchical benchmark designed to evaluate LLMs on the core cognitive capabilities expected of technology managers. TMBench comprises 3474 questions structured across three tasks and five question types. It demonstrates substantial human–LLM agreement, and robustness checks show moderate-to-high ranking consistency against potential biases in model-dependent synthesis. We have benchmarked nine open-weight and proprietary models within practical industry-specific scenarios. The models have been evaluated by varying prompt configurations and temperature settings to investigate their capabilities and characterize their error patterns. To provide a comparative standard, human baselines are established by evaluating human experts on a distributionally representative 370-item subset, enabling a performance comparison between LLMs and humans.

Our findings reveal the following:

1. LLMs exhibit differences in capabilities, with no single model ranking first across all three capability levels. On the two higher-order tasks, LLMs have more favorable point estimates, but neither LLM–human difference is significant.

2. Open-weight models show more favorable performance under CoT-based prompting in higher-order capability tasks, whereas proprietary models show more stable performance across domain and prompt transitions.

3. Under strict task constraints, the limitations of LLMs in complex tasks are associated with knowledge base, retrieval, and application deficits, contextual processing and attentional impairments, as well as cognitive reasoning and inference fallacies.

Our evaluation highlights the potential of LLMs to support analytical tasks in technology transfer and the continued need for targeted optimization in this domain. As a foundational effort to bridge the evaluation gap, TMBench provides a reference resource for researchers, developers, and practitioners aiming to support the responsible use of LLMs as analytical tools in technology management. Moreover, our benchmarking pipeline, evaluation methodology, and structured prompting frameworks offer a scalable reference for future domain-specific assessments. The limitations of this study stem from its restricted scenario coverage, a bias toward explicit knowledge over tacit practical experience, a reliance on static evaluation formats, and inherent methodological constraints, including residual biases in untranslated QA and LLM-synthesized data. To address these gaps, future work will expand the benchmark

toward multi-agent and embodied AI evaluation paradigms to capture the interactive and execution-oriented dimensions of technology transfer.

Supplementary information

The online version contains supplementary materials available at https://doi.org/10.1631/ENG.ITEE.2026.0168.

Acknowledgments

This work was supported by the National Social Science Fund of China (No. 22BTQ017).

Author contributions

Zhipeng ZHU designed the research, conducted the investigation, and drafted the paper. Guo LI provided advice on the structure and content of the paper. Mark Xuefang ZHU supervised the research, acquired funding, and reviewed the paper. Shanshan KONG curated the data. Zhenyu WANG provided technical support for the research. Wenting XIONG conducted the literature survey.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The source code and data used in this study are openly available on GitHub at https://github.com/zzpnj199535-sketch/r-abe9bf10128c.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Alibaba Group, 2025. Qwen3-Next-80B: instruction-tuned models for RAG workflows. https://qwen.ai/blog/qwen3-next-ultimate-efficiency/ [Accessed on Jan. 20, 2026].

Altermatt FR, Neyem A, Sumonte NI, et al., 2025. Evaluating the performance of large language models on the CONACEM anesthesiology certification exam: a comparison with human participants. *Appl Sci*, 15(11):6245. https://doi.org/10.3390/app15116245

Anthropic, 2026. Introducing Claude Haiku 4.5: high-efficiency iterations of the Claude family. https://www.anthropic.com/news/claude-haiku-4-5 [Accessed on Jan. 20, 2026].

Association of University Technology Managers, 2024. Technology Transfer Practice Manual. https://autm.net/surveys-and-tools/tech-transferpractices-manual [Accessed on Jan. 20, 2026].

Badenes-Olmedo C, Eyzaguirre-Barreda P, Chu-Artzt N, et al., 2026. ESQAD: a curriculum-aligned dataset for question answer generation in Spanish. *Knowl-Based Syst*, 335:115255. https://doi.org/10.1016/j.knosys.2025.115255

Bastani H, Karger E, Chen YH, et al., 2025. ForecastBench: a dynamic benchmark of AI forecasting capabilities. Proc 13th Int Conf on Learning Representations.

BenYoash N, Brief M, Ovadia O, et al., 2025. SECQUE: a benchmark for evaluating real-world financial analysis capabilities. Proc 4th Workshop on Generation, Evaluation and Metrics, p.212-230.

Betancourt J, Coral A, Fraga A, et al., 2025. Intelligent virtual assistant for calculating technology readiness levels using large language models (LLM). *IEEE Access*, 13:137709-137723. https://doi.org/10.1109/access.2025.3595699

Binegde GN, Zhang HP, 2026. Exploring cultural commonsense in multilingual large language models: a survey. *Inform Syst*, 138:102649. https://doi.org/10.1016/j.is.2025.102649

Busch K, Leopold H, 2024. Towards a benchmark for large language models for business process management tasks. https://doi.org/10.48550/arXiv.2410.03255

Center for AI Safety, Scale AI, HLE Contributors Consortium, 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139-1146. https://doi.org/10.1038/s41586-025-09962-4

Cetindamar D, Phaal R, Probert DR, 2016. Technology management as a profession and the challenges ahead. *J Eng Technol Manag*, 41:1-13. https://doi.org/10.1016/j.jengtecman.2016.05.001

Chadli K, Botterweck G, Saber T, 2026. PEARL: performance and energy aware routing for LLMs. *Fut Gener Comput Syst*, 176:108218. https://doi.org/10.1016/j.future.2025.108218

Chang CC, Chen CY, Lee HS, et al., 2024. Benchmarking cognitive domains for LLMs: insights from Taiwanese Hakka culture. Proc 27th Conf of the Oriental COCOSDA Int Committee, p.1-6. https://doi.org/10.1109/O-COCOSDA64382.2024.10800594

Chang YP, Wang X, Wang JD, et al., 2024. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*, 15(3):39. https://doi.org/10.1145/3641289

China Association of Science and Technology Evaluation and Management of Scientific and Technical Achievement (CASTEM), National Center for Science and Technology Evaluation (NCSTE), 2025. Advanced Tutorial Technical Manager. Science and Technology of China Press, Beijing, China (in Chinese).

Chiu YY, Jiang LW, Lin BY, et al., 2025. CulturalBench: a robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. Proc 63rd Annual Meeting of the Association for Computational Linguistics, p.25663-25701. https://doi.org/10.18653/v1/2025.acl-long.1247

Chowa SS, Alvi R, Rahman SS, et al., 2026. From language to action: a review of large language models as autonomous agents and tool users. *Artif Intell Rev*, 59(2):71. https://doi.org/10.1007/s10462-025-11471-9

COGR, 2021. The evolution of technology transfer. https://www.cogr.edu/sites/default/files/COGR%20Tech%20Transfer10621.pdf [Accessed on Jan. 20, 2026].

Dai H, Teehan R, Ren MY, 2025. Are LLMs prescient? A continuous evaluation using daily news as the oracle. https://doi.org/10.48550/arXiv.2411.08324

Decoupes R, Interdonato R, Roche M, et al., 2025. Evaluation of geographical distortions in language models. *Mach Learn*, 114(12):1-32. https://doi.org/10.1007/s10994-025-06916-9

DeepSeek AI, 2025. DeepSeek-V3.2: high-compute optimization and reasoning performance. https://api-docs.deepseek.com/news/news251201/ [Accessed on Jan. 20, 2026].

Deroy A, Ghosh K, Ghosh S, 2026. Investigating legal question generation using large language models. *Artif Intell Law*, 34:775-813. https://doi.org/10.1007/s10506-025-09452-y

DeVore S, 2025. Exploring the ability of LLMs to classify written proficiency levels. *Comput Speech Lang*, 90:101745. https://doi.org/10.1016/j.csl.2024.101745

Fan ZY, He JL, Li XQ, et al., 2025. CAMIEval: enhancing NLG evaluation through multidimensional comparative instruction-following analysis. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.8708-8733. https://doi.org/10.18653/v1/2025.naacl-long.438

Gao MQ, Hu XY, Yin XJ, et al., 2025. LLM-based NLG evaluation: current status and challenges. *Comput Ling*, 51(2):661-687. https://doi.org/10.1162/coli_a_00561

Genç O, 2026. Is conversational AI ready for engineering licensure? A ChatGPT performance benchmark and prompting strategy evaluation. *J Asian Archit Build Eng*, p.1-11. https://doi.org/10.1080/13467581.2025.2574556

Georgiev P, Lei VL, Burnell R, et al., 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. https://arXiv.org/abs/2403.05530.

Gignac GE, Ilić D, 2025. Psychometrically derived 60-question benchmarks: substantial efficiencies and the possibility of human-AI comparisons. *Intelligence*, 110:101922. https://doi.org/10.1016/j.intell.2025.101922

Google DeepMind, 2025. Gemini 3 Flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/ [Accessed on Jan. 20, 2026].

Gorschek T, Garre P, Larsson S, et al., 2006. A model for technology transfer in practice. *IEEE Softw*, 23(6):88-95. https://doi.org/10.1109/MS.2006.147

Guo SK, Shariatmadari AH, Xiong GZ, et al., 2025. IDEABench: benchmarking large language models for research idea generation. Proc 31st ACM SIGKDD Conf on Knowledge Discovery and Data Mining V.2, p.5888-5899. https://doi.org/10.1145/3711896.3737419

Guo X, Xia HT, Liu ZW, et al., 2025. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.6258-6292. https://doi.org/10.18653/v1/2025.naacl-long.318

Halawi D, Zhang F, Chen YH, et al., 2024. Approaching human-level forecasting with language models. Proc 38th Int Conf on Neural Information Processing Systems, p.50426-50468. https://doi.org/10.52202/079017-1598

Harrison W, 2006. Technology transfer and the tech broker. *IEEE Softw*, 23(5):5-7. https://doi.org/10.1109/MS.2006.142

Huang L, Yu WJ, Ma WT, et al., 2025. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inform Syst*, 43(2):42. https://doi.org/10.1145/3703155

Hunan University, National Center for Science and Technology Evaluation (NCSTE), 2025. Elementary Tutorial Technical Manager. Science and Technology of China Press, Beijing, China (in Chinese).

Jiang LK, Zhang CQ, Scherz PA, et al., 2025. Can large language models generate high-quality patent claims? Findings of the Association for Computational Linguistics, p.1272-1287. https://doi.org/10.18653/v1/2025.findings-naacl.70

Jung J, Lu M, Benker SC, et al., 2025. How model size, temperature, and prompt style affect LLM-human assessment score alignment. Proc Artificial Intelligence in Measurement and Education Conf, p.265-273.

Kampelopoulos D, Tsanousa A, Vrochidis S, et al., 2025. A review of LLMs and their applications in the architecture, engineering and construction industry. *Artif Intell Rev*, 58(8):250. https://doi.org/10.1007/s10462-025-11241-7

Khera B, Alamian R, Scherz PA, et al., 2025. Can large language models understand as well as apply patent regulations to pass a hands-on patent attorney test? https://doi.org/10.48550/arXiv.2507.10576

Kim J, Podlasek A, Shidara K, et al., 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Sci Rep*, 15(1):39426. https://doi.org/10.1038/s41598-025-22940-0

Kim JS, Kim BK, 2022. Examining different technology transfer capabilities and their counterpart works from two different positions. *Technol Soc*, 68:101856. https://doi.org/10.1016/j.techsoc.2021.101856

Krathwohl DR, 2002. A revision of Bloom's taxonomy: an overview. *Theory Pract*, 41(4):212-218. https://doi.org/10.1207/s15430421tip4104_2

Lane JP, 1999. Understanding technology transfer. *Assist Technol*, 11(1):5-19. https://doi.org/10.1080/10400435.1999.10131981

Li HT, Chen Y, Ai QY, et al., 2024. LexEval: a comprehensive Chinese legal benchmark for evaluating large language models. Proc 38th Int Conf on Neural Information Processing Systems, p.25061-25094. https://doi.org/10.52202/079017-0790

Li JW, Gao YZ, Yang Y, et al., 2025. Fundamental capabilities and applications of large language models: a survey. *ACM Comput Surv*, 58(2):38. https://doi.org/10.1145/3735632

Li KY, Jiang MH, Fu DY, et al., 2025. DatasetResearch: benchmarking agent systems for demand-driven dataset discovery. https://doi.org/10.48550/arXiv.2508.06960

Li SW, Jiang J, Yang XD, 2025. Preliminary assessment of large language models' performance in answering questions on developmental dysplasia of the hip. *J Child Orthop*, 19(3):207-212. https://doi.org/10.1177/18632521251331772

Li YJ, Sun Y, 2026. A fine-grained evaluation framework for language models: combining pointwise grading and pairwise comparison. *Inform Process Manag*, 63(1):104270. https://doi.org/10.1016/j.ipm.2025.104270

Liang C, Huang ZQ, Wang HF, et al., 2026. AECBench: a hierarchical benchmark for knowledge evaluation of large language models in the AEC field. *Adv Eng Inform*, 71:104314. https://doi.org/10.1016/j.aei.2026.104314

Lin CB, Liang HY, Su TY, et al., 2022. Multi-criteria decision-making approach for analyzing competency model of technology managers. *Math Probl Eng*, 2022(1):1909851. https://doi.org/10.1155/2022/1909851

Liu AX, Feng B, Xue B, et al., 2025. DeepSeek-v3 technical report. https://doi.org/10.48550/arXiv.2412.19437

Liu QY, Zou T, Xu CH, et al., 2025. Shanghai Service Manual about Technology Commercialization. Shanghai Jiao Tong University Press, Shanghai, China (in Chinese).

Longpre S, Perisetla K, Chen A, et al., 2021. Entity-based knowledge conflicts in question answering. Proc Conf on Empirical Methods in Natural Language Processing, p.7052-7063. https://doi.org/10.18653/v1/2021.emnlp-main.565

MacMillan IC, McGrath RG, 2004. Nine new roles for technology managers. *Res-Technol Manag*, 47(3):16-26. https://doi.org/10.1080/08956308.2004.11671626

Manifold, 2026. Manifold Markets: The World's Largest Prediction Market Platform. https://manifold.markets/ [Accessed on Jan. 20, 2026].

Matrinex Academy of International Innovation & Strategy (MAIIS), National Center for Science and Technology Evaluation (NCSTE), 2025. Intermediate Tutorial Technical Manager. Science and Technology of China Press, Beijing, China (in Chinese).

Meta AI, 2025. Llama 4 Maverick: sparse mixture-of-experts (MoE) architectures. https://ai.meta.com/blog/llama-4-multimodal-intelligence/ [Accessed on Jan. 20, 2026].

Miao ZY, Sun QY, Wang JY, et al., 2025. Easy Dataset: a unified and extensible framework for synthesizing LLM fine-tuning data from unstructured documents. Proc Conf on Empirical Methods in Natural Language Processing: System Demonstrations, p.960-968. https://doi.org/10.18653/v1/2025.emnlp-demos.75

Mistral AI, 2025. Mistral Small 3.1: latency-optimized 24B parameters model. https://mistral.ai/news/mistral-small-3-1/ [Accessed on Jan. 20, 2026].

Moonshot AI, 2025. Kimi-K2-0905: scaling 1T-parameter MoE architectures. https://www.moonshot.cn/news/kimi-k2-announcement/ [Accessed on Jan. 20, 2026].

Morrissey MT, Almonacid S, 2005. Rethinking technology transfer. *J Food Eng*, 67(1-2):135-145. https://doi.org/10.1016/j.jfoodeng.2004.05.057

NASA, 2020. Technology readiness assessment best practices guide. NASA Special Publication, NASA/SP-20205003605. https://ntrs.nasa.gov/citations/20205003605 [Accessed on Jan. 20, 2026].

NASA, 2026. TechPort 101. https://techport.nasa.gov/help/TechPort101 [Accessed on Jan. 20, 2026].

Naveed H, Khan AU, Qiu S, et al., 2025. A comprehensive overview of large language models. *ACM Trans Intell Syst Technol*, 16(5):106. https://doi.org/10.1145/3744746

Ni YQ, Huang JS, Song LL, 2024. Guidelines for Enhancing the Capacity of National Technology Transfer Professionals. University of Science and Technology of China Press, Hefei, China (in Chinese).

Nie Y, Yan BW, Guo TY, et al., 2025. CFinBench: a comprehensive Chinese financial benchmark for large language models. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.876-891. https://doi.org/10.18653/v1/2025.naacl-long.40

OpenAI, 2026. GPT-5-mini technical report. https://openai.com/index/introducing-gpt-5-for-developers/ [Accessed on Jan. 20, 2026].

Park J, Mendes E, Stanovsky G, et al., 2026. Look before you leap: estimating LLM benchmark scores from descriptions. https://arxiv.org/abs/2509.20645v2

Pei DS, Wu YW, He JH, et al., 2025. Methodology and benchmark for automated driving theory test of large language models. *IEEE Trans Intell Transp Syst*, 26(10):17013-17026. https://doi.org/10.1109/TITS.2025.3571213

Peters U, Chin-Yee B, 2025. Generalization bias in large language model summarization of scientific research. *R Soc Open Sci*, 12(4):241776. https://doi.org/10.1098/rsos.241776

Pham DK, Vo BQ, 2024. Towards reliable medical question answering: techniques and challenges in mitigating hallucinations in language models. https://doi.org/10.48550/arXiv.2408.13808

Pombal J, Yoon D, Fernandes P, et al., 2025. M-Prometheus: a suite of open multilingual LLM judges. https://doi.org/10.48550/arXiv.2504.04953

Rahman ABMA, Anwar S, Usman M, et al., 2025. DefAn: definitive answer dataset for LLM hallucination evaluation. *Information*, 16(11):937. https://doi.org/10.3390/info16110937

Reese ML, Smirnova A, 2024. Comparing ChatGPT and humans on world knowledge and common-sense reasoning tasks: a case study of the Japanese Winograd schema challenge. Extended Abstracts of the CHI Conf on Human Factors in Computing Systems, p.1-9. https://doi.org/10.1145/3613905.3650975

Renze M, 2024. The effect of sampling temperature on problem solving in large language models. Findings of the Association for Computational Linguistics, p.7346-7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432

Rifkin KI, Fineman M, Ruhnke CH, 1999. The human side: developing technical managers—first you need a competency model. *Res-Technol Manag*, 42(2):53-57. https://doi.org/10.1080/08956308.1999.11671273

Rogers EM, Takegami S, Yin J, 2001. Lessons learned about technology transfer. *Technovation*, 21(4):253-261. https://doi.org/10.1016/S0166-4972(00)00039-9

Saxena A, Aggarwal D, Badathala N, et al., 2025. A framework to synthetically generate fine-grained hallucinated data. *Lang Resour Eval*, 59(4):4223-4252. https://doi.org/10.1007/s10579-025-09864-x

Schaufeld J, 2015. The commercialization model: the value of a process. In: Commercializing Innovation: Turning Technology Breakthroughs into Products. Apress, Berkeley, CA, p.17-30. https://doi.org/10.1007/978-1-4302-6353-1_2

Schoenegger P, Tuminauskaite I, Park PS, 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Sci Adv*, 10(45):eadp1528. https://doi.org/10.1126/sciadv.adp1528

Servantez S, Barrow J, Hammond K, et al., 2024. Chain of logic: rule-based reasoning with large language models. Findings of the Association for Computational Linguistics, p.2721-2733. https://doi.org/10.18653/v1/2024.findings-acl.159

Sheft J, 2008. Technology transfer and idea commercialization. *Nat Biotechnol*, 26(6):711-712. https://doi.org/10.1038/nbt0608-711

Song PY, Han PR, Goodman N, 2026. Large language model reasoning failures. https://doi.org/10.48550/arXiv.2602.06176

Srivastava A, Rastogi A, Rao A, et al., 2023. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/arXiv.2206.04615

Sun LT, Han Y, Zhao ZH, et al., 2024. SciEval: a multi-level large language model evaluation benchmark for scientific research. Proc 38th AAAI Conf on Artificial Intelligence, p.19053-19061. https://doi.org/10.1609/aaai.v38i17.29872

Tak R, Pele DT, 2025. LLM-driven stock prediction: capturing market trends with LLaMA. Proc Int Conf on Busin Excell, 19:529-540. https://doi.org/10.2478/picbe-2025-0043

Tao YC, Xu HT, 2026. TheoremForge: scaling up formal data synthesis with low-budget agentic workflow. https://doi.org/10.48550/arXiv.2601.17332

Ting YS, Nguyen TD, Ghosal T, et al., 2025. AstroMLab 1: who wins astronomy jeopardy!? *Astron Comput*, 51:100893. https://doi.org/10.1016/j.ascom.2024.100893

Tofeeq K, Naseer A, Wali A, 2025. Large language models in healthcare: a systematic evaluation on medical Q/A datasets. *Health Inform Sci Syst*, 14(1):2. https://doi.org/10.1007/s13755-025-00397-9

Toroghi A, Guo W, Sanner S, 2025. CoLoTa: a dataset for entity-based commonsense reasoning over long-tail knowledge. Proc 48th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.3444-3454. https://doi.org/10.1145/3726302.3730291

Vu T, Iyyer M, Wang XZ, et al., 2024. FreshLLMs: refreshing large language models with search engine augmentation. Findings of the Association for Computational Linguistics, p.13697-13720. https://doi.org/10.18653/v1/2024.findings-acl.813

Wang MS, Li ZQ, Zhao XG, et al., 2026. Eliminate-then-select: a human-centric reasoning framework for educational question answering with LLMs. *Inform Process Manag*, 63(2):104422. https://doi.org/10.1016/j.ipm.2025.104422

Wang QY, Chen GH, Wang HB, et al., 2025. IPBench: benchmarking the knowledge of large language models in intellectual property. https://doi.org/10.48550/arXiv.2504.15524

Wang RQ, Guo JY, Gao CY, et al., 2025. Can LLMs replace human evaluators? An empirical study of LLM-as-a-Judge in software engineering. *Proc ACM Softw Eng*, 2(ISSTA):1955-1977. https://doi.org/10.1145/3728963

Wang XZ, Wei J, Schuurmans D, et al., 2023. Self-consistency improves chain of thought reasoning in language models. https://doi.org/10.48550/arXiv.2203.11171

Wei KL, Paskov P, Dev S, et al., 2025. Recommendations and reporting checklist for rigorous & transparent human baselines in model evaluations. https://doi.org/10.48550/arXiv.2506.13776

xAI, 2025. Grok 4.1 Fast and Agent Tools API. https://x.ai/news/grok-4-1-fast [Accessed on Jan. 20, 2026].

Xu K, Lian ZY, 2026. WaferSAGE: large language model-powered wafer defect analysis via synthetic data generation and rubric-guided reinforcement learning. https://doi.org/10.48550/arXiv.2604.27629

Xu LQ, Li QW, Peng TS, et al., 2025. Can large language models be good language teachers? Proc Conf on Empirical Methods in Natural Language Processing, p.23957-23971. https://doi.org/10.18653/v1/2025.emnlp-main.1222

Yang QC, Mahns S, Li SD, et al., 2025. LLM-as-a-Prophet: understanding predictive intelligence with Prophet Arena. https://doi.org/10.48550/arXiv.2510.17638

Yao SY, Yu D, Zhao J, et al., 2023. Tree of thoughts: deliberate problem solving with large language models. Proc 37th Int Conf on Neural Information Processing Systems, p.11809-11822.

Yu JF, Wang XZ, Tu SQ, et al., 2024. KoLA: carefully benchmarking world knowledge of large language models. https://doi.org/10.48550/arXiv.2306.09296

Zhang S, Huang QC, Zhang Q, et al., 2025. ElecBench: a large language model benchmark in electric power domain. *Eng Appl Artif Intell*, 162:112310. https://doi.org/10.1016/j.engappai.2025.112310

Zhao XK, Huang H, Yang T, et al., 2025. Urban planning in the age of large language models: assessing OpenAI o1's performance and capabilities across 556 tasks. *Comput Environ Urban Syst*, 121:102332. https://doi.org/10.1016/j.compenvurbsys.2025.102332

Zhao ZH, Huang ZX, Li JR, et al., 2025. SUPERChem: a multimodal reasoning benchmark in chemistry. https://doi.org/10.48550/arXiv.2512.01274

Zheng LM, Chiang WL, Sheng Y, et al., 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Proc 37th Int Conf on Neural Information Processing Systems, p.46595-46623.

Zografos G, Moussiades L, 2025. Beyond the benchmark: a customizable platform for real-time, preference-driven LLM evaluation. *Electronics*, 14(13):2577. https://doi.org/10.3390/electronics14132577