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

ENGINEERING Inform Technol Electron Eng

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

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

Corresponding author: Mark Xuefang ZHU
E-mail: xfzhu@nju.edu.cn
ORCID: https://orcid.org/0000-0002-8244-5999

Motivation

Professional requirement

Technology managers need:
Cross-domain knowledge
Technology readiness judgment
Market forecasting

+

Evaluation gap

Existing benchmarks remain fragmented, limiting a unified assessment of these capabilities.

Research question

How do large language models (LLMs) perform across the three cognitive levels, and how do their strengths and weaknesses vary by tasks?

ENGINEERING Inform Tech of Electron Eng

Main idea

- TMBench adapts Bloom's taxonomy to translate the cognitive requirements of technology managers into three measurable tasks.
- TMBench evaluates cognitive prerequisites for technology transfer.

Cognitive level	Benchmark task
Knowledge memorization and understanding	Domain knowledge question answering (QA)
Procedural application and analytical reasoning	Technology readiness assessment
Complex synthesis and predictive inference	Technology market forecasting

ENGINEERING Inform Technol Electron Eco

Evaluation framework

(Tasks, inputs, and outputs)

Task	Input	Main output	Item
Domain knowledge QA	Professional questions and answer options	Answer selection	2586
Technology readiness assessment	Technical project descriptions	Technology readiness level (TRL) judgment and rationale	554
Technology market forecasting	Question, time-bounded background, resolution Criteria, and retrieval date	Probability (0–1) and rationale	334

Three parallel evaluation branches, with 3474 questions in total.

ENGINEERING Inform Technol Electron Eng

Evaluation framework

(Metrics and scoring)

Task input → Model output → Metrics → Task score → TMscore

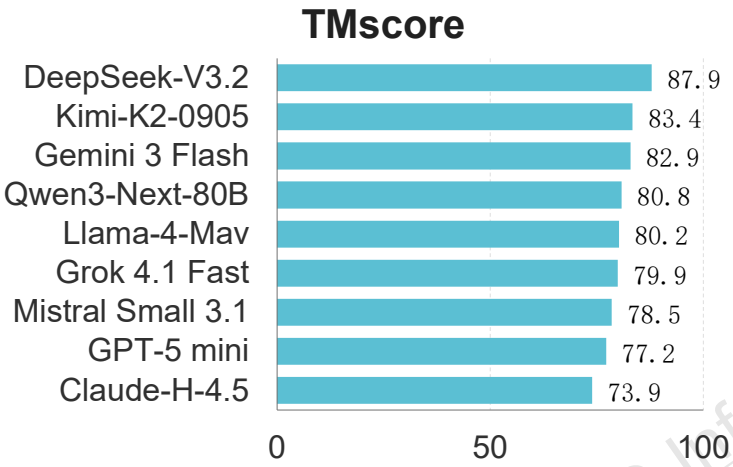
Task	Prompt configuration	Core (50%)	Supporting metrics (50%)
Domain knowledge QA	CoT	Accuracy	F1: 29.6% Score: 20.4%
Technology readiness assessment	PA1: Zero-shot PA2: Few-shot PA3: Few-shot CoT	Accuracy	RS: 35.1% MAE: 14.9%
Technology market forecasting	PF1: Zero-shot PF2: Zero-shot CoT PF3: EG zero-shot CoT	BS	AR: 14.6% ECE: 19.8% RS: 15.6%

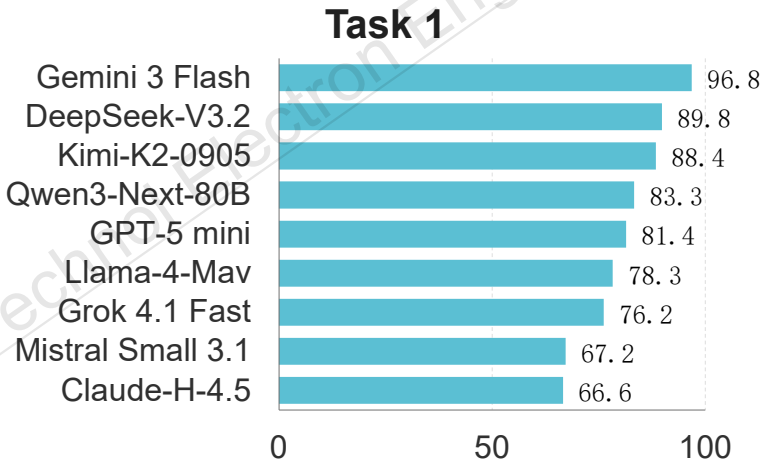
F1: option-level F1 score. Score: mean points per question. RS: reasoning score. BS: Brier score.
MAE: mean absolute error. AR: average return. ECE: expected calibration error. CoT: chain-of-thought.
EG: evidence-grounded. PF3 uses only the supplied background as evidence

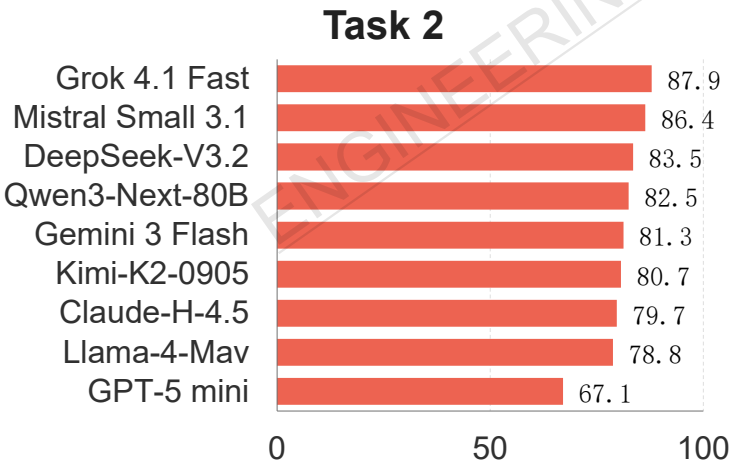
ENGINEERING Inform Technol Electron Eng

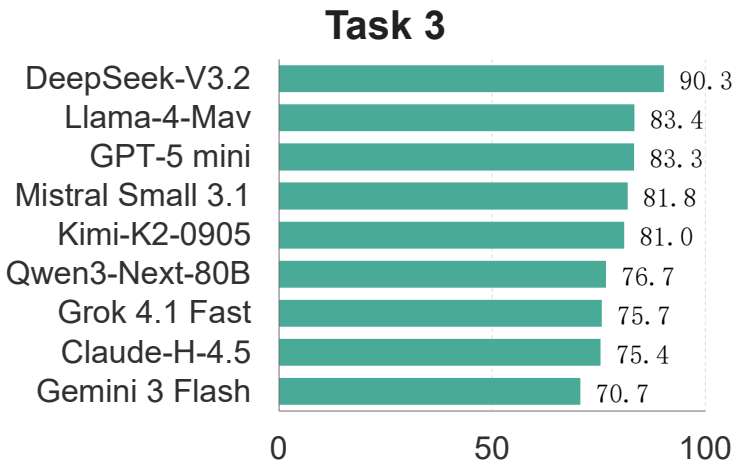
Major results

(Overall results: composite scores, temperature = 0.7)









ENGINEERING Inform Technol Electron Eng

Major results

(LLM performance on three tasks: Task-specific results)

Task	Mean result	Main finding
Domain knowledge QA	**Accuracy** True/False: 88.6% Single-select: 85.9% Multi-select: 57.3%	Selecting the complete answer set remains a major difficulty in multi-select QA.
Technology readiness assessment	**PA1 → PA3** Accuracy: 77.0% → 78.4% RS: 3.975 → 3.681	PA3 has higher mean accuracy but a lower mean RS than PA1.
Technology market forecasting	**PF1/PF2/PF3** BS: 0.252/0.242/0.235 AR: 1.370/1.372/1.368	PF3 has the lowest mean BS, without higher mean AR.

Means across nine models.

ENGINEERING Inform Technol Electron Eng

Major results

(Comparison with human baselines)

Point estimates favor LLMs on both higher-order tasks, but all three comparisons are nonsignificant after Holm correction.

Task (metric)	Item	LLM mean	Human mean
Domain knowledge QA (Accuracy)	250	77.7%	80.8%
Technology readiness assessment (Accuracy)	75	76.9%	72.5%
Technology market forecasting (BS)	45	0.250	0.259

Models and technology managers were compared on a 370-item subset

ENGINEERING Inform Technol Electron Eng

Major results

(Error analysis)

Knowledge-related deficits form the largest error category

Primary error category	Primary share	Error subcategory	Subcategory share
Knowledge base, retrieval, and application deficits	48.5%	Domain knowledge deficiency	37.3%
		External knowledge hallucination	9.0%
		Domain knowledge misapplication	2.2%
Contextual processing and attentional impairments	27.0%	Contextual association failure	19.5%
		Key constraint omission	4.9%
		Multi-source information conflict	2.6%
Cognitive reasoning and inference fallacies	24.5%	Inferential leap	9.4%
		Evidence synthesis fallacy	7.7%
		Logical inconsistency	7.4%

ENGINEERING Inform Technol Electron Eng

Conclusions

Task-dependent performance

Leading models vary across TMBench's three tasks.
Point estimates favor LLMs on both higher-order tasks, but all three LLM–human comparisons are nonsignificant after Holm correction.

+

Prompt trade-offs and error patterns

- Across prompts, better mean outcome metrics do not always coincide with higher mean RS or AR.
- Knowledge-related deficits form the largest error category.

Model selection and targeted optimization

The findings support task-specific model selection and joint assessment of outcomes and judged rationale quality. Error profiles identify targets for improving knowledge use, context handling, and reasoning.

ENGINEERING Inform Technol Electron Eng