面向技术经理人核心认知能力的大语言模型评估：一种分层基准

朱志鹏，李果，朱学芳，孔杉杉，王震宇，熊文婷
南京大学信息管理学院，中国南京市，210023

摘要：技术经理人在促进技术转移过程中发挥着重要作用。尽管大语言模型（LLMs）已在该领域展现出潜力，但由于缺乏领域专用基准，评估其在技术经理人所需核心认知能力方面的表现仍面临挑战。为此，构建了一个用于评估大语言模型在技术经理人所需认知能力方面表现的领域专用基准——TMBench。该基准包含3474个问题，覆盖多种情境，涉及3项任务和5种题型。该基准将常用评价指标与以大语言模型为评判者（LLM-as-a-Judge）的评价方法相结合，采用固定权重和熵权计算对应的维度得分和综合得分。基于该评价机制，在不同提示策略和温度设置下对9个代表性模型进行评估，以考察其能力、分析其错误特征，并将其与人类专家进行基线比较。结果表明，不同模型的能力表现存在差异，没有任何单一模型在3个能力层级上均位列第一。在高阶能力方面，LLMs的点估计优于人类基线，但差异不具有统计学显著性。在两项高阶任务中，基于思维链的提示与模型的提示敏感性相关，且开放权重模型取得更好的结果。错误分析表明，知识储备、检索和应用方面的不足是最常见的错误类别。基于上述发现，本文为LLMs在技术转移领域应用中的模型选择、提示工程和针对性优化提供参考。

关键词：技术转移；技术经理人；基准；大语言模型
https://doi.org/10.1631/ENG.ITEE.2026.0168