УДК 33

Бенчмаркинг потенциала больших языковых моделей в принятии управленческих решений

Тихонюк Алексей Александрович

Исследователь,
Национальный исследовательский университет
«Высшая школа экономики»,
101000, Российская Федерация, Москва, ул. Мясницкая, 20;
e-mail: info@hse.ru

Аннотация

В статье обоснована необходимость разработки специализированного инструмента оценки способностей больших языковых моделей (LLM) в сфере менеджмента. Предложен оригинальный фреймворк для бенчмаркинга управленческого потенциала LLM, основанный на моделях компетенций менеджера. Автор критически анализирует практику использования антропоцентричных тестов для оценки LLM, выявляя их методологические ограничения. В качестве альтернативы предлагается адаптация методологии Evidence-Centered Design (ECD) для создания специализированных оценочных инструментов, учитывающих специфику языковых моделей. Исследование включает анализ существующих подходов к оценке профессиональной компетентности в смежных областях и разработку перспективной методологии бенчмаркинга, ориентированной на будущие потребности в оценке управленческого потенциала ИИ-систем.

Для цитирования в научных исследованиях

Тихонюк А.А. Бенчмаркинг потенциала больших языковых моделей в принятии управленческих решений // Экономика: вчера, сегодня, завтра. 2025. Том 15. № 2А. С. 696-705.

Ключевые слова

Менеджмент, принятие управленческих решений, бенчмаркинг, большие языковые модели, LLM, психометрика, ECD.

Введение

В последние годы большие языковые модели (large language models, LLMs) стремительно интегрируются в различные аспекты бизнеса и общества. Их применение варьируется от автоматизации рутинных задач до поддержки принятия стратегических решений. В западных странах 29% топ-менеджеров, 15% менеджеров среднего и нижнего звена и 40% магистров бизнес-направлений постоянно используют генеративный интеллект [Rosani, Farri, 2024]. 39% российских компаний уже внедрили ИИ-технологии [Ассоциация менеджеров, 2024]. Однако, несмотря на их широкое распространение и потенциал, отсутствуют стандартизированные инструменты для оценки их способностей в специфических областях, таких как менеджмент. Зачастую в таких случаях в качестве прокси для оценки полезности модели в конкретном контексте используется результат в неспецифичных тестах [Зеленков, 2024], что позволяет оценить применимость технологии в широком спектре потенциальных приложений, но не может быть использовано для определения профиля перформанса между дисциплинами или между темами внутри дисциплин. Данная лакуна обнаруживает пробел в понимании того, насколько эффективно LLM могут выполнять управленческие функции и какие аспекты требуют дальнейшего развития. Кроме того, отсутствие унифицированных методик измерения функциональных характеристик систем ИИ препятствует усилиям по формированию единых для индустрии критериев качества [Куприков, Башкирова, 2022].

Цель данного исследования — предложить подход к разработке бенчмарка для оценки управленческого потенциала LLM.

Основное содержание

LLM — это искусственные нейронные сети, обученные на больших объемах текстовых данных для генерации связного и осмысленного текста. Они способны обрабатывать и генерировать текст на основе заданного контекста, что открывает широкие возможности для их применения в различных областях знаний. Например, чат-бот, дообученный на транскрипциях тысяч часов профессиональных коучинговых сессий и профессиональной литературе, способен выполнять роль коуча на достаточно высоком уровне.

Необходимость сравнения между собой разных версий LLM для их совершенствования объясняет появление инструментов оценки их умений — бенчмарков (реже — бенчмарок). В общем смысле бенчмарки — это инструменты, позволяющие объективно оценивать и сравнивать системы или их компоненты по определённым характеристикам, таким как производительность, надёжность и безопасность. Самый простой вариант бенчмарка в сфере оценки LLM — набор вопросов, на которые модель генерирует свой ответ, а характеристикой перформанса является процент верных ответов. Например, бенчмарк MMLU-Pro содержит более 12 тысяч вопросов по выбору одного правильного ответа по 14 предметным областям от инженерии до права, тестовые вопросы сформулированы на английском языке [Wang et al., 2024].

Основные характеристики бенчмарков можно сгруппировать следующим образом [Kistowski et al., 2015]:

– Валидность, генерализация (бенчмарк проверяет именно то, что нужно проверять стейкхолдерам процесса, а результаты теста воспроизводимы в реальных условиях).

– Воспроизводимость (бенчмаркинг даёт одни и те же результаты при использовании тех

же настроек; данная характеристика также связана с детерминированностью ответов LLM при применении соответствующих настроек).

– Универсальность (выполнение бенчмарка зависит только от способности модели выполнять задания без искусственных ограничений).
– Проверяемость (бенчмаркинг проводится в среде, которая обеспечивает оценивание на равных).
– Удобство использования (бенчмарк может быть запущен в тестовой среде пользователя, что необходимо для защиты оцениваемого алгоритма от копирования).

Многие из этих принципов также относятся и к оценке людей. Именно поэтому многие бенчмарки включают задания, полученные из таких экзаменов, как ЕГЭ или GMAT. Зачастую на основании прохождения созданного для оценки людей теста делаются выводы об способностях LLM относительно людей. Однако оценивать LLM по тем же критериям, что и человека, некорректно. Основной аргумент против сопоставления производительности LLM и людей на тестах и профессиональных экзаменах заключается в том, что большинство этих тестов проверяют знание предметной области и недооценивают реальные навыки, необходимые в практике. Иными словами, эти тесты переоценивают именно те способности, в которых языковые модели особенно сильны. Кроме того, ИИ-модели решают задачи принципиально по-разному, а результаты тестов могут не отражать действительные способности в реальных ситуациях (генерализация результатов).

При сравнении тестов для людей и бенчмарков, ориентированных на LLM, наиболее заметное отличие заключается в их размере и структуре. Бенчмарки для LLM часто содержат тысячи заданий с большим количеством вариантов ответа, тогда как тесты для людей обычно состоят из десятка вопросов с ограниченным числом ответов. Возможность обойти свойственные людям ограничения человеческого внимания и способности к фокусировке позволяют создавать бенчмарки, в которых формулировка каждого вопроса может составлять десятки тысяч слов — формат, практически нереализуемый при оценке людей.

Чтобы учесть особенности LLM, которые отличают их от людей как объектов тестирования, необходимо обратить внимание также на следующие факторы:

– Источники вариативности. На LLM не воздействуют случайные внешние факторы, как это происходит у людей. Но у них есть много других источников вариативности ответов, например, чувствительность к контексту вопроса или чувствительность к формулировкам. LLM отлично распознают тонкие закономерности в данных и чрезвычайно чувствительны к изменениям в вводимых данных, которые кажутся незначительными для людей. Например, добавление пробела в конце запроса или изменение порядка инструкций без изменения их смысла может заметно повлиять на ответ.
– Влияние гиперпараметров. Например, гиперпараметр “temperature” регулирует степень случайности в ответах, “max tokens” ограничивает длину ответа, а “top P” определяет, сколько возможных вариантов продолжения ответа будет учтено.
– Классификация типов задач. Один из способов понять, в чём LLM более успешны, — это разделение задач на те, что связаны с созданием/генерацией, и те, что требуют понимания/распознавания [Davis, 2021]. Так, задачи на генерацию текста, аудио и видео исторически легче давались ИИ. В отличие от этого, человеческие способности к созиданию, как правило, предполагают предварительное понимание предметной

области.

По этим причинам традиционные методы оценки, применяемые к людям, не всегда релевантны для искусственного интеллекта. Необходимы специфические требования и метрики, учитывающие особенности LLM. Требования к оценке LLM должны включать надежность, воспроизводимость результатов, способность к адаптации в различных ситуациях и эффективность в решении конкретных задач, и этот список не является исчерпывающим [Fodor, 2025]. В контексте менеджмента это особенно актуально, поскольку управление требует не только технических знаний, но и понимания организационного контекста, человеческих факторов, навыков эмоционального интеллекта и способности принимать решения в условиях неопределенности.

Бенчмаркинг в области менеджмента: существующие и отсутствующие решения. На сегодняшний день отсутствуют бенчмарки, специально разработанные для целостной оценки LLM в области менеджмента. Однако отдельные релевантные нашим задачам бенчмарки существуют. Мы предлагаем следующую неисчерпывающую группировку требуемых от LLM знаний, умений и навыков в области менеджмента на несколько категорий, опирающуюся на классическую компетентностную модель Р. Катца:

1) Базовые навыки работы с информацией. Данная группа бенчмарков оценивает умение LLM извлекать информацию из мультимодальных источников, критически анализировать и обобщать её, систематизировать большие объёмы данных, проверять достоверность данных, переформулировать текст с сохранением смысла и адаптацией к целевой аудитории.
2) Отраслевые знания. Данная группа бенчмарков оценивает знания LLM в области экономики предприятий, менеджмента, трудового права, финансового менеджмента, маркетинга, HR-управления, а также понимание отраслевых особенностей и применение соответствующих законодательных и нормативных актов.
3) Управленческие навыки. Данная группа бенчмарков оценивает умение LLM осуществлять эффективные управленческие коммуникации, разрабатывать стратегии мотивации и лидерства, управлять конфликтами, организовывать командную работу, планировать и распределять ресурсы, обеспечивать контроль и обратную связь.
4) Решение управленческих задач (принятие управленческих решений). Данная группа бенчмарков оценивает способность LLM анализировать комплексные бизнес-кейсы, проводить стратегическую оценку, предлагать обоснованные варианты действий в условиях неопределённости, рассчитывать риски и прогнозировать последствия, а также разрабатывать планы внедрения изменений с учётом человеческого фактора.

В настоящее время уже существуют бенчмарки, нацеленные на оценку некоторых из заявленных знаний, умений и навыков. Группа 1 покрыта ими полностью, хотя продолжают разрабатываться всё более ухищрённые способы проверки данных умений. Например, большая группа бенчмарков проверяет робастность — стабильность ответов LLM при незначительных изменениях в вопросе — речь идёт об опечатках, использовании синонимов, перестановке пунктов списка и т.д. Группа 2 покрыта частично, в первую очередь за счёт использования созданных для оценки людей тестов в качестве бенчмарка для LLM [см. обзор Rosani, Farri, 2024]. Навыки и умения групп 3 и 4 не могут быть оценены с помощью существующих инструментов за исключением немногих исключений [например, Yang et al., 2024].

Отсутствие средств оценки LLM в данных областях подтверждает необходимость создания бенчмарка профессиональных компетенций в области менеджмента. Подчеркнём, что

профессиональная компетентность не сводится к соответствию набору разрозненных требований — речь идёт скорее об интегральной компетентности, проявляющейся в холистическом анализе ситуации с использованием знаний из разных областей, применением разнообразных и адекватных ситуации способов работы с информации и в принятии решений, учитывающих контекст ситуации и вероятностных профиль рассмотренных сценариев. Многие модели компетентности менеджера включают 5 типов элементов: знания, умения, представления о себе и ценности, черты характера и мотивы [Chouhan, Srivastava, 2014]. Очевидно, что не все из этих компонентов реализуются у людей также, как и у машин. Но в то же время необходимо учитывать их, чтобы понять поведение менеджера, «усиленного» LLM. Например, в области мотивов важно понимать, является ли приоритетом модели помощь пользователю или обеспечение интересов организации. Обычно для оценки профессиональной компетентности используются открытые форматы. Например, для оценки управления конфликтами можно использовать динамические симуляции переговоров, а для оценки бизнес-мышления — задачи с неструктурированными данными, важными контекстуальными особенностями и противоречащими друг другу целями.

Подход к созданию бенчмарка в области менеджмента. Всё больше внимания уделяется проблемам бенчмаркинга: низкой генерализации результатов (например, из-за низкой робастности), контаминации данных (обучении модели на существующих бенчмарках, которые затем используются для оценки модели), неясной интерпретации полученных результатов. Если первые две проблемы решаются техническими специалистами через изменение архитектуры модели, настройку механизма внимания и чистку обучающего датасета, то проблема улучшения интерпретируемости результатов привлекла внимание специалистов по оцениванию людей в образовании. Методы оценки людей, как мы показали выше, нельзя без изменений перенести на LLM, однако выработанные в психологии, педагогике и психометрике методы позволяют оценить ИИ-модели валидным и надёжным способом.

Раскроем смысл понятий «валидность» и «надёжность» измерения. Валидность — характеристика теста, который измеряет именно того, что мы хотим измерить [Cook, Beckman, 2006]. Так, например, тест, состоящий только из вопросов по истории менеджмента, не является валидным инструментом оценивания управленческих способностей, поскольку при его прохождении тестируемый демонстрирует знание истории менеджмента, а не способность осуществлять управление. Этот пример является утрированным, рассмотрим менее очевидный случай. Упомянутый выше бенчмарк MMLU-Pro [Wang et al., 2024] расшифровывается как Massive Multitask Language Understanding; может показаться, что он оценивает понимание текстов на естественном языке. В какой-то степени это верно и консистентно с принятым в области словоупотреблением, однако задания этого бенчмарка также оценивают знания модели в профессиональных областях (математика, право…), её логику и не содержат заданий, направленных на разрешение более тонких лингвистических механизмов (например, разрешение кореференции). Таким образом, этот бенчмарк, несмотря на высокое качество его заданий, не является валидным в классическом смысле этого слова, так как оценивает не совсем то, что заявляется.

Надёжность измерения — свойство измерения давать оценку, которая близка к истинной. Приведём пример. Сравним два формата итогового теста по университетскому курсу: тест из 5 и 50 вопросов. Очевидно, что роль случайности в итоговом балле каждого студента будет выше в более коротком тесте, что означает, что его результаты менее надёжны. Размер теста — далеко не единственный аспект, который оказывает влияние на надёжность изменения, однако

очевидно, что составные бенчмарки, включающие тысячи вопросов, в этом отношении превосходят тесты, используемые для оценки людей. Заметим, что даже самый надёжный невалидный тест не даёт полезную оценку искомого качества или умения модели.

С точки зрения психометрики, разработка тестов для людей требует четкого определения целевого конструкта и экспертного создания заданий. Это обеспечивает адекватное представление целевого поведения и ясную связь между заданием и измеряемым конструктом. Психометрически обоснованные тесты фокусируются на создании и совершенствовании заданий, тогда как бенчмарки для LLM больше ориентированы на компиляцию больших наборов данных и контроль качества. Процесс разработки тестов для людей включает пилотные исследования и детальный анализ данных с последующими изменениями перед фактическим использованием. В отличие от этого, бенчмаркинг LLM больше фокусируется на анализе ошибок модели и согласованности результатов различных попыток оценки целевого конструкта и уделяет меньше внимания подробной статистической валидации.

Различия в процессах разработки обусловлены разными подходами к вопросу: как мы можем гарантировать, что задания измеряют именно то, что мы хотим, и в совокупности дают ясное представление о целевом конструкте? Подход к разработке тестов для людей основан на причинно-следственной теории измерения, где задания выбираются таким образом, чтобы исключить альтернативные объяснения поведения и охватить все аспекты целевого конструкта. Бенчмаркинговый подход следует логике представительности, формируя тест как репрезентативную выборку поведения, вызываемого измеряемым качеством [Federiakin, 2025].

Оптимальный инструмент оценки для LLM может быть разработан путем объединения сильных сторон обоих подходов — разработки тестов и бенчмаркинга. Необходимо оценивать как способности LLM к генерации и пониманию, так и учитывать особенности их функционирования, отличные от человеческого мышления. Кроме того, характеристики входных данных — такие как сложность, ясность, длина и количество инструкций — имеют значительно большее значение для LLM и должны быть тщательно продуманы.

Было разработано несколько подходов к решению данной задачи [Fang, Oberski, Nguyen, 2024; Kardanova et al., 2024]. Исследовательский проект, реализованный в НИУ ВШЭ [Кузьминов, Кручинская, 2024; Kardanova et al., 2024], позволил провести качественную оценку профессиональных знаний ряда моделей и определить, в частности, что «языковые модели GigaChat Pro и GPT-4 допускают до 50% ошибок в теоретических основах права, образования и экономики, поскольку не обладают базовыми профессиональными знаниями. Все известные методики дообучения пока не могут предложить оптимального решения» [Кузьминов, Кручинская, 2024, с. 70]. Подобный качественный вывод невозможен без холистического подхода к дизайну и разработке бенчмарка, в процессе которых учитывается необходимость последующей корректной интерпретации на самых ранних этапах создания теста. Предложенный авторами подход опирается на методологию Evidence-Centered Design (ECD), созданную в психометрике для создания тестов таким образом, чтобы принимаемые в процессе разработки теста решения (1) обеспечивали валидность теста и (2) обеспечивали корректность и соответствие запросам стейкхолдеров интерпретации результатов тестирования и использования его результатов. Авторы предлагают разрабатывать задания в рамках трёхмерной матрицы, плоскостями которой являются:

1) Предметная область (трудовое законодательство, лидерство, экономика предприятия, эмоциональный интеллект и т.д.);
2) Уровни когнитивных операций по таксономии Бенджамина Блума (воспроизведение,

понимание, применение);

3) Предполагаемый уровень трудности заданий.

Мы полагаем, что подобный подход, уже опробованный на юриспруденции, педагогике и экономике может быть успешно реализован и для создания бенчмарка управленческой деятельности. Как было показано ранее, для лучшей генерализации результатов степень достаточности, формализации и даже правдивости контекста должна варьироваться, а формат заданий должен быть приближен к реальным ситуациям: динамическим бизнес-кейсам, переговорам, задачам на принятие решений при неполной информации. Благодаря такой интеграции ECD и компетентностной модели менеджера можно получить не просто статистику правильных ответов по категориям, но выявить сильные и слабые стороны LLM по нескольким измерениям.

Заключение

Оценка LLM в области менеджмента является сложной задачей, требующей учета контекста целевых сценариев применения искусственного интеллекта. Оценить не значит понять; важно не только измерить производительность моделей, но и понять, как они могут быть интегрированы в управленческую практику. Эффективность применения LLM в менеджменте зависит от того, насколько они способны усилить работу менеджера, предоставляя дополнительную информацию, анализ или поддержку в принятии решений. Различные подходы к разработке бенчмарков отражают разные понимания роли LLM и подходы к деятельности в целом. Иногда более простой и практичный бенчмарк может быть более полезным для конкретных целей.

Необходимо учитывать, что создание подобного бенчмарка — лишь первый шаг. Необходимо также провести пилотные исследования, проверить надёжность и валидность разработанных задач, скорректировать модель при реальной апробации на разных LLM, а также удостовериться в том, что стейкхолдеры процесса — от менеджеров, принимающих решение о допустимости использования LLM в работе, до программистов, дообучающих модели на основании получаемых оценок — правильно понимают границы интерпретации и использования результатов. В долгосрочной перспективе такой проект позволит сформировать отраслевой стандарт, облегчающий интеграцию ИИ-инструментов в повседневную деятельность руководителей, повышая общий уровень качества управления и конкурентоспособность организаций.

Применение LLM в управленческом контексте должно описаться на анализ рисков, связанных с охраной данных, уязвимостью к адверсариальным атакам, PR, нарушениям этических принципов организации и т.д. [Yuan et al., 2024]. М. Камински справедливо отмечает, что обычно регулирование рисков стремится «починить» проблемы с технологией, чтобы её можно было использовать дальше, но не рассматривает возможность, что использовать её в некоторых случаях вовсе нецелесообразно. Регулирование рисков лучше всего справляется с количественно измеримыми проблемами и испытывает трудности с вредом, который сложно оценить количественно. Подобный подход ошибочно представляет принципиальные вопросы о применимости технологии как технические решения [Kaminski, 2023]. Мы полагаем, что, во-первых, оценка возможностей ИИ может становиться таким техническим решением, а потому лицам, принимающим решения, необходимо учитывать контекст при интерпретации оценок компетентности моделей; во-вторых, результаты бенчмаркинга помогут избежать

использования LLM в областях, где их компетенции недостаточны, снижая риски.

Подчеркнём, что LLM служат скорее для аугментации, а не замены менеджеров. Поэтому оценивать следует не столько навыки менеджмента, которые могут имитировать LLM, но и то, как искусственный интеллект усиливает и поддерживает менеджера в его деятельности. Внедрение LLM в сферу менеджмента создаёт новые возможности для повышения эффективности использования имеющихся в компании ресурсов [Зеленков, 2024] и принятия обоснованных решений. Однако для полноценного использования этого потенциала необходимо разработать адекватные инструменты оценки и понять, как лучше интегрировать эти технологии в существующие процессы. Создание специализированного бенчмарка является важным шагом на этом пути, позволяющим внедрять LLM именно в те процессы, где он может быть полезным и для сотрудников, принимающих решения, и для компании в целом.

Библиография

1. Ассоциация менеджеров. 39% крупных российских компаний внедрили искусственный интеллект в свои бизнес-процессы [Электронный ресурс]. – 18.07.2024. – Режим доступа: https://amr.ru/press/news/gr/39-krupnykh-rossiyskikh-kompaniy-vnedrili-iskusstvennyy-intellekt-v-svoi-biznes-protsessy/ (дата обращения: 22.01.2025).
2. Зеленков Ю. А. Управление знаниями организации и большие языковые модели // Российский журнал менеджмента. – 2024. – Т. 22, № 3. – С. 573–601.
3. Кузьминов Я. Потенциал генеративного искусственного интеллекта для решения профессиональных задач / Я. Кузьминов, Е. Кручинская // Форсайт. – 2024. – Т. 18, № 4. – С. 67–76.
4. Куприков Н. М. Вопросы стандартизации в сфере искусственного интеллекта / Н. М. Куприков, Е. А. Башкирова // Компетентность. – 2022. – № 3. – С. 14–18.
5. Chang Y. A survey on evaluation of large language models / Y. Chang, X. Wang, J. Wang [et al.] // ACM Trans actions on Intelligent Systems and Technology. – 2024. – Vol. 15, No 3. – P. 1–45.
6. Chouhan, V. S. Understanding competencies and competency modeling—A literature survey / V. S. Chouhan, S. Srivastava // IOSR Journal of Business and management. – 2014. – Vol. 16 (1). – P. 14-22.
7. Cook D. A. Current concepts in validity and reliability for psychometric instruments: theory and application / D. A. Cook, T. J. Beckman // The American Journal of Medicine. – 2006. – Vol. 119, No 2. – P. 166–e7.
8. Davis E. Using human skills taxonomies and tests as measures of artificial intelligence / E. Davis // AI and the Future of Skills, Volume 1: Capabilities and Assessments, Educational Research and Innovation / OECD. – Paris: OECD Publishing, 2021. – 334 p.
9. Egele R. AI Competitions and Benchmarks: Dataset Development [Электронный ресурс] / R. Egele, J. Junior, C. S. Jacques [et al.]. – 2024. – arXiv:2404.09703. – Режим доступа: https://arxiv.org/abs/2404.09703 (дата обращения: 15.01.2025).
10. Fang Q. PATCH! Psychometrics-AssisTed benCHmarking of Large Language Models: A Case Study of Proficiency in 8th Grade Mathematics [Электронный ресурс] / Q. Fang, D. L. Oberski, D. Nguyen. – 2024. – arXiv:2404.01799. – Режим доступа: https://arxiv.org/abs/2404.01799 (дата обращения: 17.01.2025).
11. Federiakin D. Improving LLM Leaderboards with Psychometrical Methodology [Электронный ресурс] / D. Federiakin. – 2025. – arXiv:2501.17200. – Режим доступа: https://arxiv.org/abs/2501.17200 (дата обращения: 19.01.2025).
12. Fodor J. Line Goes Up? Inherent Limitations of Benchmarks for Evaluating Large Language Models [Электронный ресурс] / J. Fodor. – 2025. – arXiv preprint arXiv:2502.14318. – Режим доступа: https://arxiv.org/abs/2502.14318 (дата обращения: 24.01.2025).
13. Kaminski M. E. Regulating the Risks of AI / M. E. Kaminski // BUL Review. – 2023. – Vol. 103. – P. 1347–1411.
14. Kardanova E. A. Novel Psychometrics-Based Approach to Developing Professional Competency Benchmark for Large Language Models [Электронный ресурс] / E. Kardanova, A. Ivanova, K. Tarasova [et al.]. – 2024. – arXiv:2411.00045. – Режим доступа: https://arxiv.org/abs/2411.00045 (дата обращения: 20.01.2025).
15. Katz R. L. Skills of an effective administrator / R. L. Katz // Harvard Business Review. – 1955. – Vol. 33, No 1. – P. 33–42.
16. Rosani G., Farri E. How the Next Generation of Managers Is Using Gen AI / G. Rosani, E. Farri. – Harvard Business Review, 20.09.2024. – Режим доступа: https://hbr.org/2024/09/how-the-next-generation-of-managers-is-using-gen-ai.
17. v. Kistowski J. How to build a benchmark / J. v. Kistowski, J. A. Arnold, K. Huppler, K. D. Lange, J. L. Henning, P. Cao // Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering. – January 2015. – P. 333–336.

18. Wang Y. MMLU-pro: A more robust and challenging multi-task language understanding benchmark / Y. Wang, X. Ma, G. Zhang [et al.] // The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. – June 2024. – arXiv:2406.01574v6.
19. Yang Z. Benchmarking llms for optimization modeling and enhancing reasoning via reverse socratic synthesis [Электронный ресурс] / Z. Yang, Y. Huang, W. Shi [et al.]. – 2024. – arXiv e-prints, arXiv:2407. – Режим доступа: https://arxiv.org/abs/2407 (дата обращения: 23.01.2025).
20. Yuan T. R-judge: Benchmarking safety risk awareness for LLM agents [Электронный ресурс] / T. Yuan, Z. He, L. Dong [et al.]. – 2024. – arXiv:2401.10019. – Режим доступа: https://arxiv.org/abs/2401.10019 (дата обращения: 25.01.2025).

Benchmarking the Potential of Large Language Models in Managerial Decision-Making

Aleksei A. Tikhonyuk

Researcher,
National Research University "Higher School of Economics",
101000, 20, Myasnitskaya str., Moscow, Russian Federation;
e-mail: info@hse.ru

Abstract

This article argues for the necessity of developing a specialized tool for assessing the capabilities of large language models (LLMs) in the field of management and proposes a framework for this task based on the competency models of managers. Despite the widespread use of LLMs in various fields, particularly in decision-making, there is currently no tool available worldwide to assess how useful these models can be for managers in their daily activities. While the standard practice is to evaluate LLMs using tests designed for humans, we note significant shortcomings of this approach and suggest the creation of new solutions for assessing LLM capabilities — specifically, an original benchmarking tool for managerial potential. We analyze existing solutions for assessing professional competence in related fields and propose a methodology for creating such a benchmark in management, relying not on the industry standards of today, but on the most promising approaches to evaluating language models, based on the adaptation of the Evidence-Centered Design (ECD) methodology to the development of tools for LLMs.

For citation

Tikhonyuk, A.A. (2025) Benchmarking potentsiala bolshikh yazykovykh modeley v prinyatii upravlencheskikh resheniy [Benchmarking the Potential of Large Language Models in Management Decision-Making]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 15 (2A), pp. 696-705.

Keywords

Management, decision-making, benchmarking, LLM, psychometrics, ECD.

References

1. Assotsiatsiya menedzherov. (2024, July 18). *39% krupnykh rossiyskikh kompaniy vnedrili iskusstvennyy intellekt v svoi biznes-protsessy* [39% of large Russian companies have implemented artificial intelligence in their business

processes]. https://amr.ru/press/news/gr/39-krupnykh-rossiyskikh-kompaniy-vnedrili-iskusstvennyy-intellekt-v-svoi-biznes-protsessy/
2. Zelenkov, Yu. A. (2024). Upravlenie znaniyami organizatsii i bolshie yazykovye modeli [Knowledge management in organizations and large language models]. Rossiyskiy zhurnal menedzhmenta [Russian Management Journal], 22(3), 573-601.
3. Kuzminov, Ya., & Kruchinskaya, E. (2024). Potentsial generativnogo iskusstvennogo intellekta dlya resheniya professionalnykh zadach [The potential of generative artificial intelligence for solving professional tasks]. Forsayt [Foresight], 18(4), 67-76.
4. Kuprikov, N. M., & Bashkirova, E. A. (2022). Voprosy standartizatsii v sfere iskusstvennogo intellekta [Issues of standardization in the field of artificial intelligence]. Kompetentnost [Competence], 3, 14-18.
5. Chang, Y., Wang, X., Wang, J., et al. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1-45.
6. Chouhan, V. S., & Srivastava, S. (2014). Understanding competencies and competency modeling—A literature survey. IOSR Journal of Business and Management, 16(1), 14-22.
7. Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. The American Journal of Medicine, 119(2), 166-e7.
8. Davis, E. (2021). Using human skills taxonomies and tests as measures of artificial intelligence. In AI and the Future of Skills, Volume 1: Capabilities and Assessments. OECD Publishing.
9. Egele, R., Junior, J., Jacques, C. S., et al. (2024). AI Competitions and Benchmarks: Dataset Development [arXiv:2404.09703]. https://arxiv.org/abs/2404.09703
10. Fang, Q., Oberski, D. L., & Nguyen, D. (2024). *PATCH! Psychometrics-AssisTed benCHmarking of Large Language Models: A Case Study of Proficiency in 8th Grade Mathematics* [arXiv:2404.01799]. https://arxiv.org/abs/2404.01799
11. Federiakin, D. (2025). Improving LLM Leaderboards with Psychometrical Methodology [arXiv:2501.17200]. https://arxiv.org/abs/2501.17200
12. Fodor, J. (2025). Line Goes Up? Inherent Limitations of Benchmarks for Evaluating Large Language Models [arXiv:2502.14318]. https://arxiv.org/abs/2502.14318
13. Kaminski, M. E. (2023). Regulating the Risks of AI. BUL Review, 103, 1347-1411.
14. Kardanova, E. A., Ivanova, A., Tarasova, K., et al. (2024). Novel Psychometrics-Based Approach to Developing Professional Competency Benchmark for Large Language Models [arXiv:2411.00045]. https://arxiv.org/abs/2411.00045
15. Katz, R. L. (1955). Skills of an effective administrator. Harvard Business Review, 33(1), 33-42.
16. Rosani, G., & Farri, E. (2024, September 20). How the Next Generation of Managers Is Using Gen AI. Harvard Business Review. https://hbr.org/2024/09/how-the-next-generation-of-managers-is-using-gen-ai
17. von Kistowski, J., Arnold, J. A., Huppler, K., Lange, K. D., Henning, J. L., & Cao, P. (2015). How to build a benchmark. *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, 333-336.
18. Wang, Y., Ma, X., Zhang, G., et al. (2024). MMLU-pro: A more robust and challenging multi-task language understanding benchmark [arXiv:2406.01574v6].
19. Yang, Z., Huang, Y., Shi, W., et al. (2024). Benchmarking llms for optimization modeling and enhancing reasoning via reverse socratic synthesis [arXiv:2407]. https://arxiv.org/abs/2407
20. Yuan, T., He, Z., Dong, L., et al. (2024). R-judge: Benchmarking safety risk awareness for LLM agents [arXiv:2401.10019]. https://arxiv.org/abs/2401.10019