International Scientific and Practical Conference Samarkand Institute of Economics and Service
Vol. 1 Issue 1 | pp. 348-351 **| ISSN:** XXXX-XXXX
Available online @ https://globalconf.net/index.php/samisi

Article

Evaluating Generative AI Assistants for Russian-Language Office Tasks: Methodology and Results for Small and Medium-Sized Enterprises

Berkinbaev Margulan Kamalbekovich*[1], Eshenkulova Gaukhar Ilyasovna[2]

1,2. Astana IT University, Astana, Kazakhstan
* Correspondence: 255239@astanait.edu.kz

Abstract: Small and medium-sized enterprises use generative AI assistants to prepare documents, yet reproducible evidence on their quality for Russian-language office tasks is scarce. The paper proposes an evaluation method based on eight office tasks with embedded traps (missing data, a duplicate item, a cancelled decision, a ban on promises) and rule-based checks of formal requirements. Six systems (GPT-6, GPT-5.6 Sol, Claude Opus 5.5, Claude Sonnet 5.5, Gemini, DeepSeek) performed each task three times through their consumer web interfaces (144 outputs). The share of passed checks was 83–100%; all violations concerned length control. Interface artefacts (service messages, the real system date) that an employee may copy into a document were identified. Practical rules for verifying AI output in organisations are proposed.

Keywords: generative artificial intelligence, large language models, office automation, quality evaluation, small and medium-sized enterprises, digital economy.

Citation: Kamalbekovich B. M., & Ilyasovna, E. G. Evaluating Generative AI Assistants for Russian-Language Office Tasks: Methodology and Results for Small and Medium-Sized Enterprises. International Scientific and Practical Conference Samarkand Institute of Economics and Service 2026, 1(1), 348-351.

Received: 12th Aug 2026
Revised: 26th Sept 2026
Accepted: 8th Oct 2026
Published: 21th Nov 2026

Copyright: © 2026 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Генеративный искусственный интеллект (ИИ) быстро становится рабочим инструментом офисных сотрудников. Эксперименты показывают, что ИИ-ассистенты сокращают время подготовки текстов и повышают производительность в поддержке клиентов [1, 2], а значительная часть профессий подвержена влиянию больших языковых моделей (БЯМ) [3]. Вместе с тем за пределами возможностей модели качество работы с ИИ может снижаться [4], а «галлюцинации» — правдоподобные, но выдуманные утверждения — остаются системным свойством БЯМ [5].

Для стран Центральной Азии вопрос имеет прикладное значение: в Узбекистане и Казахстане внедрение ИИ закреплено на уровне государственных документов [6, 7], а малый и средний бизнес (МСБ) сталкивается прежде всего с организационными, а не техническими барьерами внедрения [8, 9]. Публичные бенчмарки оценивают в основном рассуждения и программирование [10] и почти не отвечают на вопрос, насколько можно доверять ассистенту при подготовке обычных деловых документов на русском языке.

Цель исследования — разработать воспроизводимый метод оценки ИИ-ассистентов на типовых офисных задачах и апробировать его на русскоязычном наборе заданий для шести распространённых систем. Англоязычная версия метода с оценкой жюри из моделей описана в отдельной работе, находящейся на

рецензировании [11]; настоящая статья представляет русскоязычную апробацию и практические выводы для МСБ.

2. Materials and Methods

Задания. Восемь заданий охватывают четыре области офисной работы: презентации (план квартального отчёта; презентация с просьбой о пилоте), деловые тексты (отчёт на 300–350 слов; проектное предложение на 250–300 слов), управление задачами (приоритизация 12 дел; поручения из протокола встречи) и коммуникации (ответ на жалобу клиента; пересказ переписки). Все данные синтетические — выдуманная сервисная компания в Астане. В каждое задание встроена «ловушка»: например, показатель удовлетворённости за квартал не измерялся, два пункта списка дублируют друг друга, решение о переносе встречи отменено, клиент требует возврат денег, который нельзя обещать.

Системы и протокол. Задания выполнены 16–17 ноября 2026 года в пользовательских веб-интерфейсах, которыми пользуются сотрудники: GPT-6 и GPT-5.6 Sol (chatgpt.com, план Plus, временный чат, рассуждение Medium), Claude Opus 5.5 и Claude Sonnet 5.5 (claude.ai, режим инкогнито), Gemini (бесплатный план, временный чат, режим «Авто») и DeepSeek (настройки по умолчанию). Каждое задание запускалось трижды в новом чате без системных инструкций, порядок запусков перемешан с фиксированным зерном; сохранялись полный текст ответа, время, адрес чата и настройки интерфейса. Получено 144 ответа.

Проверки. Формальные требования, которые можно проверить без экспертного суждения, оценивались программно по правилам, зафиксированным до сбора данных [12]: число слайдов, объём текста, обязательные заголовки, дословная подпись, отсутствие денежных сумм и названий продуктов, наличие всех 12 дел, отсутствие отменённого решения — всего 42 проверки на систему. Дополнительно фиксировались артефакты интерфейсов — текст, не относящийся к заданию. Оценка качества содержания по рубрике и сравнение с оценками моделей-судей [13] выполняются на следующем этапе.

3. Results and Discussion

Результаты

Таблица 1. Результаты программных проверок

Система	Пройдено проверок	Доля	Ответов без нарушений
GPT-6	42 из 42	100%	24 из 24
GPT-5.6 Sol	42 из 42	100%	24 из 24
Claude Opus 5.5	42 из 42	100%	24 из 24
Claude Sonnet 5.5	40 из 42	95%	22 из 24
Gemini	35 из 42	83%	17 из 24
DeepSeek	39 из 42	93%	21 из 24

Обе модели ChatGPT и Claude Opus 5.5 прошли все 42 проверки (таблица 1, рисунок 1). Все нарушения у остальных систем связаны только с объёмом: Gemini трижды вышел за диапазон 250–300 слов в проектном предложении, дважды написал отчёт короче 300 слов и дважды превысил лимит пересказа в 120 слов; DeepSeek во всех трёх повторах и Claude Sonnet в двух написали проектное предложение короче нормы. Ни одна система не включила отменённое решение в таблицу поручений, не указала денежных сумм в презентации о неодобренном пилоте и не нарушила подпись письма.

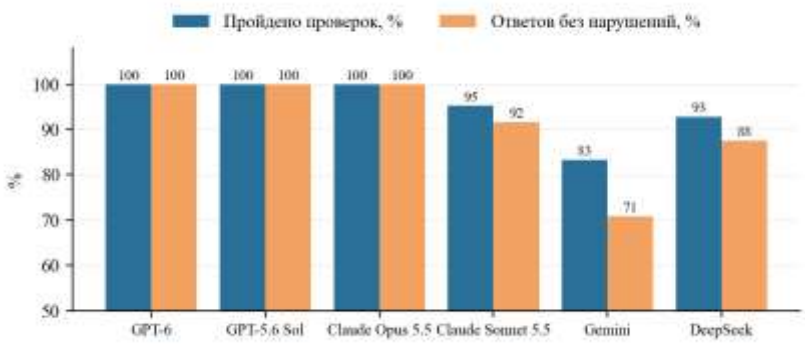

Рисунок 1. *Доля* пройденных проверок и ответов без нарушений

Рисунок 2 – Объём ответов в заданиях с ограничением *длины* (точки – повторы, *линия* – среднее, зелёная полоса – допустимый диапазон)

Рисунок 2 показывает различия в стиле. Gemini пишет самые *длинные* презентации и планы, но короче остальных — отчёты; DeepSeek систематически занижает объём проектного предложения; GPT-5.6 Sol — самая *лаконичная* система (около 80 слов в письме против примерно 150 у Claude Sonnet). *Для* сотрудника это означает разное время на сокращение или дописывание текста.

Таблица 2. Артефакты интерфейсов (число ответов из 24)

Артефакт	GPT-6	Sol	Opus	Sonnet	Gemini	DeepSeek
Кнопки-подсказки после подписи	3	3	0	0	0	0
Сообщение о режиме инкогнито	0	0	3	0	0	0
Упоминание своей системы	0	0	3	0	1	0
Реальная дата вместо даты задания	0	0	1	1	1	0

Артефакты интерфейсов (таблица 2) — свойство продукта, а не модели, но именно этот текст сотрудник скопирует в документ: подсказки интерфейса под черновиком письма, уведомления о невозможности сохранить документ, подстановка фактической даты вместо даты, заданной в условии.

Обсуждение

Результаты показывают, что формальные требования — структура, подпись, запреты — современные ассистенты соблюдают почти всегда. Основные риски другие: управление объёмом текста, посторонний служебный текст и, как показала англоязычная часть исследования [11], выдуманные статусы и решения, которые *легко* пропустить при беглом чтении. Поэтому эффективность ИИ в офисе определяется не только выбором модели, но и тем, встроена *ли* в процесс проверка результата. Это согласуется с выводом о том, что успех внедрения ИИ в МСБ зависит прежде всего от организационной готовности [8].

Для организаций Центральной Азии можно предложить три простых правила: (1) сверять цифры, даты, обещания и принятые решения в черновике с исходными данными; (2) удалять служебный текст интерфейса перед отправкой; (3) задавать объём явно и проверять его. Эти правила не требуют технических затрат и могут быть включены в шаблоны документов и внутренние регламенты использования ИИ.

Ограничения. Пользовательские интерфейсы меняются без уведомления и не раскрывают точные версии ряда моделей; восемь синтетических заданий не

охватывают всё разнообразие офисной работы; оценка качества содержания людьми и экспертами продолжается.

4. Conclusion

Предложенный метод — задания с «ловушками», повторные запуски и программные проверки — позволяет за короткое время и без затрат сравнить ИИ-ассистенты на русскоязычных офисных задачах. Шесть протестированных систем выполнили 83–100% формальных требований; различия между ними проявились в управлении объёмом текста и в служебном тексте интерфейсов. Дальнейшая работа включает экспертную оценку качества содержания, казахскую и узбекскую версии заданий и измерение времени выполнения задач сотрудниками.

REFERENCES

[1] Noy S., Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence // Science. – 2023. – Vol. 381, No. 6654. – P. 187–192.

[2] Brynjolfsson E., Li D., Raymond L. Generative AI at Work // The Quarterly Journal of Economics. – 2025. – Vol. 140, No. 2. – P. 889–942.

[3] Eloundou T., Manning S., Mishkin P., Rock D. GPTs are GPTs: Labor market impact potential of LLMs // Science. – 2024. – Vol. 384, No. 6702. – P. 1306–1308.

[4] Dell'Acqua F. et al. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. – Harvard Business School Working Paper 24-013, 2023.

[5] Ji Z. et al. Survey of hallucination in natural language generation // ACM Computing Surveys. – 2023. – Vol. 55, No. 12. – P. 1–38.

[6] Постановление Президента Республики Узбекистан от 17 февраля 2021 года № ПП-4996 «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта».

[7] Концепция развития искусственного интеллекта на 2024–2029 годы. – Утверждена постановлением Правительства Республики Казахстан от 24 июля 2024 года № 592.

[8] Oldemeyer L., Jede A., Teuteberg F. Artificial intelligence implementation challenges in SMEs: A systematic literature review // Management Review Quarterly. – 2025. – Vol. 75. – P. 1039–1092.

[9] Dwivedi Y. K. et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI // International Journal of Information Management. – 2023. – Vol. 71. – Art. 102642.

[10] Liang P. et al. Holistic evaluation of language models // Transactions on Machine Learning Research. – 2023.

[11] Berkinbayev M. TAL-Bench v2: Comparing ChatGPT, Claude and Gemini on office tasks with embedded traps, repeated runs and a cross-vendor LLM jury. – Submitted to IEEE Digital Generation Conference (DG 2027), under review.

[12] Zhou J. et al. Instruction-following evaluation for large language models. – arXiv:2311.07911, 2023.

[13] Zheng L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena // Advances in Neural Information Processing Systems. – 2023. – Vol. 36.