Летов Вячеслав, студент 3 курса бакалавриата,

Национальный исследовательский университет «Высшая школа экономики»

Проблемы обработки японского языка трансформерными языковыми моделями

На фоне продолжающегося бума больших языковых моделей (LLM) в мире и Японии возник противоречивый тренд. Хотя технология трансформеров существует уже девять лет, до сих пор остаются не решёнными до конца некоторые архитектурные ограничения обработки нейросетями японского языка, что затормаживает достижение японскими сообществами учёных и инженеров “суверенитета” в сфере искусственного интеллекта. Статья ставит перед собой следующий исследовательский вопрос: какие препятствия возникают на пути разработки японо-ориентированных больших языковых моделей и как такие ограничения могут влиять на их потенциал?

Обзор литературы. Для выявления и обобщения проблем было использовано три вида ресурсов. Во-первых, это литература, ориентирующаяся на определённых ограничениях при NLP обработке языков в целом (не считая английского) и/или японского языка в частности (Ли и др.[1]; Кодаки, Сасаки[2]; Риз[3]; Шут и др.[4]; Гамбарделла и др.[5]). Кроме того, работа Ду[6] позволила углубить понимание коренных различий между английским и японским языками. Во-вторых, это

[1] Multilingual Relative Clause Attachment Ambiguity Resolution in Large Language Models [Электронный ресурс] / S. Y. Lee, R. Scheinberg, A. Shore, A. Agrawal // arXiv preprint. — 2024.
[2] Kodaki K. Validating the Effectiveness of Fine-Tuning for Semantic Classification of Japanese Katakana Words: An Analysis of Frequency and Polysemy Effects on Accuracy [Электронный ресурс] / K. Kodaki, M. Sasaki // Big Data and Cognitive Computing. — 2026. — P. 1–8.
[3] Reese M. L. Commonsense Reasoning and World Knowledge Benchmark Tests: A Comparison of ChatGPT and Humans on the Japanese Winograd Schema Challenge [Электронный ресурс] / M. L. Reese // San Francisco State University. — 2024. — P. 19–59.
[4] Schut L. Do Multilingual LLMs Think In English? [Электронный ресурс] / L. Schut, Y. Gal, S. Farquhar // arXiv.org. — 2025.
[5] Gambardella A. Inconsistent Tokenizations Cause Language Models to be Perplexed by Japanese Grammar [Электронный ресурс] / A. Gambardella, T. Kojima, Y. Iwasawa, Y. Matsuo // arXiv.org. — 2025. — P. 1–8.
[6] Du J. Review of Language Structures and NLP Techniques for Chinese, Japanese, and English [Электронный ресурс] / J. Du // Proceedings of the 5th International Conference on Signal Processing and Machine Learning. — 2024.

статьи, посвящённые истории разработки японских локальных языковых моделей, а именно: LLM-jp-1 и LLM-jp-2 от организации LLM-jp[7] и Tsuzumi от Танака и др.[8]. Последний тип — прочие интернет-ресурсы: содержащие датасеты для дообучения японских моделей: WSCR-ja[9] и BCCWJ[10][11], содержащие информацию об используемой модели Llama-3.1-8B[12] при проведении эксперимента с логитной линзой, информация о которой взята из теоретического руководства на онлайн-форуме LessWrong[13], а реализация — из репозитория на Github[14].

Цель, задачи и методы. Используя описанные источники, исследование ставит своей целью выявление лингвистических ограничений трансформеров и стратегий их преодоления. Задачи сформированы по степени углубления в проблему:

1) описание англо-японских лингвистических различий;

2) систематизация актуальных ограничений LLM;

3) оценка существующих методов их решения;

4) проверка влияния ограничений на потенциал моделей.

В рамках исследования были проведены системный и сравнительный анализ и кейс-стади: использование логитной линзы на модели Llama-3.1-8B.

Результаты. Исследование позволило систематизировать три ключевые проблемы, ограничивающие эффективность современных трансформерных моделей при работе с японским языком. Во-первых, установлено, что доминирующая доля текстов на английском языке “видоизменяет” семантическое пространство модели, из-за чего она начинает обрабатывать

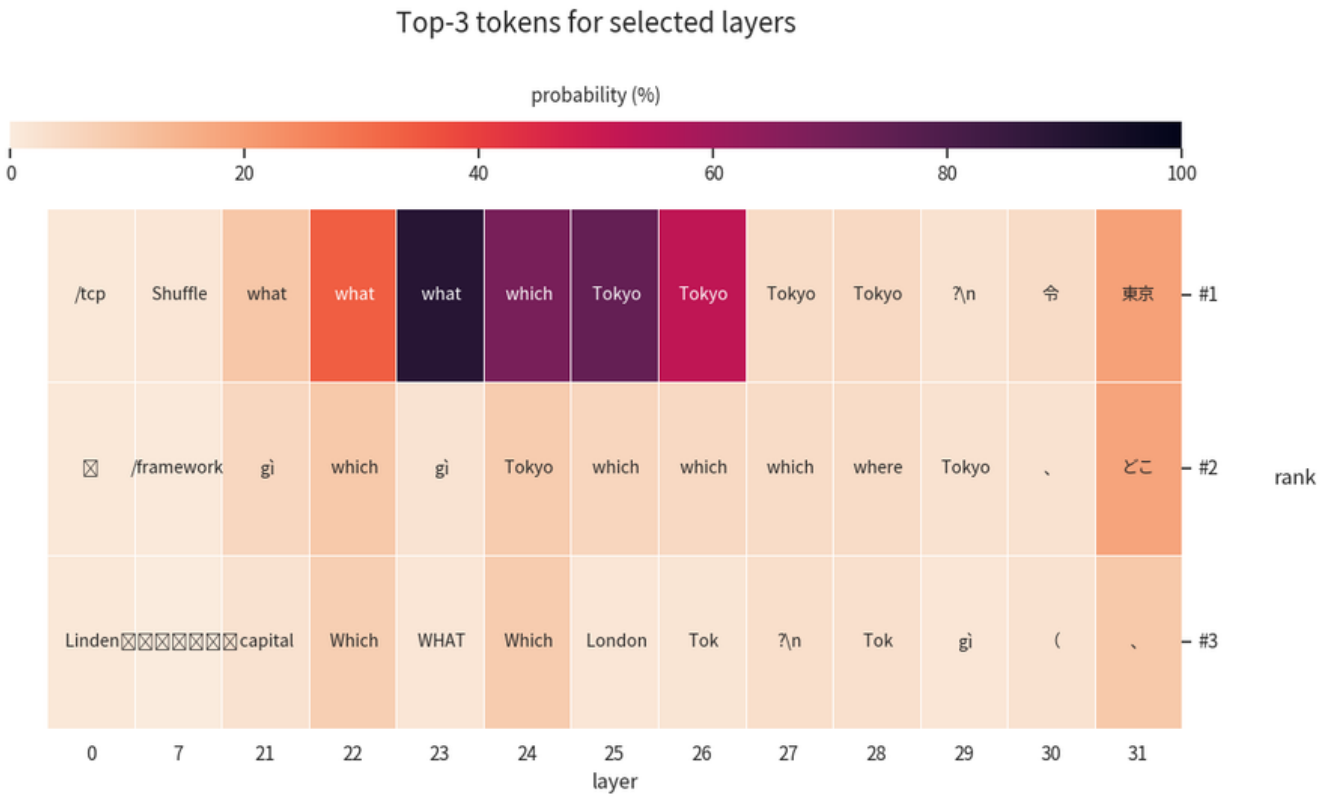
---

[7] LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs [Электронный ресурс] / LLM-jp // arXiv.org. — 2024.
[8] Shimizu K. NTT's LLM "tsuzumi" [Электронный ресурс] / K. Shimizu, K. Nishida, K. Nishida // NTT Technical Review. — 2024.
[9] KU-NLP. Winograd-Schema-Challenge-Ja [Электронный ресурс] / KU-NLP // GitHub. — 2026.
[10] Balanced Corpus of Contemporary Written Japanese [Электронный ресурс] / K. Maekawa // The 6th Workshop on Asian Language Resources. — 2008. — P. 101–102.
[11] NINJAL. Balanced Corpus of Contemporary Written Japanese [Электронный ресурс] / NINJAL // National Institute for Japanese Language and Linguistics. — 2026.
[12] Meta. Llama-3.1-8B [Электронный ресурс] / Meta // Hugging Face. — 2024.
[13] nostalgebraist. Interpreting GPT: the logit lens [Электронный ресурс] / nostalgebraist // LessWrong. — 2020.
[14] zhenyu-02. LogitLens4LLMs [Электронный ресурс] / zhenyu-02 // GitHub. — 2024.

запросы сначала на английском. Модель сначала генерирует абстрактное представление концепта на английском, а затем переводит его на японский.

Для проверки описываемого концепта автором был применён метод логитной линзы на модели Llama-3.1-8B со следующим промптом: “日本の首都は“ (яп. “Столица Японии — ”), результаты показаны на рис. 1. По оси Y перечислены слои модели (где 0 это первый слой, а 31 — последний), а по оси X — ранг токена среди наиболее вероятных вариантов продолжения (где #1 — это токен с наибольшей вероятностью, согласно softmax-распределению по словарю модели).

Рис. 1. Тепловая карта, полученная после применения логитной линзы на модели Llama-3.1-8B с промптом “日本の首都は“. Часть слоёв опущена для наглядности



Обработка моделью внутренних слоёв показала, что, несмотря на промпт, заданный на японском, модель сперва обрабатывает его на английском языке, что видно на средних слоях модели (7–29). Модель сперва сформулировала наиболее вероятный ответ на английском языке (первые ранги слоёв 22–26), и только на 31 слое перевела его на японский. Это, наравне с другими примерами, демонстрируемыми в статье Шут и др. подтверждает теорию англоцентричности моделей. В долгосрочной перспективе такая особенность архитектуры больших языковых

моделей приводит к синтаксическим искажениям и, потенциально, потере культурного контекста. Для нивелирования риска англоцентричности моделей японские корпорации и сообщества стараются обучать модели на преимущественно японском языке.

Во-вторых, эллиптическая структура японских предложений, где подлежащее, сказуемое и другие части речи могут опускаться (что особенно актуально для разговорного стиля), остается сложной задачей для трансформерных моделей. Использование специализированных датасетов, таких как Winograd-Schema-Challenge-Ja (WSCR-ja), лишь частично помогает нейросетям достраивать логическую связь между актантами действия, потому что они являются скорее “справочниками”, а не сложными алгоритмами для выявления нулевых анафор.

Наконец, большие языковые модели демонстрируют высокую частоту галлюцинаций при встрече с полисемичными словами категории *васэй-эйго*, значения которых в японском языке разошлись с английскими оригиналами (рис. 2).

Рис. 2. Семантические сдвиги заимствованной лексики *васэй-эйго*

Слово (катакана)	Оригинал на английском	Японское значение	Тип семантического сдвига
マンション (*мансён)*	Mansion (роскошный особняк)	Многоквартирный дом	Расширение
テンション (*тэнсён*)	Tension (умственное или физическое напряжение)	Эмоциональное возбуждение, высокая энергия	Сужение
マフラー (*мафура:*)	Muffler (автомобильный глушитель)	Шарф, автомобильный глушитель	Сдвиг смысла

Дообучение языковых моделей на специализированных датасетах, таких как The Balanced Corpus of Contemporary Written Japanese (BCCWJ), позволяет частично решить проблему: васэй-эйго, возникшие после сужения или сдвига смысла, модель может обучать эффективно. Однако слова, чья семантика стала шире, такие как マンション, становятся полисемичными, а точность их корректного машинного перевода понижается.

Таким образом, исследование подтверждает, что японский язык представляет собой вызов для трансформерных моделей. Выявленные ограничения носят комплексный характер, а методы их преодоления являются скорее компенсаторными, чем всеобъемлющими. Достижение подлинного технологического суверенитета Японии в сфере ИИ требует продолжения тренда на уменьшение доли обучающих датасетов на английском языке, а также усовершенствования алгоритмов определения нулевой анафоры и семантической роли некоторых заимствованных слов.

Литература

[1] Balanced Corpus of Contemporary Written Japanese [Электронный ресурс] / K. Maekawa // The 6th Workshop on Asian Languae Resources. — 2008. — P. 101–102. https://aclanthology.org/I08-7018.pdf

[2] Du J. Review of Language Structures and NLP Techniques for Chinese, Japanese, and English [Электронный ресурс] / J. Du // Proceedings of the 5th International Conference on Signal Processing and Machine Learning. — 2024. https://www.researchgate.net/publication/386118861_Review_of_Language_Structures_and_NLP_Techniques_for_Chinese_Japanese_and_English

[3] Gambardella A. Inconsistent Tokenizations Cause Language Models to be Perplexed by Japanese Grammar [Электронный ресурс] / A. Gambardella, T. Kojima, Y. Iwasawa, Y. Matsuo // arXiv.org. — 2025. — P. 1–8. https://arxiv.org/abs/2505.19599

[4] Kodaki K. Validating the Effectiveness of Fine-Tuning for Semantic Classification of Japanese Katakana Words: An Analysis of Frequency and Polysemy Effects on Accuracy [Электронный ресурс] / K. Kodaki, M. Sasaki // Big Data and Cognitive Computing. — 2026. — P. 1–8. https://www.mdpi.com/2504-2289/10/3/67

[5] LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs [Электронный ресурс] / LLM-jp // arXiv.org. — 2024. https://arxiv.org/abs/2407.03963

[6] Multilingual Relative Clause Attachment Ambiguity Resolution in Large Language Models [Электронный ресурс] / S. Y. Lee, R. Scheinberg, A. Shore, A. Agrawal // arXiv preprint. — 2024. https://arxiv.org/abs/2503.02971

[7] nostalgebraist. Interpreting GPT: the logit lens [Электронный ресурс] / nostalgebraist // LessWrong. — 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

[8] Reese M. L. Commonsense Reasoning and World Knowledge Benchmark Tests: A Comparison of ChatGPT and Humans on the Japanese Winograd Schema Challenge [Электронный ресурс] / M. L. Reese // San Francisco State University. — 2024. — P. 19–59. https://scholarworks.calstate.edu/downloads/1z40m196h

[9] Schut L. Do Multilingual LLMs Think In English? [Электронный ресурс] / L. Schut, Y. Gal, S. Farquhar // arXiv.org. — 2025. https://arxiv.org/abs/2502.15603

[10] Shimizu K. NTT's LLM "tsuzumi" [Электронный ресурс] / K. Shimizu, K. Nishida, K. Nishida // NTT Technical Review. — 2024. https://www.rd.ntt/e/research/JN202406_26651.html

Источники

[11] KU-NLP. Winograd-Schema-Challenge-Ja [Электронный ресурс] / KU-NLP // GitHub. — 2026. https://github.com/ku-nlp/Winograd-Schema-Challenge-Ja

[12] Meta. Llama-3.1-8B [Электронный ресурс] / Meta // Hugging Face. — 2024. https://huggingface.co/meta-llama/Llama-3.1-8B

[13] NINJAL. Balanced Corpus of Contemporary Written Japanese [Электронный ресурс] / NINJAL // National Institute for Japanese Language and Linguistics. — 2026. https://clrd.ninjal.ac.jp/bccwj/en/

[14] OpenAI o1 System Card [Электронный ресурс] / OpenAI // OpenAI. — 2024. https://cdn.openai.com/o1-system-card-20241205.pdf

[15] SB Intuitions Corp. [Электронный ресурс] // SB Intuitions Corp.: [сайт]. — 2026. https://www.sbintuitions.co.jp/en/

[16] zhenyu-02. LogitLens4LLMs [Электронный ресурс] / zhenyu-02 // GitHub. — 2024. https://github.com/zhenyu-02/LogitLens4LLMs