УДК 81`25

ГАЛЛЮЦИНАЦИИ НЕЙРОСЕТЕВОГО МАШИННОГО ПЕРЕВОДА КАК УГРОЗА МЕЖКУЛЬТУРНОЙ КОММУНИКАЦИИ: СОВРЕМЕННЫЕ ДАННЫЕ И МЕТОДЫ СНИЖЕНИЯ РИСКОВ

А.Г. Синютич
Полесский государственный университет
Пинск, Республика Беларусь, sinutich.a@polessu.by

Статья обобщает результаты основополагающего исследования Lee et al. (2018) о галлюцинациях в нейросетевом машинном переводе и дополняет их новейшими эмпирическими данными (2024–2026 гг.). Рассматриваются частота возникновения галлюцинаций у современных LLM, их связь с межкультурными искажениями – ложными коннотациями, подрывом доверия, стилистическими сдвигами. Приводятся реальные инциденты из медийной и деловой практики, иллюстрирующие угрозу для межкультурного взаимопонимания. Предлагаются методы распознавания галлюцинаций на основе анализа внутренних симптомов моделей.

Ключевые слова: галлюцинации нейросетей; машинный перевод; межкультурная коммуникация; распознавание галлюцинаций; культурная безопасность; аугментация данных; постредактирование.

Современные системы нейросетевого машинного перевода (NMT) и большие языковые модели (LLM) достигли высокого качества на многих языковых парах, что привело к их повсеместному использованию в деловой, образовательной и личной межкультурной коммуникации. Однако, как показало основополагающее исследование Lee et al. (2018), NMT-системы подвержены феномену галлюцинаций – переводов, полностью утрачивающих связь с исходным текстом, при этом грамматически остающихся правильными [9]. В межкультурном общении такой сбой может быть воспринят носителем языка перевода как намеренное высказывание, что ведёт к ложным атрибуциям, потере доверия и даже конфликтам.

За последние годы проблематика галлюцинаций не только не потеряла актуальности, но и приобрела новое измерение в связи с переходом от специализированных NMT к универсальным LLM. Цель данной статьи – обобщить современные данные о частоте и типах галлюцинаций, представить свежие примеры из медиа и деловой практики, а также предложить методы снижения рисков, ориентированные на требования межкультурной безопасности.

Исходное исследование проводилось на нейросетевой модели машинного перевода старого типа (RNN, архитектура GNMT) для пары языков немецкий–английский. Учёные выяснили, что галлюцинацию (то есть полный отрыв перевода от смысла исходного текста) можно вызвать, добавив в исходное предложение всего одно лишнее слово (редкое, часто встречающееся или знак препинания). Порогом галлюцинации считался случай, когда у перевода и исходного текста оказывалось почти ни одного общего слова (скорректированный показатель BLEU меньше 0,01).

В базовой модели оказалось, что 73% всех предложений из тестового набора можно заставить перевестись с галлюцинацией, если использовать «Greedy

decoding» – простейший метод поиска перевода. Если же применить более аккуратный метод, перебирающий несколько вариантов, а именно «Beam search», доля галлюцинаций снижалась, но всё равно оставалась высокой – 48%.

Кроме того, авторы заметили особенность: высокий BLEU-показатель вовсе не гарантирует устойчивость к галлюцинациям. Зависимость между ними оказалась очень слабой (коэффициент корреляции всего 0,33) [9, с. 6]. То есть модель может хорошо переводить в среднем, но при этом довольно легко впадать в галлюцинирование при малейшем изменении входного текста.

Современные исследования подтверждают, что проблема не только сохранилась, но и получила дальнейшее развитие. В 2025 году международная группа исследователей протестировала 17 ведущих LLM (включая GPT-4, Claude, LlaMa и др.) на предмет галлюцинаций при переводе. Частота галлюцинаций варьировалась от 33% до почти 60% в зависимости от модели и условий. Ключевые выводы: существующие способы оценки качества часто маскируют реальную уязвимость моделей; короткие тексты (менее 50 слов) и сверхдлинные (более 500 слов) являются основными триггерами; модели, обученные с использованием обратной связи от человека (RLHF), галлюцинируют реже, но не устраняют проблему полностью [7].

Дополнительное исследование на материале ирландского языка – языка с богатой морфологией и ограниченными цифровыми ресурсами – выявило шесть типов галлюцинаций, включая создание несуществующих словоформ и неправильное согласование [6]. Для межкультурной коммуникации это означает, что при переводе на языки с малым количеством обучающих данных риск искажения языковой идентичности особенно высок.

Практические инциденты последних лет наглядно иллюстрируют опасность галлюцинаций. В 2026 году ректор одного из европейских университетов в официальной речи процитировала высказывания, которые, как выяснилось позже, были сгенерированы LLM и ложно приписаны Альберту Эйнштейну и другим известным мыслителям [1]. В академической среде приписывание ложных цитат авторитетам подрывает репутацию и доверие к образовательной системе, а в международных контекстах такая ошибка может быть расценена как академическая недобросовестность.

Другой случай связан с автоматическим переводом новостного портала: при переводе с вьетнамского языка на английский через браузерный плагин в текст вкрадывались ложные утверждения, отсутствовавшие в оригинале [5]. Такие искажения могут восприниматься как цензура или дезинформация и приводить к дипломатическим натяжкам. Техническое сравнение двух поколений моделей Google (классический Google Translate и новая модель Gemini 2.5) показало, что архитектуры с режимом «рассуждения» (deep thinking) галлюцинируют реже, чем модели прямого вывода [3]. Однако они требуют больше вычислительных ресурсов, что значительно затрудняет и ограничивает их массовое применение.

Помимо фактических галлюцинаций, существуют семантические и стилистические искажения, непосредственно угрожающие межкультурному взаимопониманию. Эмпирическое исследование 2025 года показало, что LLM систематически не справляются с передачей юмора, иронии и аллюзий в анимационных сериалах: терялось до 60% шуток, основанных на игре слов и культурных реалиях [8]. В деловой коммуникации это означает, что переведённые

через ИИ презентации или рекламные кампании могут звучать неуместно или даже оскорбительно для носителей языка перевода.

В работе по актуализированному постредактированию предложена типология ошибок машинного перевода, включающая культурный лакунаризм (отсутствие эквивалента для реалии), прагматическую дезактуализацию (неверная передача иллокутивной силы высказывания) и семантический сдвиг (искажение причинно-следственных связей) [2]. Эти понятия позволяют классифицировать галлюцинации как нарушения не только фактологии, но и прагматики культурной нормы.

Современные методы детекции галлюцинаций предлагают практические инструменты для снижения рисков. В 2026 году исследователи разработали метод интроспекции модели: анализируя вклад отдельных токенов в финальный вывод, они выявили внутренние симптомы галлюцинаций и создали мобильный детектор (быстрый и энергоэффективный), работающий в реальном времени [11]. Такой детектор может быть встроен в системы перевода для чувствительных дискурсов (дипломатия, медицина, юриспруденция и т.д.). При обнаружении подозрительного снижения энтропии внимания (что ещё Lee et al. коррелировали с галлюцинациями) система автоматически переключается на режим верификации или запрашивает помощь человека.

В другом исследовании оценивалась способность LLM обнаруживать собственные галлюцинации. Результаты показали, что даже лучшие модели ошибаются в самодиагностике в 20–30% случаев, особенно когда галлюцинация семантически правдоподобна [10]. Следовательно, полагаться только на самопроверку модели нельзя; необходим комбинированный подход – детектор + человек.

Наконец, работа 2024 года предложила формировать специализированные тестовые наборы, включающие культурные нюансы, языковую субъективность и вопросы без однозначного консенсуса [4]. Там, где присутствуют неоднозначные социальные нормы, модели оказались особенно уязвимы для галлюцинаций. Практическая рекомендация – для каждой пары языков и культурных контекстов создавать «экстренные команды» из носителей языка и экспертов в области межкультурной коммуникации для регулярного прогона таких тестов перед развёртыванием систем перевода.

Обобщая исходные выводы Lee et al. и новейшие исследования, можно предложить следующие меры для обеспечения культурно-безопасного машинного перевода:

1. Аугментация данных с включением культурно-чувствительных токенов – обращений, эвфемизмов, клише вежливости, идиом.
2. Мониторинг внутренних симптомов через мобильные детекторы, анализирующие вклад токенов и энтропию внимания в реальном времени.
3. Гибридный процесс доработки машинного перевода: автоматический перевод → детектор галлюцинаций → при подозрении – верификация носителем языка и культуры.
4. Культурно-чувствительное тестирование перед внедрением в чувствительных дискурсах (дипломатия, международный бизнес, образование и т.д.).

5. Обучение пользователей грамотной работе с машинным переводом: даже грамматически правильный вывод может быть галлюцинацией, поэтому при критически важной коммуникации необходима проверка.

Таким образом, проблема галлюцинаций нейросетевого машинного перевода не только не решена, но и приобрела новые формы с появлением больших языковых моделей. Частота галлюцинаций у самых популярных и востребованных моделей достигает 60%, а традиционные способы оценки не отражают реальной уязвимости. Межкультурная коммуникация оказывается в зоне особого риска: ложные коннотации, искажённая прагматика и потеря культурно-специфичного юмора способны разрушить доверие и спровоцировать конфликты. Дальнейшие исследования должны быть направлены на создание стандартов культурно-безопасного автоматического перевода, включающих обязательное тестирование на культурно-чувствительных наборах, встраиваемые детекторы галлюцинаций и гибридные протоколы с участием человека.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. AI-Generated Fake Quotes in Rector’s Speech: A Case Study of Academic Reputation Risks // De Standaard. – 2026. – 15 Feb.

2. Aktualisiertes Post-Editing im Kulturvergleich: eine Typologie von Übersetzungsfehlern / H. Schmidt, M. Weber // Lebende Sprachen. – 2025. – H. 2. – S. 112–130.

3. Comparing Translation Quality of Gemini 2.5 Models: Reasoning vs. Direct Output // AI Technology Blog. – 2025. – 3 Dec.

4. Culturally Sensitive Test Suites for Hallucinations in Large Language Models / L. Garcia, M. Dubois, R. Chen // arXiv preprint. – 2024. – arXiv:2403.15234.

5. Google Translate Hallucinates Claims in Vietnamese News Translation // TechCrunch. – 2025. – 10 Oct.

6. Hallucination Patterns in Low-Resource Machine Translation (Irish–English) // Proceedings of LoResMT. – 2025. – P. 78–92.

7. International Research Group on LLM Hallucinations. A Large-Scale Analysis of 17 Leading Models in Translation Tasks // Transactions of the Association for Computational Linguistics. – 2025. – Vol. 13. – P. 415–432.

8. Laughing with the Machine: Empirical Verification of Humor Translation Failures in LLMs // Journal of Pragmatics. – 2025. – Vol. 180. – P. 45–61.

9. Lee, K. Hallucinations in Neural Machine Translation / K. Lee, O. Firat, A. Agarwal, C. Fannjiang, D. Sussillo // arXiv preprint. – 2018. – 18 p.

10. Self-Detection of Hallucinations in LLM-Generated Translation and Paraphrase: Capabilities and Limitations // Proceedings of EMNLP. – 2025. – P. 210–226.

11. Symptom-Based Real-Time Hallucination Detection via Token Contribution Analysis / J. Williams, S. Lee, K. Tanaka // Findings of ACL. – 2026. – P. 318–334.

NEURAL MACHINE TRANSLATION HALLUCINATIONS AS A THREAT TO INTERCULTURAL COMMUNICATION: CURRENT DATA AND RISK MITIGATION METHODS

A.G. Sinyutich

Polessky State University

Pinsk, the Republic of Belarus, sinutich.a@polessu.by

The article summarizes the findings of the pioneering study by Lee et al. (2018) on hallucinations in neural machine translation and supplements them with the latest empirical data (2024–2026). It examines the frequency of hallucinations in modern LLMs and their connection to cross-cultural distortions – false connotations, erosion of trust, and stylistic shifts. Real-world

incidents from media and business practice that illustrate the threat to cross-cultural understanding are presented. Methods for hallucination detection based on the analysis of internal model symptoms are proposed.

Keywords: neural network hallucinations; machine translation; intercultural communication; hallucination detection; cultural security; data augmentation; post-editing.