# Источники первого дня

Все ссылки, которые встречаются в лекциях, практике и дополнительных материалах первого дня. Нумерация общая для всех файлов дня.

## Научные публикации


- [1] Huang L., Yu W., et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. https://doi.org/10.1145/3703155
- [2] Farquhar S., Kossen J., et al. Detecting hallucinations in large language models using semantic entropy // Nature. 2024. https://doi.org/10.1038/s41586-024-07421-0
- [3] Chelli M., Descamps J., et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis // Journal of Medical Internet Research. 2024. https://doi.org/10.2196/53164
- [4] Mysior M., Iniotakis C. Evaluating large language models on text-based engineering design tasks: performance gaps and failure modes // Research in Engineering Design. 2026. https://doi.org/10.1007/s00163-026-00504-1
- [5] Omar M., Sorin V., et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support // Communications Medicine. 2025. https://doi.org/10.1038/s43856-025-01021-3
- [6] Dahl M., Magesh V., et al. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models // Journal of Legal Analysis. 2024. https://doi.org/10.1093/jla/laae003
- [7] Wei J., Wang X., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022. https://doi.org/10.48550/arXiv.2201.11903
- [8] Wang L., Xu W., et al. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. 2023. https://doi.org/10.48550/arXiv.2305.04091
- [9] Jeon S., Kim H.-G. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering // Computers in Biology and Medicine. 2025. https://doi.org/10.1016/j.compbiomed.2025.110614
- [10] Sorour S., Sahbel A. Large language models in intelligent manufacturing and mechanical engineering: a review of robotics, fault diagnosis, design, and engineering knowledge workflows // Journal of Intelligent Manufacturing. 2026. https://doi.org/10.1007/

s10845-026-02927-y
- [13] Yao S., Zhao J., et al. ReAct: Synergizing Reasoning and Acting in Language Models. 2022. https://doi.org/10.48550/arXiv.2210.03629
- [14] Zhou D., Schärli N., et al. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. 2022. https://doi.org/10.48550/arXiv.2205.10625
- [15] Kojima T., Gu S. S., et al. Large Language Models are Zero-Shot Reasoners. 2022. https://doi.org/10.48550/arXiv.2205.11916
- [17] Alansari A., Luqman H. Large Language Models Hallucination: A Comprehensive Survey. 2025. Препринт arXiv. https://doi.org/10.48550/arXiv.2510.06265
- [16] Ülkir M., Paslı B. Reference Hallucination, Citation Reliability, and Readability of Large Language Models in Anatomy-Related Question Answering // Clinical Anatomy. 2026. https://doi.org/10.1002/ca.70187
- [22] Liu N. F., Lin K., et al. Lost in the Middle: How Language Models Use Long Contexts // Transactions of the Association for Computational Linguistics. 2024. https://doi.org/10.1162/tacl_a_00638
- [23] Mirzadeh I., Alizadeh K., et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024. Препринт arXiv. https://doi.org/10.48550/arXiv.2410.05229
- [24] Długosz D. A., Oliveira A., Rodríguez N. The Importance of Being Statistically Earnest: A Critical Re-evaluation of GSM-Symbolic // Proceedings of EMNLP 2026. https://aclanthology.org/2026.emnlp-main.1487/
- [25] Fanous A., Goldberg J., et al. SycEval: Evaluating LLM Sycophancy // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2025. https://doi.org/10.1609/aies.v8i1.36598

Работы [7], [8], [13], [14], [15] и [23] доступны на arXiv.

# Отраслевые документы

- [11] NIST AI 100-2e2025. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2025. https://doi.org/10.6028/NIST.AI.100-2e2025
- [12] OWASP GenAI Security Project. Agentic AI – Threats and Mitigations. 2025. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

# Сервисы


- S1. GigaChat. Справка: начало работы, частые вопросы, документы, изображения. https://giga.chat/help/articles/how-to-start-work-with-gigachat ; https://giga.chat/help/articles/faq ; https://giga.chat/help/articles/work-with-docs ; https://giga.chat/help/articles/gigachat-load-picture
- S2. Пользовательское соглашение GigaChat. https://developers.sber.ru/docs/ru/policies/gigachat/agreement
- S3. Условия использования сервиса «Алиса AI». https://yandex.ru/legal/alice_chat/
- S4. Алиса AI. Справка: чат, опция «Алиса Плюс», режим «Эксперт», файлы. https://alice.yandex.ru/support/ru/assistant/chat-alice ; https://alice.yandex.ru/support/ru/assistant/alice-plus ; https://alice.yandex.ru/support/ru/assistant/chat/intellect ; https://alice.yandex.ru/support/ru/assistant/chat/files
- S5. Политика конфиденциальности Яндекса. https://yandex.ru/legal/confidential/
- S6. DeepSeek Privacy Policy. https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html
- S7. DeepSeek Terms of Use; анонс DeepSeek-V4 Preview. https://cdn.deepseek.com/policies/en-US/deepseek-terms-of-use.html ; https://www.deepseek.com/en/news/v4-preview/
- S8. Qwen. Соглашение об оказании услуг; страница входа; описание сервиса. https://qwen.ai/termsservice ; https://chat.qwen.ai/auth ; https://qwen.ai/home
- S9. Qwen. Политика конфиденциальности. https://qwen.ai/privacypolicy


# Выступления и вебинары


- [18] Субботин А. Платформы для совместной работы с ИИ-агентами. Доклад на ИИ-митапе ЦГУ, 06.11.2026 (презентация доклада).
- [19] Овчинников К. Инженер после AI. Доклад на ИИ-митапе ЦГУ, 06.11.2026 (презентация доклада). На слайде о парадоксе надзора докладчик ссылается на материал Anthropic «How AI Is Transforming Work at Anthropic» (2025).
- [20] Гагарин П. Разработка с агентами: от постановки задачи до деплоя. Вебинар учебного центра «МАСКОМ», 28.07.2026 (презентация вебинара).
- [26] Сессия «Безопасность ИИ-моделей: угрозы, методы защиты и подходы к регулированию». ИТ-форум «Цифровые решения – 2026», Национальный центр «Россия», 09.10.2026. Выступления А. Кулашовой («Лаборатория Касперского»), Д.

Служеникина (консорциум безопасности технологий ИИ), А. Белеванцева (ИСП РАН); модератор П. Ганелин.
- [27] Цымлов А., Филатов В. Агенты пишут код, а кто его проверяет? Вебинар 3 цикла «Качество и безопасность ПО в эпоху GenAI», 28.08.2026 (выступления и презентация).
- [28] Круглый стол «Реальное состояние рынка безопасной разработки (РБПО)». SOC Forum 2026, 28.10.2026. Выступления А. Башарина (Swordfish Security), А. Смирнова (CodeScoring); модератор В. Падарян (ИСП РАН).
- [29] ИИ-митап ЦГУ, 06.11.2026: доклад А. Морозова «Безопасность в AI и AI в безопасности»; выступление представителя банка о внедрении ИИ-агента второй линии поддержки.
- [30] Пиков В. А. LLMs – ваш цифровой помощник в преподавании. Вебинар для преподавателей, 17.09.2025 (презентация).

# Практические руководства

- [21] Мокевнин К. Agentic Design Patterns. Приложение «Фразы для AGENTS.md» (книга о паттернах агентного программирования, при поддержке Хекслета и Праксора; дата не указана, страница открыта 11.11.2026). https://mokevnin.github.io/agentic-coding-design-patterns/ru/agents-md-phrases.html

# Учебные материалы

Учебный кейс ЭС-24 составлен для курса. Узел условный, значения подобраны для упражнений.

---

**Автор курса**

Виталий Пиков

- Архитектор информационной безопасности
  • РБПО • DevSecOps • AppSec
- Лектор и автор образовательных программ по ИБ и РБПО

Телефон: +7 (963) 782-92-78 ·
E-mail: vitaly@pikov.expert ·
Сайт: https://pikov.expert ·
Telegram: @UnderLineSecurity