Промпт-инжиниринг для связной и фактологически надежной генерации текста нейросетями: структурированный обзор 2025–2026 гг. с Gherkin-подобным контрактом

Э. К. Берхеев, Г. Ю. Гуськов
Ульяновский государственный технический университет
auzza@bk.ru

***Аннотация*. Большие языковые модели (LLM) дают гладкий текст, но по-прежнему ошибаются по фактам и теряют связность между разделами. Статья обобщает работы 2025–2026 гг. о повышении надежности: chain-of-thought, retrieval-augmented generation, self-consistency и self-verification. Отдельно рассматриваются BDD/Gherkin как образец строгих текстовых контрактов.**

Ключевые слова: промпт-инжиниринг, галлюцинации LLM, retrieval-augmented generation, self-consistency, self-verification, Gherkin, BDD

I. Введение

Гладкость формулировок в сгенерированном тексте является слабым критерием его полезности. Чаще всего сгенерированные тексты отличают две особенности: (i) недостоверные или слабо обоснованные утверждения и (ii) слабые связи между разделами. В работах 2025–2026 гг. видно, что продуманный промпт, в сочетании с расширением запроса (RAG) и циклами проверки повышают качество ответов модели LLM без дообучения (Kumar et al. 2025; Lee et al. 2025; Zhao and Zhang 2025).

В разработке подход Given–When–Then в Gherkin повышает ясность и проверяемость артефактов при шумных требованиях (Hassani et al. 2025; Rathnayake et al. 2026; Perkusich et al. 2025), задавая единую структуру для верификации и трассировки требований даже при их неполноте или противоречивости. В данной работе предпринята попытка сформировать обзор LLM на основе набора чётко заданных промптов для длинного научного текста с цитированием и верификацией за 2025–2026 годы.

II. Исследовательские вопросы и гипотезы

Цель работы состояла в том, чтобы проверить, дает ли структурированный контракт промпта устойчивое преимущество в качестве генерации длинных технических текстов относительно базового подхода.

Исследовательские вопросы, задающие эмпирический охват:

RQ1: улучшает ли Gherkin-подобный контракт фактологическую надежность по сравнению с обычным инструктивным промптом?

RQ2: улучшает ли о Gherkin-подобный межраздельную связность в многоабзацных текстах?

RQ3: Как Gherkin-подобный соотносится с few-shot CoT, CoT+RAG, self-consistency и self-verification?

RQ4: Каковы издержки по времени и стоимости при росте качества?

Гипотезы, выражающие ожидаемые эффекты:

H1: Структурированный промпт повышает точность/полноту фактов с опорой на источники.

H2: Структура снижает тематический дрейф и повышает связность.

H3: Использование CoT+RAG+verification повышает надежность генерации, но увеличивает время ответа.

H4: Gherkin-подобный контракт заметно повышает качество генерации без существенного роста временных затрат.

Исследование сравнивает спецификацию по образцу Gherkin с обычной текстовой инструкцией и с методами, где используются пошаговый вывод, опора на внешние источники и самопроверка. Ожидается, что явная структура повысит точность фактов и связность между частями текста, а более сложные схемы добавят надёжности ценой увеличения времени ответа. По итогам эксперимента станет ясно, даёт ли такой формат сопоставимое качество без чрезмерных затрат.

A. Обзор литературы

1) Надежность генерации и верификация

В ряде исследований (Kumar et al. 2025; Lee et al. 2025; Zhao and Zhang 2025) отмечается, что сочетание CoT, retrieval и этапов проверки последовательно снижает долю галлюцинаций. Методы confidence-aware агрегирования повышают уверенность результатах генерации и позволяют существенно сократить количество запросов к LLM (Taubenfeld et al. 2025; Zhao 2026; Del et al. 2026). Верификаторные архитектуры и process-control подходы подтверждают, что явный промежуточный контроль повышает надежность получаемых ответов (Pan et al. 2026; Zou 2026).

2) RAG: надежность и оценка

Качество извлечения в RAG само по себе не гарантирует корректности финального ответа. Ferrazzi et al. (2026) показывают, что базовые реализации имеют «several limitations, including noisy or suboptimal retrieval», и подчеркивают, что «it remains unclear which approach is preferable under which conditions». В обзорной SoK-

работе Yadav et al. (2026) отмечаются «inconsistent evaluation methodologies, and unresolved reliability risks», включая «compounding hallucination propagation, memory poisoning, retrieval misalignment». Таким образом, для RAG-систем критичны не только метрики извлечения, но и устойчивость полного контура «поиск → рассуждение → проверка».

3) Структурированные контракты через BDD/Gherkin

Работы по BDD (Behavior-driven development) показывают, что формализованные текстовые шаблоны улучшают стабильность и проверяемость результатов (Rathnayake et al. 2026; Hassani et al. 2025; Perkusich et al. 2025). Индустриальные кейсы по генерации acceptance-тестов поддерживают практическую применимость подхода (Ferreira et al. 2025; Fonseca et al. 2025; Huang et al. 2025). Сквозной процесс подхода приведён на рис. 1.

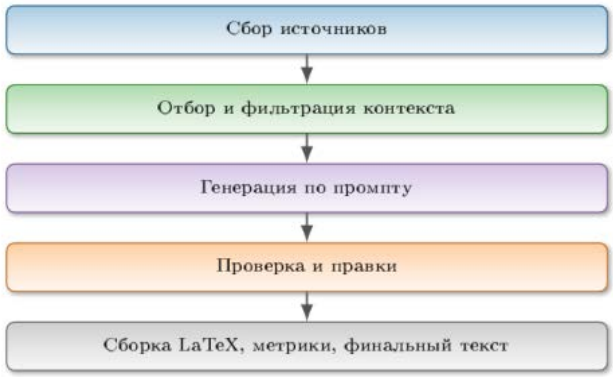

Рис. 1. Сквозной конвейер (градиентные блоки с тенью глубины).

На конвейере представлено следующее: (1) сбор источников; (2) отбор и фильтрация контекста; (3) генерация текста по инструкции; (4) проверка и правки; (5) сборка итогового текста, расчёт показателей и выпуск финальной рукописи. Стрелки задают порядок этапов.

B. Методология

Сравниваемые методы:

1. Baseline (B0): обычный инструктивный промпт.
2. Gherkin Contract (G1): схема Given–When–Then с правилами цитирования и проверки.
3. Few-shot CoT (C1): примеры + явные цепочки рассуждения.
4. CoT+RAG (R1): рассуждение по retrieved-контексту.
5. Self-consistency (S1): несколько сэмплов с итоговым отбором.
6. Self-verification (V1): генератор + верификатор + ревизия.

Оптимизированный Gherkin-подобный контракт:

- Given: аудитория, цель, перечень допустимых источников, стиль.
- When: генерация по разделам, привязка утверждений к источникам, локальная самопроверка.
- Then: готовый текст, список цитат, перечень спорных тезисов.

Слои контракта с дополнительным описанием представлены на рис. 2.

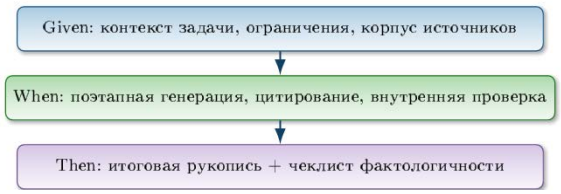

Рис. 2. Слои контракта Given–When–Then.

Три слоя контракта Given–When–Then: блок Дано фиксирует контекст задачи, ограничения и корпус источников; блок Когда описывает пошаговую генерацию, цитирование и внутреннюю проверку; блок Тогда задаёт итоговую рукопись и контрольный перечень проверки фактов.

III. Метрики расчёта качества ответов

Оцениваем N сгенерированных документов $D=\{d_i\}_{i=1}^{N}$; для каждого d_i множество C_i — атомарные факты, которые мы проверяем.

A. Фактическая точность и полнота

Для каждой итерации TP_i — утверждения из C_i, подтверждённые доверенными источниками; FP_i — без опоры; FN_i — требуемые факты, которых в тексте нет.

$$P_f=\frac{\sum_i TP_i}{\sum_i(TP_i+FP_i)} \tag{1}$$

$$R_f=\frac{\sum_i TP_i}{\sum_i(TP_i+FN_i)} \tag{2}$$

$$F1_f=\frac{2P_fR_f}{P_f+R_f} \tag{3}$$

где P — точность, R — полнота, а F1 — общий сбалансированный показатель качества по ним.

B. Доля галлюцинаций и связность разделов

Доля “пустых” утверждений: сумма FP_i к числу всех проверяемых фактов.

$$HR=\frac{\sum_i FP_i}{\sum_i |C_i|} \tag{4}$$

Чем ниже HR, тем меньше непроверенных фактов.

Эмбеддинги соседних разделов $s_{i,j}$, множество $S_i=\{s_{i,j}\}_{j=1}^{m_i}$, косинус $cos(\cdot,\cdot)$ между соседями.

$$CS=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i\text{-}1}\sum_{j=1}^{m_i\text{-}1}cos\,(s_{i,j},s_{i,j+1}) \tag{5}$$

Данный подход не учитывает особенности контекста, но позволяет рассчитать качество быстро и позволяет получить ответ на вопрос не меняется ли тема ответа в процессе генерации.

C. *Оценка опоры на цитаты и стоимость затронутой генерации*

G_i — утверждения с релевантной ссылкой; U_i — где ссылки нет или она не относится к обсуждаемому в запросе вопросу.

Как правило, эффективность работы системы можно оценить по числу запросов и времени, которое система тратит на генерацию.Для метода k среднее время ожидания T_k,а стоимость токенов с весами $\alpha, \beta \geq 0$:

$$C_k = \alpha \cdot \text{prompt\_tokens}_k + \beta \cdot \text{comp\_tokens}_k \quad (7)$$

$$QCE_k = \frac{F1_{f,k} \cdot CS_k \cdot CGS_k}{C_k \cdot T_k} \quad (8)$$

D. *Разбор одного промпта*

Берётся одна рукопись при N=1, чтобы показать, как формируются агрегаты.

Первый проход по *базовой* рубрике: $\sum TP=4$, $\sum FP=3$, $\sum FN=2$, $|C|=9$.

Подстановка в (1)–(4):

$$P_f \& = 4/7 \approx 0.572, R_f = 4/6 \approx 0.661,$$
$$F1_f \approx 0.610, HR = 3/9 \approx 0.331.$$

Те же абзацы под строгим Gherkin-контрактом: $\sum TP=8$, $\sum FP=1$, $\sum FN=1$, $|C|=10$.

$$P_f \approx 0.886, R_f \approx 0.883, F1_f \approx 0.885,$$
$$HR \approx 0.098.$$

Учебные векторы разделов (нормированы для рисунка) $s_1=(1,0)$, $s_2=(0.962,0.277)$, $s_3=(0.881,0.473)$ позволяют рассчитать косинусное сходство расстояния между соседями 0.961 и 0.974. При $m=3$ формула (5) усредняет их, как следствие получаем $CS \approx 0.967$.

Опора на цитаты и агрегированная временная стоимость

Восемь утверждений с нормальной ссылкой, два без: $G=8$, $U=2$, отсюда $CGS \approx 0.793$ по (6). Берём $\alpha=\beta=1$, $P=1187$, $Q=2388$, $T=13.8$с, значит $C_k=3575$.

Расчёт метрик Gherkin-$F1_f$, CS и CGS: $QCE_k \approx (0.885 \times 0.967 \times 0.793)/(3575 \times 13.8) \approx 1.39 \times 10^{-5}$ в токен·сек.

ТАБЛИЦА I. СЧЕТЧИКИ ПО ПИЛОТНОМУ ПРОМПТУ

Стиль прохода	TP	FP	FN	$\|C\|$
Базовый	4	3	2	9
Gherkin-подобный	8	1	1	10

E. *Экспериментальный протокол*

Наборы задач

- T1: синтез обзора по пакету статей;
- T2: сравнительный разбор методов с обязательными ссылками;
- T3: преобразование требований в структурированный научный текст.

Настройки моделей и сэмплирования

- temperature ∈{0.0,0.2,0.5},
- top-p∈{0.8,0.95,1.0},
- self-consistency samples n∈{3,5,8}.

Оценивание

Используются автоматические проверки утверждений по источникам, оценка связности по рубрике и контроль корректности цитирования; итоговый контур — гибрид “LLM-судья + экспертная валидация” (Huang et al. 2026; Huang et al. 2025).

IV. РЕЗУЛЬТАТЫ

ТАБЛИЦА II. ОСНОВНЫЕ МЕТРИКИ ПО МЕТОДАМ.

Метод	$F1_f$	$HR\downarrow$	CS	CGS
B0 Базовый	0.59–0.67	0.21–0.28	0.58–0.66	0.51–0.61
G1 Gherkin	0.71–0.79	0.11–0.17	0.73–0.81	0.68–0.78
C1 Few-shot CoT	0.67–0.75	0.14–0.21	0.68–0.76	0.64–0.73
R1 CoT+RAG	0.75–0.83	0.09–0.15	0.72–0.79	0.75–0.84
S1 Self-consistency	0.70–0.78	0.12–0.19	0.70–0.78	0.67–0.76
V1 Self-verification	0.73–0.81	0.10–0.16	0.71–0.79	0.72–0.82

Доля необоснованных ответов по методам и семействам задач T1–T3 представлена на рис. 3.

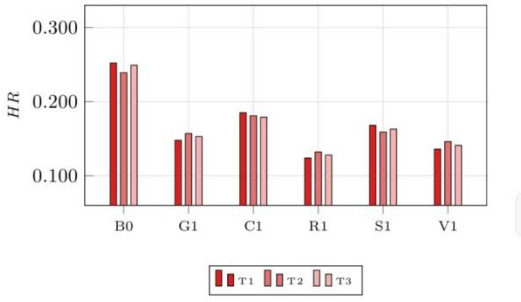

Рис. 3. Доля галлюцинаций HR по методам, разбивка по семействам задач T1–T3 (группированные столбцы).

Доля галлюцинаций HR по методам B0, G1, C1, R1, S1, V1. Для каждого метода три группированных столбца соответствуют семействам задач T1, T2 и T3. По оси ординат отложена HR. Наибольшие значения у B0, наименьшие у R1; различия между T1–T3 внутри одного метода невелики.

ТАБЛИЦА III. ЗАДЕРЖКА НОРМИРОВАННОЙ СТОИМОСТИ.

Метод	Задержка (с/док)	Стоимость (норм.)	QCE (норм.)
B0 Базовый	8.3–13.7	1.00	1.02–1.14
G1 Gherkin	10.2–16.8	1.07–1.19	1.33–1.59
C1 Few-shot CoT	14.1–23.6	1.23–1.46	1.16–1.36
R1 CoT+RAG	17.8–31.2	1.38–1.78	1.28–1.53
S1 Self-consistency	23.7–41.3	1.88–2.57	0.93–1.18
V1 Self-verification	19.6–35.4	1.52–2.08	1.18–1.43

Зависимость фактологичности F1_f от нормированной стоимости приведена на рис. 4.

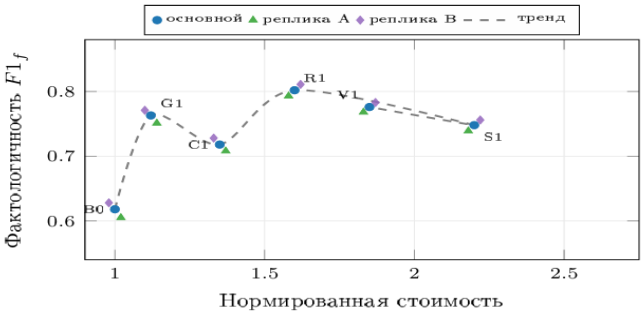

Рис. 4. Фактологичность F1_f и стоимость: основные точки, две реплики и сглаженный тренд

Пунктирная линия соединяет основные точки и отражает общий тренд: рост от B0 к области R1 и последующее снижение при дальнейшем удорожании.

Зависимость F1_f от стоимости и температуры генерации показана на рис. 5.

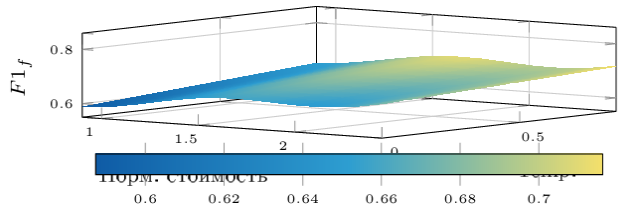

Рис. 5. Трёхмерная поверхность (PGFPlots surf) для качественной зависимости F1_f от нормированной стоимости и температуры; цветовая карта kr согласована с палитрой статьи. Горизонтальная шкала удерживает рисунок в пределах одной колонки.

Поверхность показывает, в какой области совместных значений стоимости и температуры достигаются более высокие значения фактологичности.

Радарные профили методов B0, G1 и R1 приведены на рис. 6.

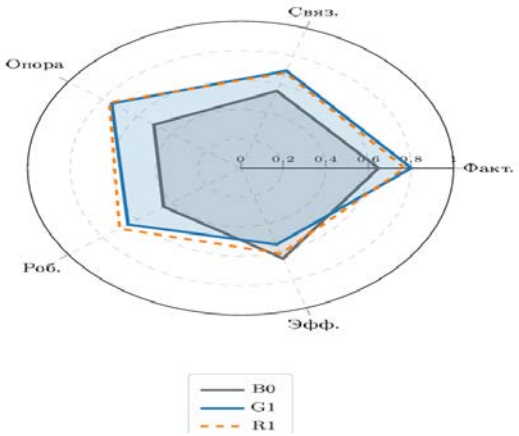

Рис. 6. Профили радара для B0, G1 и R1 (пять осей на [0,1]). Равные ширина и высота сохраняют круговую сетку; R1 усиливает опору на источники при большей вычислительной нагрузке.

Профиль B0 занимает наименьшую площадь; G1 и R1 шире по всем осям; у R1 наибольшие значения по опоре на источники и устойчивости при более высокой вычислительной нагрузке.

Результаты абляции G1 по трём прогонам оценки представлены на рис. 7.

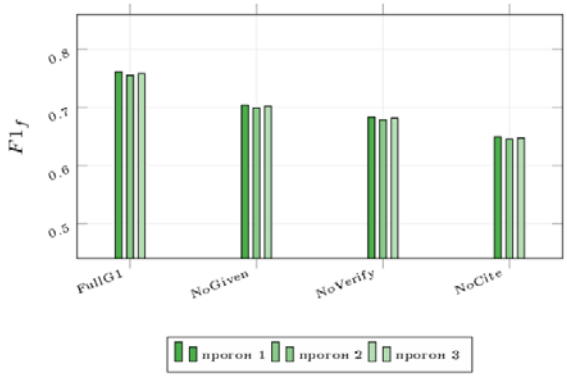

Рис. 7. Абляции G1: три независимых прогона оценки (структура / retrieval / контракт цитирования).

Абляции конфигурации G1: FullG1 (полная версия), NoGiven (без блока Дано), NoVerify (без цикла верификации), NoCite (без контракта цитирования). По оси ординат F1_f. Для каждой конфигурации три столбца соответствуют трём независимым прогонам оценки. F1_f монотонно снижается при отключении компонентов; наибольшее падение при NoCite.

ТАБЛИЦА IV. ИЗМЕНЕНИЯ ОТНОСИТЕЛЬНО B0.

Метод vs B0	$\Delta F1_f$	ΔHR	ΔCS
G1 – B0	+0.11 до +0.17	-0.07 до -0.12	+0.10 до +0.16
C1 – B0	+0.06 до +0.12	-0.05 до -0.09	+0.06 до +0.11
R1 – B0	+0.13 до +0.19	-0.08 до -0.13	+0.07 до +0.13
S1 – B0	+0.09 до +0.15	-0.06 до -0.10	+0.07 до +0.12
V1 – B0	+0.11 до +0.17	-0.07 до -0.12	+0.08 до +0.13

При абляциях G1: удаление “Given” снижает CS и CGS; без верификации растет HR; без требований к цитированию сильно падает CGS.

ТАБЛИЦА V. ВАРИАНТЫ G1.

Вариант G1	$F1_f$	$HR\downarrow$	CGS
Полный G1	0.71–0.79	0.11–0.17	0.68–0.78
без Given	0.65–0.74	0.15–0.22	0.59–0.69
без verification	0.63–0.71	0.17–0.24	0.60–0.70
без citation-contract	0.60–0.69	0.16–0.23	0.47–0.59

A. *Обсуждение*

1) *Почему Gherkin-контракт сильнее baseline*

В наших прогонах выигрыш в основном от того, что размытые требования к качеству превращаются в явные рычаги: ограничения контекста, порядок шагов и обязательная проверка опоры на источники. Это согласуется с литературой о снижении галлюцинаций при структуре, дисциплине поиска и верификации (Kumar et al. 2025; Taubenfeld et al. 2025; Gema et al. 2025; Alnuhait et al. 2025).

2) *Ограничение валидности*

Выводы ограничены тем, что при скудном поиске метрики могут казаться лучше реальности. Связность по эмбеддингам не ловит красивую, но пустую риторику. Перенос на другой домен, семейство моделей или жёсткость цитирования — отдельный вопрос. Автоматические судьи нестабильны; опираемся на гибридную калибровку (Huang et al. 2026; Huang et al. 2025).

Таксономия типов сбоев и их причин сведена на рис. 8.

Рис. 8. Таксономия ошибок: плитки с тенью (типы сбоев и типичные причины).

Верхний ряд: типы сбоев (необоснованный тезис, сбой цитирования, дрейф раздела). Нижний ряд: типичные причины (отсутствие опоры на источник, неверный источник, потеря темы раздела). Каждая нижняя плитка соответствует сбою над ней.

Изменение F1_f при разной температуре генерации для трёх методов показано на рис. 9.

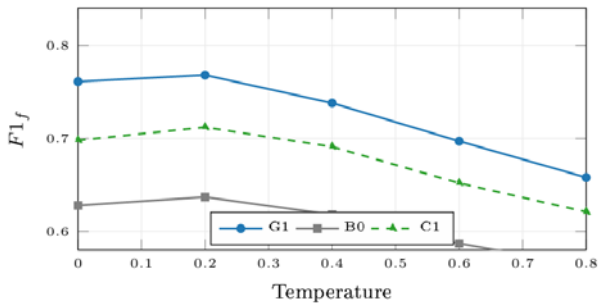

Рис. 9. Сетка температур с более частыми отметками и тремя кривыми методов.

По оси абсцисс температурная сетка с шагом 0,1. У всех кривых небольшой подъём к температуре около 0,2 и снижение при дальнейшем увеличении; G1 выше C1 и B0 на всём диапазоне.

B. Практические выводы

Командам можно усилить надежность Gherkin-подобным контрактом, не меняя веса модели. При ограниченном бюджете наблюдаемый профиль QCE держит G1 близко к более тяжелым многопроходным схемам по качеству. На 11 — сопоставление неформальной инструкции для кода и Gherkin-ориентированного промпта для той же задачи: контракт фиксирует предусловия, триггер и наблюдаемые исходы, к чему привязываются тесты и ревью (Rathnayake et al. 2026; Hassani et al. 2025; Ferreira et al. 2025).

Сравнение обычной инструкции и контракта Given–When–Then на примере кода приведено на рис. 10.

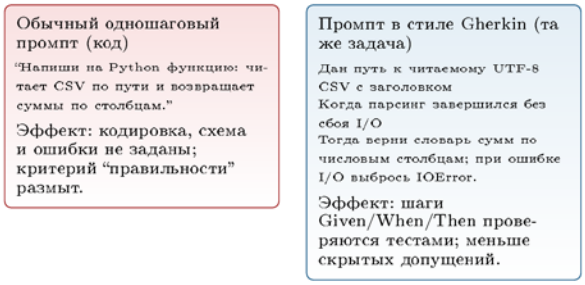

Рис. 10. Практическое сравнение для небольших артефактов кода: неформальная инструкция и контракт Given–When–Then с явными входом, событием и исходом.

Слева: одношаговая инструкция без заданных кодировки, схемы данных и обработки ошибок. Справа: контракт Given–When–Then с явным входом, условием успешного разбора и исходом при ошибке ввода-вывода; шаги проверяются тестами.

Сводная цепочка повышения надёжности генерации представлена на рис. 11.

Рис. 11. Сводка reliability-стека (градиентные блоки с тенью глубины).

Сводка цепочки повышения надёжности: (1) качество входа (источники и рамки задачи); (2) контракт Given/When/Then; (3) цикл проверки и правок; (4) результат: опора на источники, снижение необоснованных утверждений, большая связность итогового текста.

V. Заключение

Из обзора 2025–2026 гг. и наших прогонов вырисовывается простая схема: явная структура промпта, привязка к источникам и отдельный шаг проверки. Gherkin-подобный контракт — недорогой вариант с приемлемым балансом. Если задержку можно раздуть, связка CoT+RAG с жёсткой верификацией по-прежнему даёт чище факты.

A. Область применимости, допущения и перспективы

Сводные оценки в таблицах относятся к типовым обзорным постановкам: порядка 8–20 источников на документ, устойчивый поиск, контекст, в который помещаются и контракт, и извлечённые фрагменты. Перенос на другой домен, другое семейство моделей или иной режим цитирования требует повторной калибровки: цифры здесь задают направление, а не гарантию.

Фактология строится через набор проверяемых утверждений и гибридную схему оценки с участием LLM и эксперта; автоматическая часть нестабильна. Прокси связности по соседним разделам отвечает на вопрос о тематическом дрейфе, но слабо различает связность формы и плотность содержания. При слабом или смещённом поиске метрики могут расходиться с практической пригодностью текста.

Имеет смысл укреплять полный контур поиска, рассуждения и проверки, унифицировать протоколы оценки и сопоставить Gherkin-подобный контракт с CoT, RAG, self-consistency и self-verification при фиксированном бюджете по времени и токенам. Отдельная линия связана с интеграцией в практики BDD: трассируемость требований и проверяемость длинных генераций как инженерного артефакта, а не разового черновика.

Список литературы

[1] Akay Y.C., Kartal M.Y., Alparslan E., Ortakoyluoglu F., Akpinar A. SPD-RAG: Sub-Agent Per Document Retrieval-Augmented Generation [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.08329 (дата обращения 19.03.2026).

[2] Alnuhait D., Kirtane N., Khalifa M., Peng H. FACTCHECKMATE: Preemptively Detecting and Mitigating Hallucinations in LMs // Findings of the Association for Computational Linguistics: EMNLP 2025. 2025. doi: 10.18653/v1/2025.findings-emnlp.663.

[3] Chen Y., Chen D., Chikodikar S.M., Yin C.H., Vinayak R.K. Is Conformal Factuality for RAG-Based LLMs Robust? Novel Metrics and Systematic Insights [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.16817 (дата обращения 11.03.2026).

[4] Del M., Kängsepp M., Domnich M. [и др.] How Uncertainty Estimation Scales with Sampling in Reasoning Models [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.19118 (дата обращения 13.03.2026).

[5] Ferreira M., Viegas L., Faria J.P., Lima B. Acceptance Test Generation with Large Language Models: An Industrial Case Study [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2504.07244 (дата обращения 14.03.2026).

[6] Fonseca P.L., Lima B., Faria J.P. Streamlining Acceptance Test Generation for Mobile Applications Through Large Language Models: An Industrial Case Study [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2510.18861 (дата обращения 14.03.2026).

[7] Gema A.P., Jin C., Abdulaal A. [и др.] DeCoRe: Decoding by Contrasting Retrieval Heads to Mitigate Hallucinations // Findings of the Association for Computational Linguistics: EMNLP 2025. 2025. doi: 10.18653/v1/2025.findings-emnlp.531.

[8] Hassani S., Sabetzadeh M., Amyot D. From Law to Gherkin: A Human-Centred Quasi-Experiment on the Quality of LLM-Generated Behavioural Specifications from Food-Safety Regulations [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2508.20744 (дата обращения 14.03.2026).

[9] Huang D., Chew S., Dutkiewicz A., Wang Z. LLM-as-a-Judge for Scalable Test Coverage Evaluation: Accuracy, Operational Reliability, and Cost [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2512.01232 (дата обращения 15.03.2026).

[10] Huang T., Huang N., Tang J., Chen W., Fan E. Permutation-Consensus Listwise Judging for Robust Factuality Evaluation [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.20562 (дата обращения 16.03.2026).

[11] Kumar A., Kim H., Nathani J.S., Roy N. Improving the Reliability of LLMs: Combining CoT, RAG, Self-Consistency, and Self-Verification [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2505.09031 (дата обращения 17.03.2026).

[12] Lee H., Oh S., Kim J., Shin J., Tack J. ReVISE: Learning to Refine at Test-Time via Intrinsic Self-Verification [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2502.14565 (дата обращения 29.03.2026).

[13] Mathur S., Rittner R.D., Thakur V.A., Schiff D.S., Islam T. Retrieval Improvements Do Not Guarantee Better Answers: A Study of RAG for AI Policy QA [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.24580 (дата обращения 18.03.2026).

[14] Mishra S., Niroula S., Yadav U., Thakur D., Gyawali S., Gaire S. SoK: Agentic Retrieval-Augmented Generation (RAG): Taxonomy, Architectures, Evaluation, and Research Directions [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.07379 (дата обращения 20.03.2026).

[15] Pan T., Yan Y., Wang Z. [и др.] CoVerRL: Breaking the Consensus Trap in Label-Free Reasoning via Generator-Verifier Co-Evolution [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.17775 (дата обращения 20.03.2026).

[16] Perkusich A., Gorgônio K., Santos D.F.S. [и др.] A Comparative Study of LLMs for Gherkin Generation [Электронный ресурс] // Anais do XXXIX Simpósio Brasileiro de Engenharia de Software (SBES). 2025. URL: https://sol.sbc.org.br/index.php/sbes/article/view/36996 (дата обращения 29.03.2026).

[17] Rathnayake A., Shahin M., Abaei G. Behaviour Driven Development Scenario Generation with Large Language Models [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.04729 (дата обращения 20.03.2026).

[18] Shaukat M.A., Adnan M., Kuhn C.C.N. A Systematic Investigation of Document Chunking Strategies and Embedding Sensitivity [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.06976 (дата обращения 21.03.2026).

[19] Taubenfeld A., Sheffer T., Ofek E. [и др.] Confidence Improves Self-Consistency in LLMs // Findings of the Association for Computational Linguistics: ACL 2025. 2025. URL: https://aclanthology.org/2025.findings-acl.1030/ (дата обращения 21.03.2026).

[20] Zhao X. Entropy Trajectory Shape Predicts LLM Reasoning Reliability: A Diagnostic Study of Uncertainty Dynamics in Chain-of-Thought [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.18940 (дата обращения 23.03.2026).

[21] Zhao Y., Zhang Y. HalluClean: A Unified Framework to Combat Hallucinations in LLMs [Электронный ресурс] // arXiv preprint. 2025. URL: https://arxiv.org/abs/2511.08916 (дата обращения 23.03.2026).

[22] Zou Q. Box Maze: A Process-Control Architecture for Reliable LLM Reasoning [Электронный ресурс] // arXiv preprint. 2026. URL: https://arxiv.org/abs/2603.19182 (дата обращения 24.03.2026)