УДК 004.896:515.1

Мухамадиева Кибриё Баходировна

Преподаватель кафедры «Современные информационные технологии и искусственный интеллект»

Узбекский государственный университет мировых языков

АВТОНОМНОЕ САМОВОССТАНОВЛЕНИЕ LLM-АГЕНТОВ ЧЕРЕЗ ТОПОЛОГИЧЕСКУЮ ФИЛЬТРАЦИЮ ГАЛЛЮЦИНАЦИЙ

Аннотация: Фундаментальным ограничением применения больших языковых моделей (LLM) в высоконагруженных вычислительных и аналитических средах остается их неустранимая склонность к логическим галлюцинациям и феномен алгоритмической «слепоты к собственным ошибкам». В данной статье предлагается новая когнитивная архитектура полностью автономного восстановления, исключающая оператора-человека из контура корректировки. Методологической основой выступает интеграция строгой иерархии специализированных ИИ-агентов, где стохастическая генерация гипотез жестко контролируется детерминированными механизмами проверки. Представлен инновационный механизм перехвата искажений, базирующийся на трансляции свободных рассуждений модели в абстрактные синтаксические деревья (AST) с их последующей топологической фильтрацией в изолированной программной песочнице. Предложенная архитектура обеспечивает смысловую и фактологическую достоверность, критически необходимую для адаптивного обучения, глубокого морфологического анализа и интеграции ИИ в гуманитарные и филологические дисциплины.

Ключевые слова: Большие языковые модели (LLM), мультиагентные системы, топологическая фильтрация, абстрактные синтаксические деревья (AST), автономная корректировка (self-healing), галлюцинации нейросетей, детерминированная валидация, когнитивные архитектуры.

UDC 004.896:515.1

Mukhamadieva Kibriyo Bakhodirovna

Lecturer at the Department of Modern Information Technologies and Artificial Intelligence, Uzbek State World Languages University

AUTONOMOUS SELF-HEALING OF LLM AGENTS THROUGH TOPOLOGICAL FILTERING OF HALLUCINATIONS

Abstract: A fundamental limitation of applying Large Language Models (LLMs) in high-load computational and analytical environments remains their inherent tendency toward logical hallucinations and the phenomenon of algorithmic "blindness to their own errors." This paper proposes a novel cognitive architecture for fully autonomous recovery, eliminating the human operator from the correction loop. The methodological foundation lies in the integration of a strict hierarchy of specialized AI agents, where the stochastic generation of hypotheses is tightly controlled by deterministic verification mechanisms. An innovative distortion interception mechanism is presented, based on translating the model's free-form reasoning into Abstract Syntax Trees (ASTs) with their subsequent topological filtering within an isolated software sandbox. The proposed architecture ensures the semantic and factual accuracy critically required for adaptive learning, deep morphological analysis, and the integration of AI into the humanities and philological disciplines.

Keywords: Large Language Models (LLMs), multi-agent systems, topological filtering, Abstract Syntax Trees (AST), autonomous self-healing, neural network hallucinations, deterministic validation, cognitive architectures.

Введение

Стремительное развитие больших языковых моделей (LLM) открыло новые горизонты в области искусственного интеллекта, продемонстрировав выдающиеся способности к генерации текста, анализу данных и решению

сложных когнитивных задач. Однако по мере интеграции этих моделей в критически важные процессы исследователи столкнулись с фундаментальным ограничением, заложенным в самой стохастической природе трансформерных архитектур - неустранимой склонностью к смысловым и логическим галлюцинациям. Генерируя синтаксически безупречные, но фактически или алгоритмически неверные конструкции, базовые модели демонстрируют выраженный феномен «слепоты к собственным ошибкам». В условиях, когда нейросеть не способна самостоятельно отличить достоверный факт от статистически правдоподобного вымысла, применение монолитных LLM в сложных аналитических и вычислительных средах становится источником непредсказуемых рисков.

Ответом на этот методологический вызов становится отказ от монолитных структур в пользу интеграции мультиагентных систем, где когнитивная нагрузка распределяется внутри организованной иерархии специализированных LLM [1]. Вместо попыток обучить одну универсальную и безошибочную модель, передовой подход предполагает оркестрацию ансамбля агентов, разделяющих между собой функции генерации, валидации и критики. В такой экосистеме стохастическая свобода «творческого» агента, необходимая для генерации гипотез, жестко уравновешивается детерминированными механизмами проверяющих алгоритмов. Подобное разделение ролей позволяет выстроить замкнутый цикл внутреннего контроля, где система самостоятельно фиксирует противоречие, анализирует причину сбоя и инициирует протокол алгоритмического восстановления.

Именно процесс разработки и математического обоснования такого автономного контура корректировки составляет ядро данного исследования. В основе предлагаемой концепции лежит использование строгих методов парсинга - в частности, трансляции рассуждений агента в абстрактные синтаксические деревья (AST) - для автоматической топологической фильтрации ошибок в

изолированной среде. Когда агент-валидатор обнаруживает логический разрыв или выдуманный факт, он автоматически формирует trace-лог ошибки и передает его агенту-критику. Последний синтезирует корректирующий вектор, который принудительно возвращает генератора в рамки объективной логики. Модель, столкнувшись с непреодолимым программным ограничением собственной песочницы, динамически перестраивает контекст и «выздоравливает» от галлюцинации, переписывая исключительно сбойный участок рассуждений.

Целью работы является обзор современных подходов к автономной коррекции и верификации рассуждений больших языковых моделей и формирование на этой основе концептуальной архитектуры, объединяющей генерацию, структурную верификацию и локальное восстановление. Реализация такого подхода не только радикально повышает фактологическую точность предикта следующего шага, но и открывает путь к созданию саморегулирующихся, оптимизированных ИИ-систем, способных к непрерывному автономному функционированию в условиях высокой неопределенности.

Обзор литературы

Формирование теоретической базы для решения проблемы логических искажений опирается на фундаментальные этапы развития когнитивных архитектур, задокументированные в ведущих академических трудах последних лет. Ранние попытки минимизировать галлюцинации базировались на алгоритмах обучения с подкреплением на основе отзывов людей, что было детально описано в основополагающей работе Ouyang et al. (2022) при создании семейства моделей InstructGPT [1]. Хотя этот метод позволил нейросетям лучше улавливать контекст пользовательских инструкций, исследователи быстро признали его ограниченность при масштабировании на сложные аналитические задачи. В ответ на это научное сообщество предложило заставить модель разбивать задачу на промежуточные этапы: в исследовании Wei et al. (2022) была формализована

концепция «цепочек рассуждений» (Chain-of-Thought) [2]. Это значительно улучшило качество ответов, однако не решило проблему каскадных вымыслов: если модель допускала ошибку на первом шаге, вся последующая цепочка неизбежно приводила к ложному, но грамматически убедительному результату.

Стремясь автоматизировать процесс проверки без участия человека, следующие поколения исследователей сфокусировались на методах внутреннего самосовершенствования. В работах Shinn et al. (2023), предложивших архитектуру вербального подкрепления Reflexion [3], и в исследовании Madaan et al. (2023), описавшем алгоритм итеративного самосовершенствования Self-Refine [4], предполагалось, что генеративная сеть способна выступать собственным критиком, переписывая свои ответы на основе системных инструкций. Однако масштабное критическое исследование Huang et al. (2023), показательно озаглавленное «Большие языковые модели пока не способны самостоятельно корректировать рассуждения», обсуждение опровергло эту гипотезу. На основании анализа существующих подходов предлагается концептуальная модель, утверждающая что монолитные архитектуры подвержены системной слепоте к собственным ошибкам: оценивая свой текст, модель использует те же искаженные вероятностные веса, что и при его генерации. В результате рефлексия вырождается в цикличную конфабуляцию [5].

Выявленные алгоритмические тупики стимулировали переход к многоагентным парадигмам, вдохновленным теориями распределенного когнитивного труда. Знаковым прорывом стали эксперименты Park et al. (2023), где автономные генеративные агенты, наделенные модулями памяти и планирования, взаимодействовали в виртуальной среде, демонстрируя сложное эмерджентное поведение [6]. Развивая идею функционального разделения, Du et al. (2023) исследовали системы многоагентных дебатов, в которых агенты с разными системными установками критиковали ответы друг друга [7]. Эта ролевая изоляция снизила процент смысловых ошибок, так как критикующий

алгоритм не был скован первичным контекстом создателя. Тем не менее, актуальный анализ показывает, что подобные системы все еще полагаются на вероятностные текстовые взаимодействия. В сложных задачах агенты периодически впадают в состояние «коллективной галлюцинации», плавно убеждая друг друга в достоверности ошибочного решения через стохастический диалог.

Анализ современной литературы выявляет существенный пробел на стыке нейросетевой генерации и строгих детерминированных проверок, применяемых в классической программной инженерии. Важным шагом в этом направлении стала работа Gao et al. (2023) о программно-ассистируемых моделях (Program-Aided Language Models), где нейросети впервые начали делегировать логические вычисления интерпретаторам кода [8]. Однако перенос принципов непрерывной интеграции и автоматизированного восстановления логики в когнитивную плоскость искусственного интеллекта остается малоизученным.

Методология

Архитектурный фундамент предлагаемой методологии строится на адаптации принципов непрерывной интеграции и автоматического восстановления пайплайнов, традиционно применяемых в программной инженерии, к когнитивным процессам нейросетевых систем [4]. Центральная гипотеза исследования заключается в том, что стохастическая природа больших языковых моделей может быть эффективно стабилизирована путем внедрения строгой функциональной иерархии вычислительных агентов. В рамках разрабатываемого фреймворка когнитивная задача не передается единому монолитному алгоритму, а маршрутизируется через трехуровневую систему, состоящую из агента-генератора, детерминированного валидатора и агента-критика. Подобная ролевая изоляция гарантирует отсутствие конфликта интересов: модель, выдвигающая первоначальную гипотезу, лишается

возможности самостоятельно подтверждать ее истинность, уступая эту функцию математически строгим алгоритмам проверки (Рисунок 1.).

Ключевым механизмом перехвата и локализации логических искажений выступает алгоритм топологической фильтрации, базирующийся на трансляции естественного языка в структурированные вычислительные графы. На первом этапе агент-генератор, работающий с высокой степенью параметрической свободы, формирует первичное текстовое решение задачи. Инновационность подхода заключается в том, что этот вывод не принимается как финальный результат, а подвергается глубокому семантическому парсингу, в ходе которого свободные рассуждения преобразуются в абстрактное синтаксическое дерево. Эта процедура позволяет разложить расплывчатую языковую семантику на четкие алгоритмические узлы: логические операторы, декларируемые факты, переменные и хронологические зависимости. Переход от неструктурированного текста к графовой топологии создает математически измеримую среду, где каждый шаг рассуждения предстает в виде отдельного функционального блока, поддающегося строгой алгоритмической оценке.

Формальная математическая модель топологической фильтрации

Процесс автономной корректировки моделируется как дискретный оптимизационный процесс в пространстве синтаксических деревьев.

Определение 1. Трансляция семантики в топологию графа

Пусть $R_i \in \Sigma^i$ - сгенерированный агентом-генератором текст на итерации i. Определим функцию семантического парсинга f_{parse}, которая преобразует неструктурированный текст в направленный ациклический граф. AST - это дерево, а дерево является частным случаем DAG

$$G_i(V, E) = f_{parse}(R_i)$$

где $V = \{v_1, v_2, \ldots, v_n\}$ - множество узлов, представляющих элементарные утверждения, переменные или логические операторы, а

$E \subseteq V \times V$ - множество ребер, задающих причинно-следственные связи.

Определение 2. Детерминированная валидация в изолированной среде

Вводится индикаторная функция $\Phi: V \to \{0,1\}$, выполняемая внутри программной песочницы. Данная функция действует как строгий фильтр топологии:

$$\Phi(v_k) = \begin{cases} 1, \text{если } v_k \text{ валиден (не противоречит объективной базе знаний)} \\ 2, \text{если } v_k \text{ является смысловой галлюцинацией или логическим сбоем} \end{cases}$$

Состояние всего графа рассуждений G_i считается валидным тогда и только тогда, когда истинны все его узлы:

$$\Phi(G_i) = \prod_{k=1}^{|V|} \Phi(v_k)$$

Определение 3. Вектор ошибки и рефлексивная обратная связь

Если $\exists\, v_k \in V : \Phi(v_k) = 0$, среда прерывает обход графа и генерирует лог ошибки $e_k = \{v_k, traceback\}$. Агент-критик применяет функцию трансформации $g_{critic}(e_k)$ для синтеза корректирующего вектора ΔC_i. Следующее когнитивное состояние генератора определяется через функцию обновления контекста:

$$R_{i+1} = ¿\, M_{gen}(Q \oplus C_0 \oplus \Delta C_i)$$

где Q - исходная задача, а $\oplus$ - оператор конкатенации контекста.

Гипотеза о сходимости контура автономного восстановления

При условии использования изолированного детерминированного валидатора Φ и наличии жесткого лимита итераций I_{max}, процесс топологической фильтрации AST строго предотвращает распространение каскадных галлюцинаций, гарантированно сходясь либо к валидному графу $\Phi(G_m) = 1$ за $m \le I_{max}$ шагов, либо переходя в безопасное состояние вычислительного тупика.

Доказательство:

1. Рассмотрим стандартную монолитную языковую модель (baseline). Ее процесс саморефлексии описывается марковской цепью, где вероятность перехода в состояние генерации следующего ложного факта зависит от предыдущего ложного факта: $P(R_{i+1}^{hallucination} \mid R_i^{hallucination}) > \delta$. Так как модель использует

одни и те же веса для генерации и проверки, δ 1, что формирует замкнутый цикл конфабуляций (эхо-камеру) [5].

2. В предложенной архитектуре вводится внешний детерминированный разрыв. Если на шаге i генерируется граф G_i с ложным узлом v_k, песочница фиксирует $\Phi(v_k)=0$.

3. Вектор обратной связи ΔC_i действует как жесткое ограничение в латентном пространстве генератора, пенализируя вероятность повторного создания узла v_k. Следовательно, вероятность генерации идентичной топологической ошибки на следующем шаге стремится к нулю: $P\left(f_{parse}^{-1}(v_k) \mid Q \oplus \Delta C_i\right)<\epsilon$, где ϵ - пренебрежимо малая величина.

4. Поскольку размерность пространства возможных логических шагов для конкретной задачи ограничена, а обратная связь монотонно сужает множество допустимых вариантов, последовательность графов $\{G_0, G_1, \ldots\}$ направленно эволюционирует.

5. Наличие внешнего прерывателя I_{max} гарантирует конечность процесса. Таким образом, цепочка рассуждений либо достигает состояния $\Phi(G_m)=1$, либо принудительно обрывается. Бесконечная генерация каскадных галлюцинаций становится математически невозможной.

Алгоритм 1. Теоретическая модель автономного самовосстановления **Вход:** Q (задача), I_{max} (лимит итераций). **Выход:** Валидированный ответ R_{final} или сигнал *Deadlock*.

1: Установить $i=0$, $C_0=Q$

2: Пока $i<I_{max}$:

3: Сгенерировать гипотезу R_i агентом-генератором на основе C_i

4: Сформировать граф $G_i(V, E)=f_{parse}(R_i)$

5: Для каждого узла $v_k \in V$:

6: Если $\Phi(v_k)=¿0$:

7: Сформировать лог ошибки e_k и прервать обход графа

8: Если лог ошибки пуст, вернуть R_i

9: Иначе:

10: Синтезировать вектор ΔC_i агентом-критиком из e_k

11: Обновить контекст генератора $C_{i+1} = C_i \oplus \Delta C_i$

12: $i = i + 1$

13: Вернуть *Deadlock* (предотвращение зацикливания)

ВХОД: КОГНИТИВНАЯ ЗАДАЧА
Q – постановка задачи, условия, контекст
1. АГЕНТ-ГЕНЕРАТОР
Стохастическая генерация первичного решения
• Высокая параметрическая свобода
• Формирование гипотезы и рассуждений
Выход: текстовое решение S_0
2. СЕМАНТИЧЕСКИЙ ПАРСЕР
Глубокий семантический анализ естественного языка
• Выделение сущностей, фактов, отношений
• Преобразование рассуждений в структурированное представление
Выход: семантический граф G_0
3. ПОСТРОЕНИЕ АБСТРАКТНОГО СИНТАКСИЧЕСКОГО ДЕРЕВА (AST)
Трансляция семантического графа в абстрактное синтаксическое дерево (вычислительный граф рассуждений)
• Узлы: факты, операторы, переменные
• Рёбра: логические и причинно-следственные связи
Выход: AST / топологический граф A_0
4. ДЕТЕРМИНИРОВАННЫЙ АГЕНТ-ВАЛИДАТОР
Детерминированная проверка топологии решения
• Проверка фактов с базами данных и источниками знаний
• Проверка логических правил и математических законов
• Проверка причинно-следственных связей и условий задачи
ПРОВЕРЯЕМЫЕ АСПЕКТЫ
Существование объектов и фактов
Логическая непротиворечивость
Причинно-следственные зависимости
Математическая корректность
Соответствие условиям задачи
РЕЗУЛЬТАТ ПРОВЕРКИ
ОШИБОК НЕ ОБНАРУЖЕНО
Решение верифицировано
ОШИБКИ ОБНАРУЖЕНЫ
Формируется отчёт об ошибке E
ФИНАЛЬНОЕ РЕШЕНИЕ
Верифицированное решение S^*
Реконструкция ответа из AST
5. АГЕНТ-КРИТИК
Анализ отчёта об ошибке
• Определение причины сбоя
• Локализация ошибочного узла
• Формирование корректирующего вектора (Δ)
6. ЛОКАЛЬНОЕ ВОССТАНОВЛЕНИЕ
Локальная модификация AST
• Изменяется только ошибочный фрагмент
• Сохранение корректной части рассуждений
• Интеграция корректирующего вектора Δ
ОБНОВЛЁННЫЙ AST
A_{i+1} – новая версия структуры рассуждений
ВСПОМОГАТЕЛЬНЫЕ ИСТОЧНИКИ И ОГРАНИЧЕНИЯ
Базы данных и знания
Математические законы
Логические правила
Условия задачи и ограничения
ПОВТОРНЫЙ ЦИКЛ
до полной верификации

Рис. 1. Архитектура автономного контура корректировки с использованием AST-парсинга

Обсуждения и результаты

Для будущей эмпирической валидации предложенной концептуальной архитектуры предлагается использовать стандартизированные выборки логических задач (например, датасеты уровня HotpotQA). В качестве экспериментального протокола целесообразно использовать связку открытых LLM (например, семейства Llama) в роли агента-генератора и детерминированных песочниц (таких как изолированные Python-контейнеры с модулем AST) в роли валидатора.

Для объективной оценки эффективности предложенного контура по сравнению с традиционными методами (Zero-Shot генерацией и текстовой саморефлексией) предлагается ввести следующие метрики:

- Autonomous Recovery Rate (ARR): доля галлюцинаций, успешно исправленных системой без вмешательства оператора.
- Topological Exact Match (TEM): показатель структурной целостности финального графа рассуждений.
- Average Iterations: среднее количество циклов, необходимых агентам для достижения консенсуса и выхода из цикла ошибок.

Теоретический анализ показывает, что переход к AST-фильтрации должен продемонстрировать фундаментальное превосходство над монолитными моделями. Ожидается, что базовая нейросеть (Zero-Shot), лишенная среды проверки, будет показывать высокий процент каскадных ошибок, где первичная неточность разрушает всю цепочку рассуждений [6]. Метод текстовой саморефлексии (Textual Self-Refine) способен лишь

незначительно улучшить ситуацию из-за системной слепоты модели к собственным ошибкам [8].

В свою очередь, внедрение детерминированного валидатора и механизма трансляции текста в синтаксические деревья должно свести вероятность каскадных галлюцинаций к минимуму. Теоретическая модель предполагает, что локализация исправлений на уровне конкретных блоков графа, а не всего текстового массива, обеспечит беспрецедентную вычислительную эффективность. Системе не потребуется перегенерировать весь объем рассуждений с нуля, что радикально сократит количество итераций, необходимых для достижения окончательной топологической точности.

Дальнейшее развитие методологии автономной корректировки и топологического парсинга открывает широкие перспективы для внедрения подобных саморегулирующихся систем в прикладные высоконагруженные среды. Подобная адаптивная архитектура критически востребована при развертывании инфраструктуры современных лабораторий искусственного интеллекта и проектировании интеллектуальных образовательных платформ. Способность нейросети безошибочно прогнозировать индивидуальные когнитивные траектории и адаптировать учебные сценарии под конкретного пользователя, полностью исключая риск предоставления искаженных фактов, знаменует переход от экспериментальных диалоговых интерфейсов к надежным автономным когнитивным агентам, способным безопасно функционировать в академической и исследовательской практике.

Заключение

Обзор современных подходов показывает, что преодоление смысловых галлюцинаций LLM невозможно в рамках монолитных вероятностных архитектур. Решением этой проблемы выступает концептуальная модель

делегирования когнитивной нагрузки иерархии специализированных агентов. Ожидается, что строгая функциональная субординация процессов генерации, валидации и критики позволит системе автономно распознавать и устранять фактологические искажения без участия человека.

Ключевым механизмом предложенного архитектурного контура является трансляция свободных рассуждений в абстрактные синтаксические деревья (AST). Перевод процесса верификации из плоскости словесных многоагентных дебатов в плоскость детерминированной алгоритмической фильтрации внутри изолированной песочницы создает механизм возвращения нейросети в рамки объективной логики. Локальное исправление исключительно сбойных узлов графа направлено на купирование эффекта каскадных галлюцинаций и обеспечение высокой вычислительной эффективности при оптимизации моделей.

Таким образом, концептуальный переход от вероятностной словесной рефлексии к топологической фильтрации рассуждений знаменует важный эволюционный шаг в развитии когнитивных архитектур. Представленная модель утверждает жизнеспособность полностью автономных и самооптимизирующихся мультиагентных систем, закладывая теоретическую основу для безопасного функционирования ИИ в условиях высокой алгоритмической неопределенности.

Список литературы:

1. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744. https://doi.org/10.48550/arXiv.2203.02155.

2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models.

Advances in Neural Information Processing Systems, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903.

3. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2303.11366.

4. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36.

5. Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). Large language models cannot self-correct reasoning yet. arXiv preprint. https://doi.org/10.48550/arXiv.2310.01798.

6. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. B Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (стр. 1-22). https://doi.org/10.1145/3586183.3606763.

7. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv preprint. https://doi.org/10.48550/arXiv.2305.14325

8. Gao, L., Maman, A., Banerjee, S., Goswami, A., Kamalloo, E., Mackey, L., & Blanco, E. (2023). PAL: Program-aided language models. B International Conference on Machine Learning (стр. 10764-10799). PMLR. https://doi.org/10.48550/arXiv.2211.10435