УДК 004.89+004.82+004.855.5

ДЕТЕКЦИЯ ГАЛЛЮЦИНАЦИЙ НА ОСНОВЕ ВНУТРЕННИХ СОСТОЯНИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Т. Р. Айсин[1] [0009-0001-5863-3252], Т. В. Шамардина[2] [0009-0008-9033-6646]

[1, 2]Московский физико-технический институт, г. Долгопрудный, Московская обл., Россия

[1]aysin.timur@gmail.ru, [2]shamardina.tatiana@gmail.com

Аннотация

В последние годы большие языковые модели (Large Language Models, LLM) достигли значительных успехов в области обработки естественного языка и стали ключевым инструментом для решения широкого спектра прикладных и исследовательских задач. Однако с ростом их масштабов и возможностей все более острой становится проблема галлюцинаций – генерации ложной, недостоверной или несуществующей информации, представленной в достоверной форме. В связи с этим вопросы анализа природы галлюцинаций и разработки методов их выявления приобретают особую научную и практическую значимость.

В работе изучен феномен галлюцинаций в больших языковых моделях, рассмотрены их существующая классификация и возможные причины. На базе модели Flan-T5 также исследованы различия внутренних состоянии модели при генерации галлюцинаций и верных ответов. На основе этих расхождений представлены два способа детектирования галлюцинаций: с помощью карт внимания и скрытых состояний модели. Эти методы протестированы на данных из бенчмарков HaluEval и Shroom 2024 в задачах суммаризации, ответов на вопросы, перефразирования, машинного перевода и генерации определений. Кроме того, исследована переносимость обученных детекторов между различными типами галлюцинаций, что позволило оценить универсальность предложенных методов для различных типов задач.

Ключевые слова*: большие языковые модели, галлюцинации, детекция, Flan-T5, обработка естественного языка, карты внимания, внутренние состояния, HaluEval, Shroom.*

© Т. Р. Айсин, Т. В. Шамардина, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Развитие технологий обработки естественного языка (Natural Language Processing, NLP) в последние годы неразрывно связано с появлением и совершенствованием больших языковых моделей (Large Language Models, LLM), основанных на архитектуре Трансформер (Transformer, [1]). LLM способны моделировать сложные семантические зависимости и демонстрируют высокий уровень генеративных возможностей.

Однако стремительное развитие таких систем выявило и ряд новых проблем, среди которых особое место занимает феномен галлюцинаций [2]. Под галлюцинациями понимаются случаи, когда модель генерирует фактически неверную, вымышленную или логически несогласованную информацию, которая при этом выглядит правдоподобно и убедительно. В результате галлюцинации снижают достоверность создаваемых систем, что критически важно при использовании LLM в сферах, где ошибки недопустимы, например в медицине или праве.

Актуальность настоящей работы обусловлена необходимостью разработки методов детекции галлюцинаций, позволяющих повысить надежность и прозрачность работы языковых моделей. Существующие подходы зачастую ограничиваются анализом выходного текста без учета внутренних механизмов работы модели или используют обращения к внешним источникам данных при генерации. Между тем анализ внутренних представлений, таких как скрытые состояния или карты внимания, может дать дополнительную информацию о том, как модель формирует распределение вероятностей токенов и в какой момент формируется недостоверный контент.

Целью исследования являются разработка и экспериментальная проверка методов детекции галлюцинаций, основанных на внутренних состояниях и механизмах внимания слоев моделей. В рамках работы исследована взаимосвязь между внутренними признаками модели и достоверностью генерируемого текста, а также разработаны способы выявления галлюцинаций без обращения к внешним источникам данных. Эксперименты проведены на открытых датасетах HaluEval [3] и Shroom 2024 [4], что позволило оценить эффективность и переносимость предложенных методов на различных типах задач.

ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Проблема галлюцинаций в больших языковых моделях имеет комплексный характер и обусловлена сочетанием факторов, связанных с данными [5–7], процессом обучения [8, 9] и особенностями генерации текста [10, 11]. В зависимости от характера и степени отклонения от достоверности галлюцинации можно разделить на два основных типа [2]:

1) фактические галлюцинации (factuality): ответы модели расходятся с фактами реального мира, которые можно проверить;

2) верность инструкциям или контексту (faithfulness): ответы модели расходятся с инструкциями пользователя или контекстом.

Детектирование фактических галлюцинаций

Для детектирования фактических галлюцинаций можно использовать пайплайны контроля данных [12], которые выполняют разбиение сгенерированного ответа на отдельные фактические утверждения и осуществляют их автоматизированную проверку с опорой на источник (например, Википедию). Недостатком таких методов является необходимость в специальной инфраструктуре и дополнительной разработке специализированных компонентов для реализации указанных этапов обработки.

Когда нет возможности обращаться к другим источникам, можно использовать методы, анализирующие неуверенность модели во время генерации ответа [13]. Для работы этих методов необходим непосредственный доступ к весам и активациям моделей. В ситуациях, когда веса моделей недоступны, имеется несколько техник детекции галлюцинаций на основе промптинга [14] или использования LLM-«экзаменатора» [15].

Детекция галлюцинаций, не соответствующих инструкциям или контексту

При детекции галлюцинаций, связанных с искажением контекста или инструкций пользователя, применяются в том числе методы на основе анализа неуверенности модели путем проверки энтропии сгенерированного текста [16] или вероятностей токенов на различных стадиях генерации ответов [17]. Известны также различные техники промптинга, в которых используются дополнительные

запросы, сформированные по специальной структуре для оценивания соответствия ответа запросу [18].

Дополнительно можно определять различные показатели, которые указывают, насколько пересекаются сгенерированный контент и исходный запрос. Эти показатели могут отражать количество *n*-грамм [19], именованных сущностей (*NER*) [20] и отношений между этими сущностями [21]. Кроме того, можно использовать модели, обученные на задаче текстового следования [22], исходя из предположения, что верный ответ модели должен являться продолжением запроса пользователя.

При детекции галлюцинаций данного типа можно также рассматривать паттерны карт внимания, которые модель распределяет между словами входной и выходной последовательностей. В [23] авторы отмечают, что при возникновении галлюцинаций в задаче машинного перевода модель склонна фокусироваться лишь на небольшом числе начальных слов входной последовательности. В [24] предложено выделять ключевые слова в исходном тексте и анализировать, насколько большое внимание им уделялось при генерации ответа. В [25] использована мера доли внимания, направленного на входные токены, относительно общего объема внимания. Такой подход позволяет не только выявлять галлюцинации после генерации, но и интегрировать механизм контроля в сам процесс декодирования, предотвращая их появление на ранних этапах генерации.

Существуют также подходы, направленные на исследование взаимосвязи между внутренними состояниями моделей и проявлениями галлюцинаций. В [26] авторы анализируют корреляцию между внутренними представлениями модели и ее уверенностью в собственных ответах. В [27] выявлены структурные закономерности в пространстве скрытых состояний, соответствующих ответам на бинарные («да/нет») вопросы. В [28] предложен фреймворк для детекции галлюцинаций, включающий автоматическую генерацию данных с галлюцинациями и последующий анализ различий во внутренних состояниях модели при обработке сгенерированных последовательностей.

ИСПОЛЬЗОВАНИЕ ВНУТРЕННИХ СОСТОЯНИЙ МОДЕЛЕЙ ПРИ ДЕТЕКЦИИ ГАЛЛЮЦИНАЦИЙ

Настоящая работа посвящена детекции галлюцинаций, не соответствующих инструкциям или контексту. Ранее для проверки корректности ответов преимущественно использовались статистические критерии или крупные специализированные модели. В отличие от них, нашей целью является исследование пространства внутренних состояний языковых моделей при генерации и использование этих состояний в качестве признаков для обучения легковесных нейросетевых классификаторов. С их помощью решается задача бинарной классификации: определить, является ли сгенерированный ответ галлюцинацией или корректным.

В качестве признаков рассмотрим два типа внутренних состояний: карты перекрестного внимания в соответствующих блоках и выходы различных слоев модели при генерации. Для каждого из этих типов обучен специальный классификатор и проведена оценка его работы в различных задачах.

ИСПОЛЬЗУЕМЫЕ ДАТАСЕТЫ И МОДЕЛИ

Для детекции галлюцинаций были использованы два датасета: данные из соревнования Shroom 2024 [4] и бенчмарк HaluEval [3].

Данные из соревнования Shroom содержат сложнодетектируемые примеры галлюцинаций в отдельных задачах, таких как моделирование определений (definition modeling, DM), машинный перевод (machine translation, MT) и перефразирование (paraphrase generation, PG). Несмотря на большое количество данных в этом соревновании, число размеченных примеров невелико – 400 примеров каждого вида.

HaluEval – это масштабный бенчмарк для оценки галлюцинаций больших языковых моделей. Мы использовали подвыборки этого бенчмарка, соответствующие задачам «вопрос – ответ» (question answering, QA) и суммаризации (summarization), в каждой из которых содержится 10000 размеченных примеров. Каждый пример подвыборки состоит из запроса, передаваемого на вход модели, примера корректного ответа и примера галлюцинации для данного запроса.

В качестве моделей для экспериментов были использованы модели из семейства Flan-T5 [29], представляющие собой улучшенные версии модели T5 (Text-To-Text-Transfer-Transformer). Модели этого семейства, как и большинство больших языковых моделей на сегодняшний день, были обучены для решения большого числа различных задач в области NLP, поэтому внутренние состояния этих моделей потенциально репрезентативны и имеют развитую структуру, что позволяет использовать их для анализа. Благодаря этому также возможно оценить работу методов при различных сценариях. Для экспериментов была использована версия flan-t5-base, которая имеет 248 млн параметров.

ПРИЗНАКИ НА ОСНОВЕ CROSS-ATTENTION

В рамках экспериментов было установлено, что в случаях галлюцинаций и корректного ответа перекрестное внимание (cross-attention) модели фокусируется на разных частях входной последовательности.

Каждый декодер-слой $\mathrm{l} \in 1, \ldots, L$ в сети модели имеет несколько голов (heads) $\left[\mathrm{A^l}_1,\ \mathrm{A^l}_2,\ \ldots, \mathrm{A^l}_\mathrm{h}\right]$, где h – количество голов, $\mathrm{A}_\mathrm{i}^\mathrm{l}$ – матрица, у которой количество столбцов равно числу токенов входной последовательности, а строк – количеству токенов, сгенерированных в данный момент времени. Каждый элемент этой матрицы означает *важность* входного токена для выходного, которая выражается вещественным числом от 0 до 1. Нахождение максимума среди слоев и голов для входного токена a и выходного токена b – $\max_{\mathrm{i=1..h,\ j=1..L}} \left(\mathrm{A}_\mathrm{i}^\mathrm{j}\right)_{b,a}$ – позволяет выделить важные токены из контекста.

Ниже на двух примерах данных из HaluEval QA визуализированы различия в работе механизма перекрестного внимания при генерациях корректных ответов и галлюцинаций (см. рис. 1 и 2).

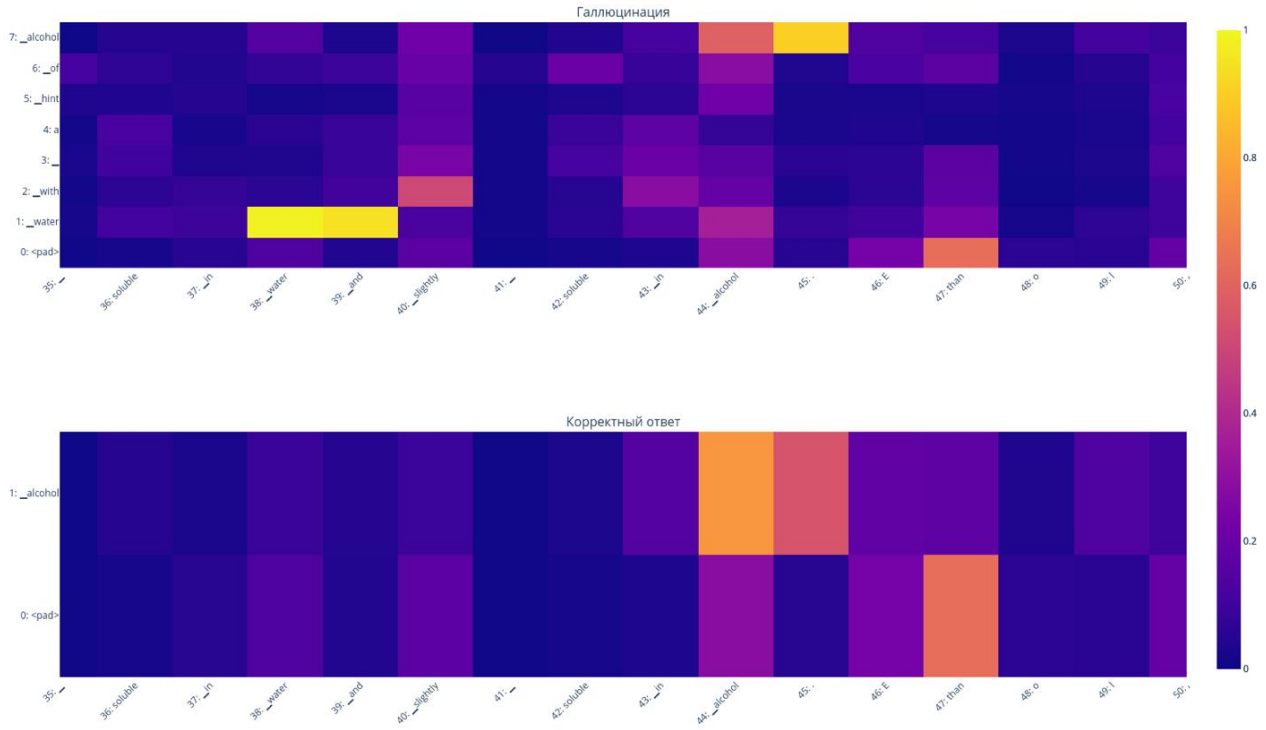

Рис. 1. При галлюцинации модель сильно фокусировалась на 38-м токене *‘_water’* и 39-м токене *‘_and’*, а при корректном ответе на 44-м токене *‘_alcohol’.*

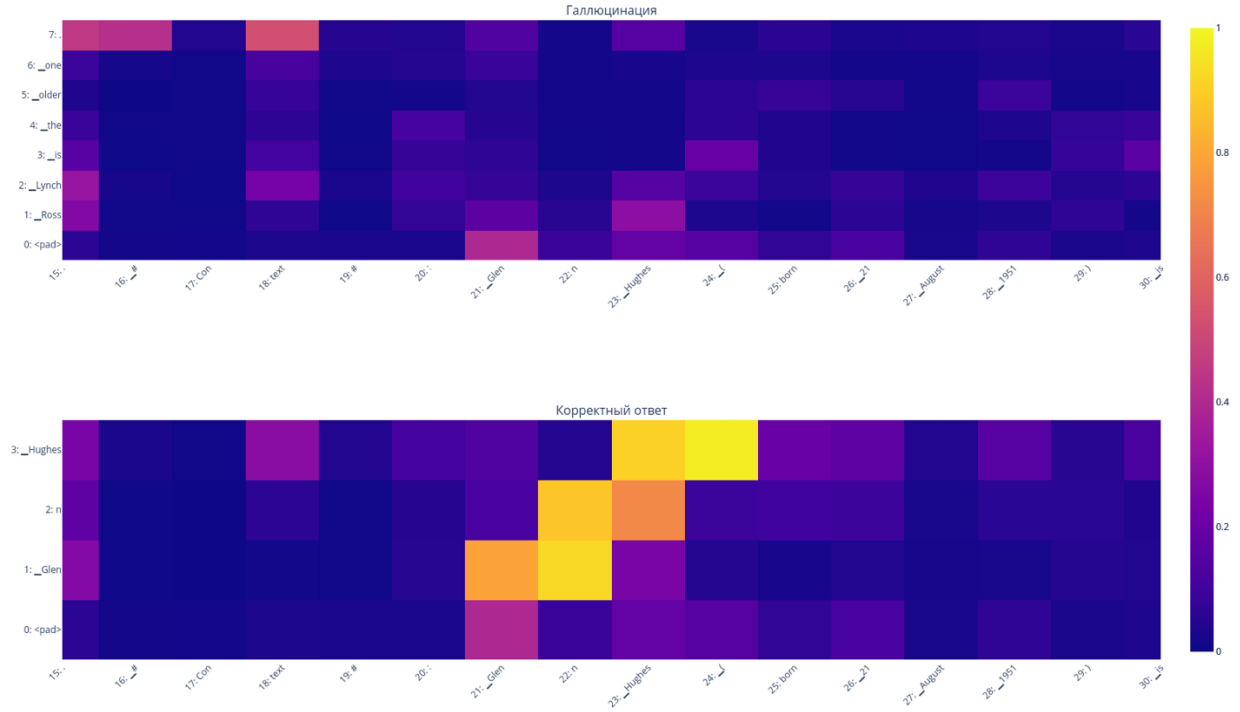

Рис. 2. При корректном ответе модель должна сильно фокусироваться на токенах 21 – 24: *‘_Glen’, ‘n’, ‘_Hughes’, ‘_(‘.* В случае галлюцинации модель не уделила этому сегменту текста внимания.

Карты перекрестного внимания заметно различаются между корректными ответами и галлюцинациями. Поэтому для проверки гипотезы об их использовании для детекции галлюцинаций была предложена следующая архитектура (см. рис. 3).

1. Входные данные x представляют собой тензор, состоящий из карт активаций внимания $[[\mathrm{A}^{\mathrm{l_i}}{}_1, \mathrm{A}^{\mathrm{l_i}}{}_2, \ldots, \mathrm{A}^{\mathrm{l_i}}{}_\mathrm{h}], \ldots, [\mathrm{A}^{\mathrm{l_{i+k}}}{}_1, \mathrm{A}^{\mathrm{l_{i+k}}}{}_2, \ldots, \mathrm{A}^{\mathrm{l_{i+k}}}{}_\mathrm{h}]]$ между токенами выходной и входной последовательностей, взятых с k последовательных слоев сети.

2. Первый блок детектора обрабатывает карты внимания различных голов модели. Он состоит из линейных преобразований, размер которых последовательно уменьшается (2790 -> 1395 -> 697 -> 348 -> 174), с функцией активации LeakyReLU между ними. В результате получается тензор x_r, в котором информация из карт внимания различных голов агрегирована в вектор меньшей размерности послойно.

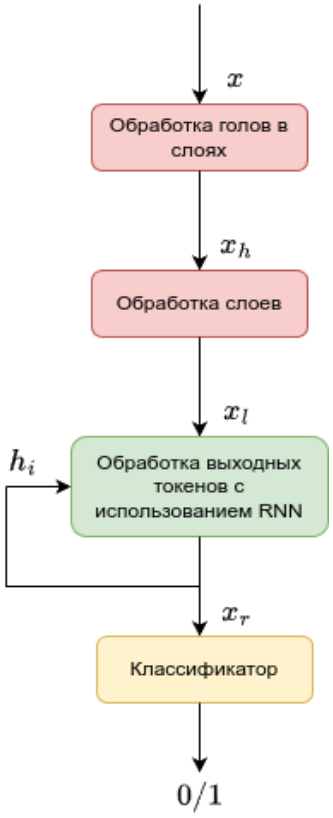

Рис. 3. Архитектура классификатора на основе перекрестного внимания.

3. После обработки голов получившиеся тензоры конкатенируются послойно, и для них аналогично применяются линейные трансформации с размерами (512 -> 256) и функции активации, как в предыдущем пункте. Таким образом, в x_l для каждого выходного токена собрана информация о важных токенах из контекста из разных слоев языковой модели.

4. Затем рекуррентная сеть LSTM [30] обрабатывает последовательности векторов из предыдущего пункта для каждого выходного токена. Такой способ позволяет анализировать, как использование информации из контекста изменяется во время генерации новых токенов.

5. Последнее внутреннее состояние данного слоя передается в классификатор. На выходе классификатор формирует вероятность того, что ответ является галлюцинацией.

Так как длина ответов может быть разной, необходимо обрезать слишком длинные или дополнять короткие ответы до одинаковой длины. В HaluEval QA галлюцинированные ответы в среднем длиннее, поэтому был также проведен эксперимент с обрезанием ответа до 4 выходных токенов.

Результаты данного алгоритма на HaluEval QA в зависимости от выбранных слоев представлены в табл. 1. Все эксперименты проводились в конфигурации, указанной в табл. 2.

Табл. 1. Результаты различных конфигураций алгоритма на HaluEval QA

Длина выходной последовательности	Слои flan-t5-base	F1-Score
32	[0, 1, 2, 3]	0.978
	[4, 5, 6, 7]	0.983
	[8, 9, 10, 11]	0.974
4	[0, 1, 2, 3]	0.852
	[4, 5, 6, 7]	0.883
	[8, 9, 10, 11]	0.849

Получившаяся сеть имеет порядка 22M параметров, что составляет примерно 9 % от размера flan-t5-base. Такой подход к детекции галлюцинаций не требует переобучения большой модели и показывает высокое качество классификации по F1-мере на тестовой выборке.

Несмотря на высокую точность, алгоритм имеет архитектурное ограничение: входные векторы чрезвычайно разрежены и обладают большой размерностью, что усложняет обработку длинных последовательностей. Поэтому для корректной работы с длинными текстами необходимы дополнительные изменения (например, разделение текста на куски более короткой длины и агрегирование результатов), выбор которых требует отдельного исследования.

Табл. 1. Параметры обучения

Параметр	Значение
Размер батча	64
Количество эпох	10
Размер обучающей выборки	16000
Размер тестовой выборки	4000
Скорость обучения (*learning rate*)	0.0001
Размер входной последовательности (с учетом паддинга)	465

ПРИЗНАКИ НА ОСНОВЕ ВЫХОДОВ РАЗЛИЧНЫХ СЛОЕВ МОДЕЛЕЙ

В ходе экспериментов было установлено, что выходные представления слоев модели при генерации галлюцинаций и корректных ответов также различаются. В отличие от карт перекрестного внимания, размер этих векторов не зависит от длины входной и выходной последовательностей и фиксируется на этапе обучения языковой модели. Благодаря этому использование признаков, построенных на основе внутренних представлений, не накладывает ограничений на размер контекста и позволяет эффективно детектировать галлюцинации в задачах с большим объемом текста, например, в HaluEval Summarization.

На рис. 4 представлены проекции скрытых внутренних состояний модели при генерации последнего токена в 1000 примерах суммаризации, полученные

с помощью PCA [31] (0-й слой – это слой, куда приходят выходы энкодера). Красным цветом выделены состояния, соответствующие галлюцинациями, синим – корректным ответам.

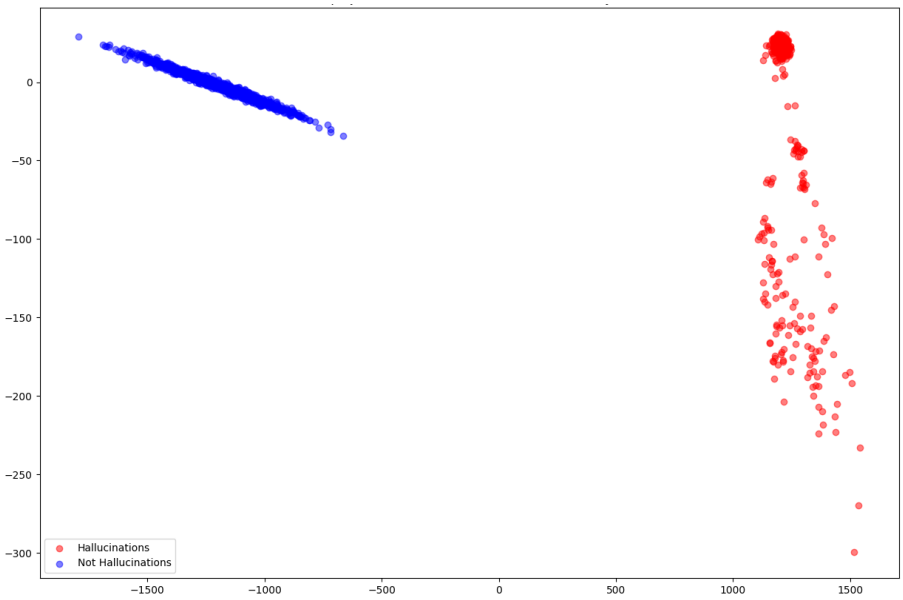

Рис. 4. PCA: Выходы 0-го слоя

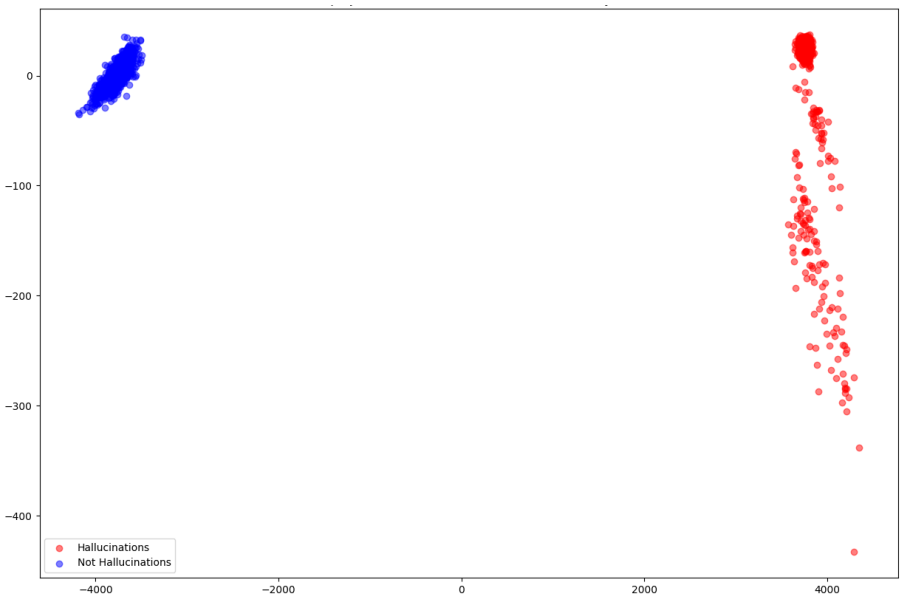

Рис. 5. PCA: Выходы 1-го слоя.

Как видно из проекций, внутренние состояния оказались довольно хорошо различимы. На основе этих значений также был обучен детектор галлюцинаций. Детектор представляет собой нейронную сеть, состоящую из 3 линейных слоев и функций активации LeakyReLU между ними, на вход которой приходят активации с различных слоев при генерации последнего токена. В табл. 3 и 4Ниже представлены избранные результаты детектора, использующего активации с различных слоев в качестве входных данных, на задаче HaluEval Summarization.

Табл. 2. Результаты детектора на HaluEval Summarization в зависимости от слоя модели flan-t5-base (нумерация с 0)

Номер слоя	F1-Score
0	0.305
1	0.757
2	0.859
6	0.924
7	0.944
8	0.951
11	0.861

Табл. 3. Параметры обучения

Параметр	Значение
Размер батча	128
Количество эпох	10
Размер обучающей выборки	18000
Размер тестовой выборки	2000
Скорость обучения (*learning rate*)	0.0001
Размер входной последовательности (с учетом паддинга)	1600
Размер выходной последовательности (с учетом паддинга)	128

Из результатов эксперимента видно, что распределения внутренних состояний слоев сильно отличаются при генерации галлюцинаций и верных ответов почти во всех слоях сети. Лучшие результаты получаются для слоев 6–8. Кроме того, достоинство этого метода заключается в том, что он работает независимо от длин входной и выходной последовательностей: выходы слоев всегда имеют одну размерность, задаваемую архитектурой исходной языковой модели. Получившийся детектор имеет порядка 566 тыс. параметров, что составляет примерно 0.2 % от размера исходной языковой модели.

ПЕРЕНОСИМОСТЬ МЕЖДУ РАЗЛИЧНЫМИ ТИПАМИ ГАЛЛЮЦИНАЦИЙ

Для проверки переносимости результатов детекторы, обученные на задаче HaluEval Summarization, были запущены на галлюцинациях других типов: HaluEval QA, Shroom MT, Shroom PG, Shroom DM (табл. 5 и 6).

Табл. 4. Результаты классификатора (F1-Score), обученного на HaluEval Summarization, на других задачах.

Номер слоя	HaluEval Summarization	HaluEval QA	Shroom PG	Shroom DM	Shroom MT
0	0.305	0.002	0.362	0.661	0.562
1	0.757	0.194	0.337	0.531	0.553
2	0.859	0.365	0.267	0.294	0.464
6	0.924	0.56	0.256	0.094	0.46
7	0.944	0.632	0.358	0.472	0.531
8	0.951	0.595	0.283	0.313	0.482
11	0.861	0.514	0.283	0.607	0.554

Эксперименты показали, что особенности распределений слоев сети, присущие галлюцинациям в задаче суммаризации, плохо переносятся на галлюцинации других типов. Таким образом можно сделать вывод, что состояния модели при генерации галлюцинаций различных типов имеют разное распределение, и для построения «универсального» детектора необходимы примеры галлюцинаций в разных задачах.

Табл. 5. Результаты классификатора (F1-Score), обученного на HaluEval QA, на других задачах.

Номер слоя	HaluEval Summarization	HaluEval QA	Shroom PG	Shroom DM	Shroom MT
0	0.666	0.665	0.362	0.662	0.562
1	0.662	0.910	0.352	0.658	0.534
2	0.665	0.959	0.362	0.66	0.561
6	0.666	0.975	0.361	0.661	0.562
7	0.666	0.978	0.361	0.661	0.562
8	0.667	0.980	0.362	0.662	0.561
11	0.663	0.965	0.36	0.661	0.561

При обучении детектора на задаче HaluEval QA также наблюдаются различия в зависимости используемых слоев, но в значительно меньшей степени, чем в задаче суммаризации. Аналогично задаче QA, детекторы на основе слоев с 6 по 8 демонстрируют наилучшие результаты. Лучшее качество получается при использовании 8-го слоя, что незначительно меньше, чем качество, полученное с помощью детектора на основе перекрестного внимания: 0.98 (табл. 6) против 0.983 (табл. 1).

При последующем применении данного детектора к другим задачах была снова подтверждена плохая переносимость между различными типами галлюцинаций. Примечательно, что при использовании классификатора, обученного на задаче QA, выбор входного слоя модели практически не влияет на качество детекции в других задачах. Это может свидетельствовать о том, что при генерации суммаризаций распределения активаций по слоям отличаются между собой значительно сильнее (табл. 5), чем в задаче QA.

ЗАКЛЮЧЕНИЕ

Рассмотрен феномен галлюцинаций в больших языковых моделях с фокусом на случаи, возникающие из-за несоответствия между входными и выходными данными. Для автоматической детекции таких галлюцинаций на датасетах

Shroom и HaluEval была использована внешняя модель flan-t5-base, а также предложены и исследованы подходы на основе скрытых состояний и карт внимания.

Проведенные эксперименты показали, что внутренние представления модели при генерации галлюцинированных и корректных ответов имеют различия, которые можно использовать для построения эффективных детекторов. В частности, методы, опирающиеся на скрытые состояния, продемонстрировали более широкую применимость и аналогичную точность по сравнению с подходами на основе внимания, особенно при работе с длинными входами, где последние требуют предварительной агрегации признаков. Это указывает на перспективность использования скрытых представлений как более универсального признакового пространства для задач детекции. В ходе экспериментов было установлено, что скрытые состояния промежуточных слоев (6–8 для модели *flan-t5-base)* наиболее информативны для задачи детекции галлюцинаций.

В рамках работы был реализован и обучен классификатор, использующий скрытые состояния модели и демонстрирующий высокое качество детекции галлюцинаций на ряде задач. Кроме того, была предпринята попытка построения модели на основе LSTM, принимающей визуализации карт внимания в качестве входа. Этот подход, несмотря на высокое качество классификации, имеет ограничения при работе с длинными последовательностями.

Отдельное внимание было уделено вопросу переносимости: детекторы, обученные на одном типе галлюцинаций, показывают ограниченную способность к переносу на другие задачи, что подчеркивает необходимость в более универсальных подходах и корпусах для обучения. Это направление остается открытым и требует дальнейшего изучения.

Одним из потенциально перспективных путей продолжения работы является применение статистических методов анализа и использование специальных выборок для изучения внутренних состояний моделей. Такие подходы могут способствовать выявлению более общих закономерностей и повышению переносимости детекторов между задачами.

СПИСОК ЛИТЕРАТУРЫ

1. *Vaswani A., Shazeer N., Parmar N. et al.* Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 30.

2. *Huang L., Yu W., Ma W. et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55. https://doi.org/10.1145/3703155

3. *Li J., Cheng X., Zhao W. X. et al.* HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397

4. *Mickus T., Zosa E., Vázquez R. et al.* SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes // International Workshop on Semantic Evaluation. 2024. https://doi.org/10.18653/v1/2024.semeval-1.273

5. *Carlini N., Ippolito D., Jagielski M. et al.* Quantifying Memorization Across Neural Language Models // The Eleventh International Conference on Learning Representations. 2023. https://doi.org/10.48550/arXiv.2202.07646

6. *Lin S., Hilton J., Evans Q.* TruthfulQA: Measuring How Models Mimic Human Falsehoods // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022. Vol. 1. P. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

7. *Li D., Rawat A.S., Zaheer M. et al.* Large Language Models with Controllable Working Memory // Findings of the Association for Computational Linguistics: ACL 2023. 2023. P. 1774–1793. https://doi.org/10.18653/v1/2023.findings-acl.112

8. *Sharma M., Tong M., Korbak T. et al.* Towards Understanding Sycophancy in Language Models // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2310.13548

9. Reinforcement Learning from Human Feedback: Progress and Challenges // YouTube. URL: https://www.youtube.com/watch?v=hhiLw5Q_UFg (дата обращения: 04.05.2025)

10. *Chuang Y.S., Xie Y., Luo H. et al.* DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models // ArXiv. 2023. Vol. abs/2309.03883. https://doi.org/10.48550/arXiv.2309.03883

11. *Voita E., Talbot D., Moiseev F. et al.* Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5797–5808. https://doi.org/10.18653/v1/P19-1580

12. *Min S., Krishna K., Lyu X. et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

13. *Luo Z., Xie Q., Ananiadou S.* ChatGPT as a Factual Inconsistency Evaluator for Text Summarization // ArXiv. 2023. Vol. abs/2303.15621. https://doi.org/10.48550/arXiv.2303.15621

14. *Manakul P., Liusie A., Gales M.J.* SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

15. *Cohen R., Hamri M., Geva M. et al.* LM vs LM: Detecting Factual Errors via Cross Examination // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 12621–12640. https://doi.org/10.18653/v1/2023.emnlp-main.778

16. *Xiao Y., Wang W.Y.* On Hallucination and Predictive Uncertainty in Conditional Language Generation // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021. P. 2734–2744. https://doi.org/10.18653/v1/2021.eacl-main.236

17. *Miao N., Teh Y.W., Rainforth T.* SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2308.00436

18. *Adlakha V., BehnamGhader P., Lu X.H. et al.* Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering // Transactions of the Association for Computational Linguistics. 2024. Vol. 12. P. 681–699. https://doi.org/10.1162/tacl_a_00667

19. *Lin Chin-Yew.* ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. 2004. P. 74-81. ISBN: 9781932432466

20. *Venkit P.N., Gautam S., Panchanadikar R. et al.* Nationality Bias in Text Generation // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023. P. 116–122. https://doi.org/10.18653/v1/2023.eacl-main.9

21. *Goodrich B., Rao V., Liu P.J. et al.* Assessing The Factual Accuracy of Generated Text // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. P. 166–175. https://doi.org/10.1145/3292500.3330955

22. *Laban P., Schnabel T., Bennett P.N. et al.* SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 163–177. https://doi.org/10.1162/tacl_a_00453

23. *Xu W., Agrawal S., Briakou E. et al.* Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection // Transactions of the Association for Computational Linguistics. 2023. Vol. 11. P. 546–564. https://doi.org/10.1162/tacl_a_00563

24. *Zhang T., Qiu L., Guo Q. et al.* Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus // Conference on Empirical Methods in Natural Language Processing. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.58

25. *Chuang Y.S., Qiu L., Hsieh C.Y. et al.* Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. P. 1419–1436. https://doi.org/10.18653/v1/2024.emnlp-main.84

26. *Yin Z., Sun Q., Guo Q. et al.* Do Large Language Models Know What They Don't Know? // Annual Meeting of the Association for Computational Linguistics. 2023. https://doi.org/10.18653/v1/2023.findings-acl.551

27. *Marks S., Tegmark M.* The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets // First Conference on Language Modeling. 2024. https://doi.org/10.48550/arXiv.2310.06824

28. *Su W., Wang C., Ai Q. et al.* Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models // Annual Meeting of the Association for Computational Linguistics. 2024. https://doi.org/10.48550/arXiv.2403.06448

29. *Chung H.W., Hou L., Longpre S. et al.* Scaling Instruction-Finetuned Language Models // Journal of Machine Learning Research. 2024. Vol. 25, No. 70. P. 1–53. https://doi.org/10.5555/3722577.3722647

30. *Hochreiter Sepp, Schmidhuber Jürgen.* Long Short-Term Memory // Neural Computation. 1997. Vol. 9, No. 8. P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

31. PCA // Wikipedia. URL: https://en.wikipedia.org/wiki/Principal_component_analysis (дата обращения: 13.06.2025).

DETECTION OF HALLUCINATIONS BASED ON THE INTERNAL STATES OF LARGE LANGUAGE MODELS

T. R. Aisin[1] [0009-0001-5863-3252], T. V. Shamardina[2] [0009-0008-9033-6646]

[1, 2]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

[1]aysin.timur@gmail.ru, [2]shamardina.tatiana@gmail.com

Abstract

In recent years, large language models (LLMs) have achieved substantial progress in natural language processing tasks and have become key instruments for addressing a wide range of applied and research problems. However, as their scale and

capabilities grow, the issue of hallucinations — i.e., the generation of false, unreliable, or nonexistent information presented in a credible manner—has become increasingly acute. Consequently, analyzing the nature of hallucinations and developing methods for their detection has acquired both scientific and practical significance.

This study examines the phenomenon of hallucinations in large language models, reviews their existing classification, and investigates potential causes. Using the Flan-T5 model, we analyze differences in the model's internal states when generating hallucinations versus correct responses. Based on these discrepancies, we propose two approaches for hallucination detection: one leveraging attention maps and the other utilizing the model's hidden states. These methods are evaluated on data from HaluEval and Shroom 2024 benchmarks in tasks such as summarization, question answering, paraphrasing, machine translation, and definition generation. Additionally, we assess the transferability of the trained detectors across different hallucination types, in order to evaluate the robustness of the proposed methods.

Keywords*: large language models, hallucinations, detection, Flan-T5, natural language processing, attention maps, hidden states, HaluEval, Shroom.*

REFERENCES

1. *Vaswani A., Shazeer N., Parmar N. et al.* Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 30.

2. *Huang L., Yu W., Ma W. et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55. https://doi.org/10.1145/3703155

3. *Li J., Cheng X., Zhao W. X. et al.* HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397

4. *Mickus T., Zosa E., Vázquez R. et al.* SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes // International Workshop on Semantic Evaluation. 2024. https://doi.org/10.18653/v1/2024.semeval-1.273

5. *Carlini N., Ippolito D., Jagielski M. et al.* Quantifying Memorization Across Neural Language Models // The Eleventh International Conference on Learning Representations. 2023. https://doi.org/10.48550/arXiv.2202.07646

6. *Lin S., Hilton J., Evans Q.* TruthfulQA: Measuring How Models Mimic Human Falsehoods // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022. Vol. 1. P. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

7. *Li D., Rawat A.S., Zaheer M. et al.* Large Language Models with Controllable Working Memory // Findings of the Association for Computational Linguistics: ACL 2023. 2023. P. 1774–1793. https://doi.org/10.18653/v1/2023.findings-acl.112

8. *Sharma M., Tong M., Korbak T. et al.* Towards Understanding Sycophancy in Language Models // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2310.13548

9. Reinforcement Learning from Human Feedback: Progress and Challenges // YouTube. URL: https://www.youtube.com/watch?v=hhiLw5Q_UFg (дата обращения: 04.05.2025)

10. *Chuang Y.S., Xie Y., Luo H. et al.* DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models // ArXiv. 2023. Vol. abs/2309.03883. https://doi.org/10.48550/arXiv.2309.03883

11. *Voita E., Talbot D., Moiseev F. et al.* Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5797–5808. https://doi.org/10.18653/v1/P19-1580

12. *Min S., Krishna K., Lyu X. et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

13. *Luo Z., Xie Q., Ananiadou S.* ChatGPT as a Factual Inconsistency Evaluator for Text Summarization // ArXiv. 2023. Vol. abs/2303.15621. https://doi.org/10.48550/arXiv.2303.15621

14. *Manakul P., Liusie A., Gales M.J.* SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

15. *Cohen R., Hamri M., Geva M. et al.* LM vs LM: Detecting Factual Errors via Cross Examination // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 12621–12640. https://doi.org/10.18653/v1/2023.emnlp-main.778

16. *Xiao Y., Wang W.Y.* On Hallucination and Predictive Uncertainty in Conditional Language Generation // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021. P. 2734–2744. https://doi.org/10.18653/v1/2021.eacl-main.236

17. *Miao N., Teh Y.W., Rainforth T.* SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2308.00436

18. *Adlakha V., BehnamGhader P., Lu X.H. et al.* Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering // Transactions of the Association for Computational Linguistics. 2024. Vol. 12. P. 681–699. https://doi.org/10.1162/tacl_a_00667

19. *Lin Chin-Yew.* ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. 2004. P. 74-81. ISBN: 9781932432466

20. *Venkit P.N., Gautam S., Panchanadikar R. et al.* Nationality Bias in Text Generation // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023. P. 116–122. https://doi.org/10.18653/v1/2023.eacl-main.9

21. *Goodrich B., Rao V., Liu P.J. et al.* Assessing The Factual Accuracy of Generated Text // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. P. 166–175. https://doi.org/10.1145/3292500.3330955

22. *Laban P., Schnabel T., Bennett P.N. et al.* SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 163–177.

https://doi.org/10.1162/tacl_a_00453

23. *Xu W., Agrawal S., Briakou E. et al.* Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection // Transactions of the Association for Computational Linguistics. 2023. Vol. 11. P. 546–564. https://doi.org/10.1162/tacl_a_00563

24. *Zhang T., Qiu L., Guo Q. et al.* Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus // Conference on Empirical Methods in Natural Language Processing. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.58

25. *Chuang Y.S., Qiu L., Hsieh C.Y. et al.* Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. P. 1419–1436. https://doi.org/10.18653/v1/2024.emnlp-main.84

26. *Yin Z., Sun Q., Guo Q. et al.* Do Large Language Models Know What They Don't Know? // Annual Meeting of the Association for Computational Linguistics. 2023. https://doi.org/10.18653/v1/2023.findings-acl.551

27. *Marks S., Tegmark M.* The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets // First Conference on Language Modeling. 2024. https://doi.org/10.48550/arXiv.2310.06824

28. *Su W., Wang C., Ai Q. et al.* Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models // Annual Meeting of the Association for Computational Linguistics. 2024. https://doi.org/10.48550/arXiv.2403.06448

29. *Chung H.W., Hou L., Longpre S. et al.* Scaling Instruction-Finetuned Language Models // Journal of Machine Learning Research. 2024. Vol. 25, No. 70. P. 1–53. https://doi.org/10.5555/3722577.3722647

30. *Hochreiter Sepp, Schmidhuber Jürgen.* Long Short-Term Memory // Neural Computation. 1997. Vol. 9, No. 8. P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

31. PCA // Wikipedia. URL: https://en.wikipedia.org/wiki/Principal_component_analysis (дата обращения: 13.06.2025).

СВЕДЕНИЯ ОБ АВТОРАХ

АЙСИН Тимур Рустемович – инженер машинного обучения и исследователь в области обработки естественного языка. Область научных интересов: оценка языковых моделей, большие языковые модели, интерпретируемость моделей, оптимизация инференса языковых моделей.

Timur Rustemovich AISIN – ML Engineer & NLP researcher. Research interests: language models evaluation, large language models, benchmarking, language models interpretability, inference optimization.

email: aysin.timur@gmail.com

ORCID: 0009-0001-5863-3252

ШАМАРДИНА Татьяна Вячеславовна – исследователь в области обработки естественного языка. Область научных интересов: оценка языковых моделей, большие языковые модели, интерпретируемость моделей. Число научных публикаций – 3.

Tatiana Vyacheslavovna SHAMARDINA – NLP researcher. Research interests: language models evaluation, large language models, benchmarking, language models interpretability. The number of publications – 3.

email: shamardina.tatiana@gmail.com

ORCID: 0009-0008-9033-6646

Материал поступил в редакцию 6 ноября 2025 года