DOI 10.37882/2223-2966.2025.07.10

МЕТОДЫ ОЦЕНКИ ИЗВЛЕЧЕНИЯ ГЛУБИННОЙ СЕМАНТИКИ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ

METHODS FOR ASSESSING DEEP SEMANTICS EXTRACTION BY LARGE LANGUAGE MODELS

D. Gromozdov
Yu. Gapanyuk
G. Afanasyev

Summary. The widespread use of textual data and the gradual increase in the complexity of natural language processing tasks that can be solved with large language models leads to the discovery of non-trivial facts, or underlying semantics, in the data. Creating false facts to produce an answer for the user is also a common characteristic of language models, known as hallucinations. It is necessary to develop a system for evaluating the truth of a fact obtained from a model. The aim of this paper is to identify a metric suitable for the evaluation of deep semantics. A statistical semantic content estimation method and a method based on machine learning were applied to the output of a large language model, which was tasked with identifying non-trivial fact from texts. The possibility of combining the approaches, to obtain a more accurate and interpretable metric, is discussed. The conducted research defines the limits of applicability and interpretability of modern metrics for the tasks of semantic analysis solved by means of large language models. We propose possible ways of solving the problem of selecting a semantic metric based on the widely spread method of intelligent agents, vector databases used for retrieval augmented generation, and other software solutions for analyzing the truth of facts or their adequacy to the information contained in the corpus of texts.

Keywords: large language models, metric, evaluation models, deep semantics, metagraph.

Громоздов Данила Романович
Аспирант, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
gromozdovdr@student.bmstu.ru
Гапанюк Юрий Евгеньевич
К.т.н., доцент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
gapyu@bmstu.ru
Афанасьев Геннадий Иванович
К.т.н., доцент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
gaipcs@bmstu.ru

Аннотация. Широкое использование текстовых данных и постепенное усложнение задач обработки естественного языка, которые возможно решить с помощью больших языковых моделей, приводит к выявлению в данных нетривиальных фактов, или глубинной семантики. Создание ложных фактов для выдачи ответа пользователю — также распространённая характеристика языковых моделей, известная как галлюцинации. Необходимо разработать систему оценки истинности полученного от модели факта. Целью данной работы является определение метрики, подходящей для оценки глубинной семантики. Метод статистической оценки семантического содержания и метод, основанный на машинном обучении, были применены к выводам большой языковой модели, которой была поставлена задача выявления нетривиального факта из текстов. Обсуждается возможность совмещения подходов, для получения более точной и интерпретируемой метрики. Проведённое исследование определяет границы применимости и интерпретируемости современных метрик для задач семантического анализа, решаемых средствами больших языковых моделей. Нами предложены возможные пути решения проблемы подбора семантических метрик, основанные на широко распространяемом методе интеллектуальных агентов, векторных баз данных, используемых для синтеза ответа дополненного поиском, и других программных решениях для анализа истинности фактов или их адекватности информации, заключённой в корпусе текстов.

Ключевые слова: большие языковые модели, метрики, оценочные модели, глубинная семантика, метаграфы.

Введение

Создание архитектуры, основанной на механизме внимания, открыло возможность для скачка в эффективности обработки текстов на естественном языке. На основе данного подхода к обработке данных, содержащих последовательности, создано большинство современных больших языковых моделей или LLM (Large Language Models). Данный тип моделей позволил далеко продвинуться в работе с пользовательскими запросами на естественном языке.

Благодаря развитию технологий обработки естественного языка, текстовые данные становятся частью больших данных и могут быть обработаны программными средствами, а не только с помощью экспертов. Важную роль в процессе анализа семантики в производственных масштабах играет валидация полученного

результата — необходимо убедиться, что в полученные большой языковой моделью данные не внедрилась галлюцинация (факты, созданные LLM для выдачи ответа, не имеющие основы в реальности) или ложные факты. Таким образом, существует необходимость в системе метрик, которые могли бы с достаточной степенью уверенности оценивать семантические характеристики ответа используемой модели.

Современные методы для оценки семантики ответов больших языковых моделей можно подразделить на две основные группы: статистические методы и методы модель-оценивает-модель. Статистические метрики основываются на числовых оценках схожести последовательностей, полученных в результате ответа на пользовательский запрос и представляющих собой эталон ответа. Рассматриваемый тип оценок даёт достаточно строгие результаты вычисления метрики, но также требует чёткости критериев истинности полученного текста. Оценки на основе статистики оказываются привязаны к точным формулировкам, заданным экспертами, и к конкретным терминам, семантика переданная перефразированно не будет улавливаться при измерении эффективности.

Метрики на основе машинного обучения используют специально обученные языковые модели, которые выполняют запросы, требующие у них выдачи оценки ответу основной большой языковой модели. В таком случае нет возможности отследить чёткие критерии, по которым модель-асессор оценивает истинность ответа, но заложенные в параметры модели внешние знания позволяют ей проводить семантический анализ текста на естественном языке более гибко, не привязываясь к конкретным терминам или словам в эталонах ответов.

В данном исследовании анализируются подходы к анализу глубинной семантики. Глубинная семантика — это нетривиальные факты, которые в скрытом виде могут содержаться в корпусе текстов, а также объединять информацию из разных документов. Для отображения такой иерархической структуры мы предлагаем использовать метаграфовую модель, разрешающую объединение элементов в метавершины более высокого порядка и создание метарёбер, связывающих элементы на более высоком уровне.

Таким образом, задача оценки ответов больших языковых моделей, работающих с глубинной семантикой является нетривиальной и требует разработки новых методов на основе уже имеющихся в нашем распоряжении.

Методы

1. *Статистические методы оценки*

Помимо непосредственно метода анализа текстов с помощью больших языковых моделей необходимо также разработать подход к оценке качества его работы. Поскольку мы имеем дело с нечёткой системой ответов моделей, в отличие, например, от задачи классификации, где имеется конечное множество возможных результатов, при работе с большими языковыми моделями количество вариантов удовлетворительного ответа модели не фиксировано заранее при постановке задачи. В условиях большой вариативности необходимо подобрать более лабильный метод оценки работы. Далее будут рассмотрены существующие подходы к показателям качества работы моделей семантического анализа с применением больших языковых моделей.

Первая группа метрик, рассматриваемых нами для проведения оценки эффективности семантического анализа большими языковыми моделями, использует числовые статистические методы для измерения схожести содержания полученного от модели ответа с представленным исследователем эталоном.

$$Q(y) = f\left(y, \hat{y}\right),\ \hat{y} = \left\{y_i * g\left(y_i|\beta\right)\right\}, \quad (1)$$

где $Q(y)$ — это метрика, или оценочная функция для эффективности модели; $f(y, \hat{y})$ — функция, которая может иметь дополнительные параметры: она служит для установления соотношения между верными ответами модели и несоответствующими заданным критериям истинности; y — это множество ответов, данных обученной моделью, а $\hat{y}$ — множество верных ответов; функция $g(y|\beta)$ — сопоставление i-ого ответа модели с соответствующим параметром β, отражающим принятые исследователями или экспертами критерии истинности.

Функция $g(y)$ принимает значение 1, если элемент совпадает с верным ответом, и 0, если не соответствует параметру истинности. Чаще всего функция $Q(y)$ является обратной зависимостью, связывающей верные ответы с совокупностью всех выводов модели.

Параметр β может быть не выражен явно, а заключаться, например, в экспертности человека, составляющего набор данных, или в правильности выполнения программы, если мы говорим о частой проблеме, решаемой в сфере семантического анализа — написание SQL запроса на основе пользовательского обращения на естественном языке. Для технических данных это может быть соответствие теоретически рассчитанному по адекватному задаче набору формул значению.

В случае же глубинной семантики определить критерии истинности β гораздо сложнее в виду вариативности возможных верных интерпретаций набора данных и использования системой внешних знаний [7].

Часто справедливо и следующее утверждение, касательно метрики Q:

$$Q \in [a,b], \quad (2)$$

где a и b — это нижняя и верхняя граница значений, которые может принимать метрика.

Действительно, производить интерпретацию и сравнивать оценки для различных моделей легче, когда у нас есть чёткие границы, обозначающие экстремальные значения, которые можно интерпретировать как наилучшее и наихудшее соответствие заданным критериям истинности.

Наиболее простые метрики проверяют точное соответствие полученного ответа эталону. Такие оценки подходят там, где ответ может быть чётко и однозначно формализован, например, в задаче генерации SQL-запроса на основе пользовательского запроса на естественном языке. В таком случае семантика, заключённая в сообщении, представляется в виде кода и соответствует определённым функции или атрибуту в запросе к базе данных, однако и в решении такой задачи, по замечаниям исследователей, может проявляться вариативность формулировки запроса при достаточно сложных отношениях таблиц и операций над ними. Например, это может быть метрика точности (accuracy), представляющая собой отношение количество верных (в точности совпадающих с эталоном) ответов к общему количеству выводов модели [15].

Если вспомнить о метрике в геометрическом смысле, как о функции вводящей расстояние между элементами, то стоит вспомнить и о расстоянии Левенштейна. Это метрика, которая вводит расстояние между двумя последовательностями символов, которое характеризуется количеством замен и других преобразований (имеющих разный вес), которое необходимо сделать, чтобы из одной получить другую. Глубинную семантику, имеющую высокую лабильность способов выражения очевидно не получится измерить с помощью расстояний Левенштейна, но это важный этап перехода от точных совпадений к сопоставлению символьных последовательностей между собой [5].

Одна из первых метрик для автономной оценки качества работы моделей обработки естественного языка — Помощник в Оценке Двуязычных Результатов/BiLingual Evaluation Understudy (BLEU). Изначально данная метрика создавалась в 2001 году командой авторов из IBM для оценки в задачах машинного перевода [8]. В приложении к задаче синтаксического парсинга, например, мы можем представить, что это задача перевода текста с естественного языка на язык грамматики и синтаксиса языка. В таком случае, становится возможным применение метрики BLEU при наличии эталонного примера синтаксического разбора текста.

Тем не менее, в результате тщательного исследования применения метрики BLEU для оценки качества Э. Райтер в статье [10] представляет критику использования данного подхода для решения задач обработки естественного языка за исключением машинного перевода. Если говорить о задаче синтаксического парсинга, то в данном случае мы имеем возможность получить точное соответствие текста на естественном языке структурам в понятиях его грамматики. Однако в случае семантического анализа получить такое соответствие гораздо сложнее, а следовательно метод оценки качества BLUE не подходит для формирования отклика о качестве работы исследуемой модели.

В 2019 году Соном Линьфэно и соавтором была предложена [14] вариация метрики BLUE, расширяющий её применение на язык AMR (Abstract Meaning Representation, Представления Абстрактного Смысла), — SemBLUE. Вслед за оригинальной метрикой BLEU поиск соответствий ищется в пределах N элементов от исследуемого слова, но в случае SemBLEU поиск осуществляется в ширину по дереву представления семантики. В качестве элементов выступают вершины графа, соответствие которых и проверяется аналогично оригинальному методу. В данном случае поиск в пределах N элементов более эффективный, поскольку несомненно затрагивает слова, связанные друг с другом, что отнюдь не гарантируется контекстом в текстовой форме предложения, особенно при малых значениях N.

SemBLUE хорошо подходит и для семантического, и для синтаксического анализа, однако для его применения необходимо наличие базы данных, содержащей AMR представления всех предложений, содержащихся в корпусе текстов, что очевидно трудно выполнимо, хотя и возможно благодаря современным синтаксическим анализаторам, или парсерам. Тем не менее проблема учёта синонимии и вариативности форм представления одной семантики актуальна и для этого метода, который также зависит от формулировок экспертов. А проверка разных вариантов эталонного решения займёт ещё больше времени из-за необходимости применения алгоритма поиска в ширину по дереву, которое может иметь достаточно разветвлённую структуру.

Из недостатков статистических метрик можно указать то, что они эффективны в основном для измерения адекватности модели задаче до её внедрения в производственные процессы. Требуется иметь заранее подготовленный набор эталонных ответов для тестовых вариантов задач (бенчмарки), чтобы оценить работу модели в производстве, будет необходимо дополнительно создать массив истинных фактов, соответствующих текстам, содержащимся в реальной базе данных компании. Разметка данных, тем более содержащих сложные семантические связи, значительно повышает затраты людских ресурсов.

2. *Методы на основе оценивающих моделей машинного обучения*

Среди современных технологий по оценке эффективности больших языковых моделей можно выделить методы, основанные на использовании других моделей машинного обучения, также называемые «Большая языковая модель, как судья» (LLM-as-a-judge). Создание оценочной нейронной сети производится путём дообучения уже имеющихся моделей, способных к работе с естественным языком, использоваться может даже вариант той же модели, ответы которой подвергаются оценке [13].

Для дополнительного обучения формулируется особая форма запроса, в которой содержатся:

- Общее описание желаемого отчёта о проведённом анализе семантики, элементы ответа и требования к оценке: ограничение на целые значения, диапазон оценивания;
- Запрос, который был предоставлен большой языковой модели для выдачи соответствующего ответа;
- Ответ, который был выдан моделью на пользовательский запрос;
- Эталонный вариант ответа, предоставленный производящим оценивание программистом;
- Рубрикация оценок: для каждой возможной оценки программистом или другим экспертом предоставляется текстовое описание критериев присвоения соответствующего значения реакции большой языковой модели [9].

Очевидно преимущество данного метода — он сам может оценивать синонимичность замен или вариаций в тексте относительно данного эталона ответа. Модель машинного обучения, особенно большая языковая, имеет навыки анализа текстов, которые может применять к своим собственным ответам. Предоставленные во время дообучения на человеческих примерах оценивания знания могут быть использованы для более строгого анализа данных запроса и ответа, а также установка на выявление ошибок позволяет уменьшить шанс того, что модель будет только «хвалить» ответы оцениваемой модели.

Если возвратиться к формуле 1 из предыдущего раздела, то можно сказать, что в качестве критерия «истинности» данных β будут выступать непосредственно сами параметры оценивающей модели.

Таким образом, исходя из общего вида запроса натренированной модели, обратившись к формуле 2, в случае предобученных оценивающих моделей это выражение будет иметь следующий вид: $Q \in [1, 5]$, $Q \in \mathbb{Z}$. Функция оценки в таком случае становится дискретной и принимает конечное число значений: 5, в отличие от статистических метрик, которые чаще всего непрерывны в пределах своей области определения.

При этом возможно задать оценивающей модели и дробные значения на отрезке [0; 1], но тогда мы не можем точно интерпретировать присвоение того или иного нецелого значения метрики ответу. Описание критериев оценки доступно только для конечного числа реперных точек, и даже если модель предоставит некоторые пояснения к своему решению, точного соответствия качественного критерия для текста на естественном языке и численного значения не будет предоставлено [16].

Отсюда вытекает один из недостатков методов на основе машинного обучения — их малая математическая интерпретируемость. Мы можем только согласиться с обоснованиями модели для присвоения определённого значения метрике или же их отвергнуть, но это потребует дополнительного анализа от человека-эксперта. Для более точного соответствия определению метрики может быть необходимо дополнение метода более точными техниками анализа данных.

3. *Представление глубинной семантики*

Метаграф — это графовая иерархическая структура, описываемая следующим образом:

$$M = \langle V, MV, E, ME \rangle, \tag{3}$$

где V — это множество вершин метаграфа, MV — множество его метавершин, E — множество рёбер метаграфа, ME — его метарёбер [1, 4].

Если говорить о семантическом применении данной модели, то можно сопоставить этим элементам определённые значения. Вершины чаще всего обозначают простые сущности, обозначенные в текстах как субъекты и объекты действий. Сами действия соответствуют рёбрам метаграфа, они описывают связи и взаимодействия сущностей, выраженные на естественном языке в документах. Метавершины могут использоваться для обозначения более абстрактных и семантически широких понятий, включающих в себя конкретные объекты-вершины. Метарёбра могут объединять действия в более широко плановые события.

Атрибуты всех элементов чаще всего служат для типизации или характеристики действия или сущности. Конечно, к этому не сводятся все возможные области применения метаграфовых структур: например, метавершины могут отображать и предложения, тексты, документы, метарёбра — связи внутри корпуса.

Возможно использование методов на основе оценки качества и количества вершин графа и их степеней

(количество связей), качество рёбер (с теми ли вершинами образовались связи), в целом ряд методов оценки связности графа при необходимости может быть переложен на семантический анализ [11]. К сожалению, мы не располагаем достаточной графовой теорией для метаграфов, чтобы адекватно экстраполировать понятие связности графа и её исследование на иерархические структуры нашей модели.

Метод Smatch, предложенный Шу Цаем, может служить примером графовой семантической метрики, которая осуществляет алгоритм жадного поиска с восхождением к вершине (hill-climbing). Он применяется к AMR представлениям текстов, предложений и фактов [2]. В отличие от SemBLEU, о котором говорилось выше, он не принимает во внимание вершины синтаксического графа, а только проходит по связям наиболее оптимальным путём и оценивает метки на рёбрах. В некотором роде это делает его более гибким по отношению к сущностям, которые будут обозначать семантику вершин, но гораздо более зависимым от точности воспроизведения семантики связей.

Глубинные семантические связи всё же и этим методом уловить будет сложно, поскольку для этого нужен очень точный поиск вглубь дерева, а также установление связей между деревьями, которого не предусмотрено в Smatch. Дополнительную трудность создаёт то, что имеющаяся концепция метаграфовой XML модели имеет мало общего с AMR представлениями фактов и предложений.

4. *Постановка эксперимента*

В качестве большой языковой модели использовалась GPT-4o. Для проверки метрик был создан искусственный набор данных (toy dataset), содержащий комплекты из трёх небольших документов (2–3 предложения), которые связаны между собой тематикой и скрытыми семантическими связями. На этих данных строился метаграф для нетривиального факта, который включал бы в себя связь между документами.

Данная задача решалась большой языковой моделью по методу one-shot — в запросе предоставляется один общий пример подготовленного нами метаграфа на основе одного комплекта документов из набора данных, а модель создаёт метаграфы для остальных комплектов по предоставленному образцу.

Для вычисления метрики BLEU использовалась функция *sentence_bleu*() из модуля *nltk* языка Python для обработки естественного языка. Разбиение текста на русском языке на отдельные слова и элементы текста производилось с помощью токенизатора модуля *razdel* языка Python.

В качестве модели-асессора также использовалась оценивающая модель GPT. Параметры, описание и критерии для каждой дискретной оценки в виде примеров метаграфа достойного присвоения конкретного значения предоставлялись вместе с запросом на оценку, включавшем комплект документов и полученный ранее от модели ответ в виде XML документа, содержащего факт в метаграфовом виде.

Результаты

При применении статистического метода для оценки эффективности большой языковой модели GPT-4o было получено среднее значение метрики BLEU = 0,7437.

Применение GPT-модели для оценки результата генерации факта в метаграфовом виде большой языковой моделью дало среднее значение нашей метрики $Q_{\text{модели-асессора}} = 4{,}8$.

Коэффициент корреляции Пирсона между оценкой, данной моделью машинного обучения, и значением статистической метрики $r = 0{,}7894$.

Обсуждение

В целом полученное значение BLEU = 0,7437, является неплохим результатом, поскольку максимум метрики — единица. Однако этот показатель не является высоким для SotA (State-of-the-Art, эталонной) большой языковой модели.

Мы видим достаточно высокую среднюю оценку от модели-судьи (4,8 балла), но не можем однозначно её интерпретировать — возможно модель просто «хвалит» другую большую языковую модель за хорошие ответы, не особо обращая внимания на недостатки точности семантики ответа [12]. Высокую оценку также можно объяснить способностью больших языковых моделей улавливать перефразирование, обращаться с синонимами и выявлять основную мысль и смысл, очищая от вспомогательных элементов. В целом, модель достаточно точно следовала данным ей критериям оценки.

Достаточна высокая положительная степень корреляции по Пирсону ($r = 0{,}7894$, значения r лежат в диапазоне $[-1; 1]$) позволяет предложить наличие прямой линейной связи между оценками статистической и большой языковой модели, что говорит в пользу того, что ответы модели-судьи носят не случайный, а закономерный характер и соотносятся с более точной математически метрикой.

Исходя из имеющихся преимуществ и недостатков современных подходов к оценкам текстовых данных, выдаваемых большими языковыми моделями, можно

предположить, что необходим некоторый новый тип метрики, направленный на включение глубинной семантики в процесс измерения эффективности используемых программных средств на основе машинного обучения.

Дообученные модели оценки могут послужить основой для такого метода, однако неоднозначность интерпретации их работы не позволяет сделать их единственным этапом в процессе анализа эффективности. Для уточнения и упорядочения критериев истинности можно обратиться к современным методам борьбы с галлюцинациями и ошибками в работе модели — системам *интеллектуальных агентов*.

Агенты предоставляют большим языковым моделям программные инструменты для выполнения действий, требующих большей точности [3]. Использование инструментов активируется через анализ инструкций и запроса пользователя по ключевым словам, фразам или шаблонам. В качестве инструментов для модели-асессора можно представить программную реализацию вычисления статистических метрик.

В связи с высокой популярностью генерации, дополненной поиском/Retrieval Augmented Generation (RAG) широко распространено использование векторных баз данных. Векторизованные данные (эмбеддинги) для текстов так же могут использоваться для проверки наличия определённых токенов и их комбинаций из фактов, выданных моделью, в самом корпусе [6].

Возможные вариации в формулировке фактологической информации можно учесть, используя встроенную и хорошо работающую функцию больших языковых моделей — нахождение синонимичных слов к заданному пользователю. Статистические методы, такие как BLEU таким образом, можно будет дополнить подборкой синонимов, наличие которых, даже используя простой модуль регулярных выражений *re*, можно проверить для текста, выданного моделью. На основе полученных данных можно собрать эталон ответа, включающий верные термины, повышая эффективность применения статистической метрики для проверки общей структуры документа.

В общем схему оценки ответа модели можно представить как последовательность шагов:

1. Проведение оценки модели с помощью, по возможности, нескольких больших языковых моделей в качестве судьи. Сравнение оценок этих моделей.
2. Применение статистических метрик к данным. Сравнение оценок от модели и от математических расчётов с помощью коэффициента корреляции.
3. Если корреляция незначительная или оценки между моделями машинного обучения не согласуются, то вызываем интеллектуальных агентов, которые могут предложить подходящий инструмент.
4. Агент осуществляет поиск синонимов и составляем новый эталон для статистической метрики.
5. Агент осуществляет поиск других доступных стат. оценок, которые могли бы подойти нам для решения конкретной задачи оценивания.
6. Агент инициирует поиск совпадений сущностей по векторным базам данных и дополняет ими оценочный запрос по необходимости.
7. Агент выносит окончательное решение об истинности или ложности выделенного факта, получив (или не получив) достоверное подтверждение на этапах 4–6.

Заключение

Проведено исследование подходов к оценке результатов извлечения глубинной семантики, содержащей нетривиальные факты, из текстов на естественном языке. Рассмотрены два основных подхода к пониманию метрик для больших языковых моделей: статистические методы (расстояние Левенштейна, BLEU и т.д.) и основанные на машинном обучении, использующие языковые модели в качестве асессоров или судей (LLM-as-a-judge). В обеих группах есть как свои положительные (связанные с точностью и гибкостью) стороны для применения к решению исследуемой задачи, так и отрицательные (связанные со слабой интерпретируемостью и привязкой к точным формулировкам). Успешно использована метаграфовая структура, как способ отображения глубинной семантики текстов.

Получены и усреднены значения метрики BLEU и оценки модели-судьи для искусственно созданного набора данных, состоящего из комплектов небольших текстов разной тематики. Модель показала высокую эффективность в решении задачи извлечения глубинной семантики, однако модель-асессор оценила работу относительно более высокими значениями своей метрики. Тем не менее анализ коэффициента корреляции Пирсона для двух метрик показал, что они в достаточной степени согласуются между собой в выставленных оценках.

Предложен возможный путь развития метрик для глубинной семантики на основе интеллектуальных агентов, сочетающих инструменты, основанные на статистических методах,

ЛИТЕРАТУРА

1. Белянова М.А., Ревунков Г.И., Афанасьев Г.И., Гапанюк Ю.Е. Автоматическая генерация вопросов на основе текстов и графов знаний // Динамика сложных систем — XXI век. 2020. Т. 14. № 4. С. 55–64.
2. Cai S., Knight K. Smatch: an evaluation metric for semantic feature structures //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013. p. 748–752.
3. Feng Y. et al. DiverseAgentEntropy: Quantifying Black-Box LLM Uncertainty through Diverse Perspectives and Multi-Agent Interaction //arXiv preprint. arXiv:2412.09572. 2024. 28p.
4. Gapanyuk Y.E. The semantic complex event processing based on metagraph approach //Biologically Inspired Cognitive Architectures 2019: Proceedings of the Tenth Annual Meeting of the BICA Society 10. — Springer International Publishing. 2020. — p. 99–104.
5. Janardhana Rao P. et al. An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance //International Conference on Communications and Cyber Physical Engineering 2018. Singapore: Springer Nature Singapore, 2024. p. 161–174.
6. Kornai A. Vector semantics. — Springer Nature, 2023. 273p.
7. Laskar M.T.R. et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations //Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. pp. 13785–13816.
8. Papineni K. et al. BLEU: a method for automatic evaluation of machine translation//ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002. pp. 311–318.
9. Pombal J. et al. M-Prometheus: A Suite of Open Multilingual LLM Judges //arXiv preprint. arXiv:2504.04953. 2025. 68p.
10. Reiter E. A structured review of the validity of BLEU //Computational Linguistics. 2018. V. 44(3). pp. 1–12.
11. Schneider P. et al. Evaluating large language models in semantic parsing for conversational question answering over knowledge graphs //arXiv preprint. arXiv:2401.01711. 2024. 11p.
12. Schroeder K., Wood-Doughty Z. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge //arXiv preprint. arXiv:2412.12509. 2024.12p.
13. Shankar S. et al. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences //Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. 2024. pp. 1–14.
14. Song L., Gildea D. SemBleu: A robust metric for AMR parsing evaluation //arXiv preprint. arXiv:1905.10726. 2019. 6p.
15. Stengel-Eskin E., Van Durme B. Calibrated interpretation: Confidence estimation in semantic parsing //Transactions of the Association for Computational Linguistics. 2023. V. 11. pp. 1213–1231.
16. Sun L. et al. Scieval: A multi-level large language model evaluation benchmark for scientific research //Proceedings of the AAAI Conference on Artificial Intelligence. 2024. v. 38. №. 17. pp. 19053–19061.