УДК 004.81

УВЕРЕННОСТЬ LLM ПРИ ПОСТРОЕНИИ СЕМАНТИЧЕСКИХ КЛАССИФИКАЦИЙ АРГУМЕНТОВ

Д. С. Ларионов[1] [0000-0001-9225-0901], Е. Н. Никитина[2] [0000-0002-6207-8693], И. В. Смирнов[3] [0000-0003-4490-2017]

[1–3]*Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия*

[3]*Российский университет дружбы народов имени Патриса Лумумбы, г. Москва, Россия*

[1]dslarionov@isa.ru, [2]yelenon@mail.ru, [3]ivs@isa.ru

Аннотация

Исследована проблема квантификации уверенности больших языковых моделей (Large Language Model, LLM) при автоматической семантической классификации аргументов при эмотивных предикатах. На материале русскоязычных сообщений социальных сетей проанализированы глаголы страха (*пугать, бояться* и др.) и эмоционального отношения (*нравиться, любить*) с семантическими ролями экспериенцера, каузатора и объекта. В работе дано сравнение самооценки уверенности LLM Claude Sonnet 4.5 с экспертной оценкой текстов рассуждений модели при классификации аргументов по тематической области «Здравоохранение». В эксперименте использована стратифицированная выборка из 300 примеров с применением цепочки рассуждений на русском языке и четырехступенчатой шкалы уверенности. Результаты показали умеренную корреляцию Спирмена между оценками эксперта и модели. Статистически значимая связь установлена только между самооценкой модели и фактической корректностью классификации, тогда как экспертная оценка лингвистических характеристик рассуждений не зависит от точности. Сделан вывод о том, что эксплицитные рассуждения LLM не связаны напрямую с самооценкой по степени уверенности и не влияют на процесс принятия решений; они могут являться важной функциональной частью пользовательского интерфейса, но не исследовательского.

***Ключевые слова**: семантическая роль, классификация аргументов, эмотивный предикат, большие языковые модели, рассуждение LLM, уверенность LLM.*

ВВЕДЕНИЕ

Автоматический анализ семантических ролей при эмотивных предикатах представляет особый интерес для задач мониторинга общественного мнения в социальных сетях и является ответом на запрос специалистов в области изучения психологии больших групп.

Эмотивные глагольные предикаты в широком смысле слова распадаются на разные семантические группы. Мы рассматриваем глаголы страха типа *пугать – пугаться, страшить – страшиться, ужасать – ужасаться* и др., а также их лексико-грамматические дериваты, и глаголы эмоционального отношения *любить, нравиться*. Они характеризуются специфической аргументной структурой с семантическими ролями ***экспериенцера*** (носителя эмоции) и ***каузатора*** (причины эмоции) либо ***объекта***, на который направлена эмоция, ср.: *Его*#Experiencer *пугают* [*темпы развития ИИ*]#Causator – *Нам*#Experiencer *нравится Коломна*#Object. Особенностью собственно эмотивных глаголов (к ним относятся исследуемые нами глаголы страха) являются наличие каузативного компонента в семантической структуре, значительная морфосинтаксическая вариативность [1], а также пропозитивность каузативного компонента (см. [2]), что усложняет процесс экспертного описания и автоматической обработки аргументно-предикатных структур. Однако, если эти особенности приняты во внимание, обращение к аргументам и их семантическим ролям позволяет не только проводить мониторинг эмоциональной составляющей текста, но и автоматически определять, с какими жизненными реалиями связаны проявления тех или иных эмоций [3]. Эти «жизненные реалии» можно подвергать семантическим (тематическим) классификациям с помощью больших языковых моделей (Large Language Model, LLM), обобщая предметные области, которые оказались связаны с обнаруживаемым эмоциональным фоном.

Современные подходы к разметке семантических ролей (Semantic Role Labeling, SRL) с использованием LLM демонстрируют перспективные результаты: исследования показали возможность применения моделей семейства GPT

и Claude для классификации текстов без дообучения [4]. При этом ключевой проблемой остается оценка достоверности предсказаний: модели склонны к чрезмерной уверенности даже при выдаче ошибочных ответов [5]. Метод цепочки рассуждений (chain-of-thought prompting) позволяет повысить качество классификации за счет эксплицитного пошагового обоснования [6].

Проблема квантификации уверенности LLM посредством промптинга активно изучается в последние годы. В работе [7] показано, что для моделей, дообученных с помощью RLHF (ChatGPT, GPT-4, Claude), вербализованные оценки уверенности, порожденные в виде выходных токенов, обладают лучшей калибровкой, чем условные вероятности модели, снижая ожидаемую ошибку калибровки примерно на 50%. В масштабном бенчмарк-исследовании [8] вербализованные оценки уверенности протестированы на 10 датасетах и 11 моделях; установлено, что надежность таких оценок существенно зависит от способа формулирования запроса: для небольших моделей оптимальны простые формулировки промптов, тогда как для крупных моделей комбинирование нескольких стратегий позволяет достичь отклонения калибровки порядка 7–10%. В [9] исследована неопределенность в объяснениях, непосредственно в порождаемых LLM (chain-of-thought); обнаружено, что модели демонстрируют сверхуверенность при вербализации неопределенности в тексте рассуждений: средняя вербализованная уверенность составила 94.46% по всем экспериментам. Отметим, что все перечисленные исследования проведены на английском языке: промпты, инструкции и генерируемые рассуждения были англоязычными. В настоящей работе промпт и генерируемые моделью рассуждения были составлены на русском языке, что позволило исследовать проявления уверенности и неуверенности модели в русскоязычном дискурсе и оценить, воспроизводятся ли выявленные закономерности, в частности разрыв между лингвистическими характеристиками рассуждений и самооценкой модели, при работе с типологически иным языком.

Настоящая работа является продолжением исследования [10], посвященного применению LLM в семантических классификациях аргументов. Ранее изучалось использование самооценки LLM по параметру уверенности/неуверенности, а данная работа фокусируется на сравнении экспертной оценки рассуждения LLM и самооценки модели, анализе тех ее решений, которые соответствуют зоне колебания, неопределенности экспертной оценки. Такой подход позволит

выявить преимущества и слабости в работе LLM и человека при классификации множества элементов, относящихся к определенным предметных областям. Исследование проведено на примере темы «Медицина и здравоохранение», обсуждаемой пользователями соцсетей (примеры даются в авторской орфографии), с применением LLM Claude Sonnet 4.5 для классификации аргументов, обозначающих причину или объект эмоции, при эмотивных предикатах.

МЕТОД И ДАННЫЕ

Постановка задачи и данные для эксперимента

Задача классификации сформулирована следующим образом: для заданного предложения с выделенным эмотивным предикатом и его аргументом (каузатором или объектом) требуется определить, относится ли данный аргумент к тематической области «Здравоохранение». Входными данными служат исходное предложение, предикат, аргумент и его семантическая роль (каузатор или объект). Выходом является бинарная метка принадлежности к целевой области. Мы сосредоточились именно на каузативной и объектной составляющих эмотивных высказываний, поскольку семантика экспериенцера носит более независимый от темы обсуждения и, соответственно, более универсальный характер. Это проявляется в том, что часть высказываний с экспериенцерами 1-го и 2-го лиц обслуживает отношения между говорящим и его адресатом [11], часть носит собственно тематический характер, что отражается в семантике каузатора или объекта, а в позиции экспериенцера, кроме вышеперечисленных лиц, появляются экспериенцеры 3-го синтаксического лица, самые интересные из которых – квантифицированные субъекты «все», «люди», «никто» и др. [12], а также медицинские исполнители либо власти.

Для решения задачи в промпт были включены: (1) список из 997 лемматизированных ключевых слов области здравоохранения; (2) критерии отнесения (совпадение, синонимы, гипонимы) и исключения (другая область, неоднозначность, абстрактные слова); (3) инструкция chain-of-thought с рассуждением на русском языке; (4) XML-разметка аргумента; (5) JSON-вывод с полями reasoning, related и confidence, где модель сначала генерирует текст рассуждения, а затем в отдельном поле выводит оценку уверенности по шкале: Certain (однозначные случаи), Likely Certain (сильные аргументы), Likely Uncertain (слабый контекст), Uncertain (неопределенность). Таким образом, самооценка уверенности не

включена в текст рассуждений, а порождается моделью как самостоятельный элемент ответа. В инструкции подчеркнут приоритет точной квантификации неопределенности. Метод разметки семантических ролей опирается на [13] и описан в [3].

Материалом исследования послужила выборка из корпуса сообщений из социальных сетей одного региона, собранных в рамках проекта по изучению семантических ролей при эмотивных предикатах. Исходный массив содержал 4973 примера целевых аргументов. После предобработки (исключения прямых совпадений и нецелевых категорий) осталось 1762 примера. Для эксперимента методом стратифицированной выборки с оптимизацией баланса классов было отобрано 300 примеров.

Выборка включает предикаты групп *бояться* (34.7%), *нравиться* (28.3%), *любить* (25.7%), *пугать* (8.0%) и др. По семантическим ролям: Object – 54.0%, Causator – 46.0%. Аргументы-каузаторы (т. е. зависимые предикатов страха) чаще относятся к здравоохранению (13.0% против 3.7%), что объясняется семантикой предикатов и временными рамками корпуса (период Covid-19 в 2020–2022 гг.): страхи часто вызваны болезнями, медицинскими процедурами и социальными мерами, связанными с ковидом.

Далее исследуемые аргументы (`#Causator`, `#Object`) были размечены экспертом-лингвистом по критерию отнесенности/отсутствия отнесенности к области здравоохранения и медицины с использованием расширенной схемы меток (см. табл. 1), предполагающих сомнение либо отсутствие однозначного решения.

Табл. 1. Схема разметки и количество размеченных примеров.

Метка	Значение	Количество примеров
`0`	Не относится к здравоохранению	221 (73.7%)
`1`	Относится к здравоохранению	19 (6.3%)
`Wra`	Неверно выделенный аргумент, не относится	21 (7.0%)
`1wra`	Неверно выделенный аргумент, относится	5 (1.7%)
`0?`	Неоднозначно, скорее не относится	17 (5.7%)
`1?`	Неоднозначно, скорее относится	16 (5.3%)

Метками `wra` отмечаются случаи, когда на этапе автоматического выделения аргументов была допущена ошибка (например, захвачен неполный или некорректный фрагмент), однако эксперт все же мог оценить принадлежность к тематической области. Метки с вопросительным знаком (`0?`, `1?`) обозначают случаи неоднозначности (*А мне Скотарева нравится, хоть грубовата но все объяснит что да как* – `1?`) либо неуверенности (*Пусть объяснит эти числа, посмотрим, кого она больше боится – губернатора или ФМБА* – `0?`).

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Сопоставление самооценки модели с экспертной оценкой уверенности рассуждений

В табл. 2 приведены показатели эксплицитной уверенности, выдаваемой моделью, т. е. самооценки модели по признаку уверенность/неуверенности, шкалированному в промпте.

Табл. 2. Уверенность модели.

Уровень уверенности	*N*	Доля
Certain	226	85.0%
Likely Certain	28	10.5%
Likely Uncertain	5	1.9%
Uncertain	7	2.6%

С целью независимой валидации самооценки уверенности модели был проведен слепой эксперимент по экспертной аннотации. Модель выводила собственную оценку уверенности в процессе генерации ответа в соответствии с инструкцией. Эксперт-лингвист получал только текст рассуждений модели без доступа к оцениваемому моделью контексту, итоговому ответу модели, ее собственной оценке уверенности и экспертной метке классификации. Задача эксперта состояла в (1) эмпирическом выборе критериев оценки степени уверенности модели и (2) собственно оценке степени уверенности, проявленной моделью в ходе рассуждения.

Модальности уверенности/неуверенности в естественном языке выделяются на том основании, что эти значения соотносятся с семантикой ирреальных

наклонений [14] и могут выражаться в русском языке лексически: *может быть, кажется, точно, наверняка, вероятно, возможно* и т. д. Выражение неуверенности говорящего относительно содержания собственного высказывания соотносимо с эпистемической возможностью, которая «выражает неполноту знаний говорящего» [14], ср.: ***Похоже**, вода холодная; А настоящего Мюллера так и не нашли, **вероятно**, он нашел свою тихую ферму с бассейном* [Форум: 17 мгновений весны (2005–2010)].

В представленном для аннотирования материале рассуждений модели практически не обнаружено признаков колебания, сомнения, неуверенности и подобных ментальных состояний, маркированных стандартными языковыми средствами; рассуждения довольно определенны и четки. Поэтому экспертом была разработана шкала степеней уверенности, опирающаяся на те лингвистические характеристики, которые представлены в оцениваемых рассуждениях. Эта шкала состоит из трех ступеней: **уу** – абсолютная уверенность, **у** – уверенность, **у-н** – недостаточная уверенность (точнее, уверенность, что принять окончательное решение невозможно). Шкала формализована следующим образом. Ступень **уу** – рассуждение содержит особые оценочные слова, усилительные элементы – лексические показатели уверенности (категоричности): «*явно* указывает», «*однозначный* случай», «*ни один* из элементов не связан *напрямую*», «не имеет *никакого* отношения» и др. Ступень **у** – рассуждение не содержит признаков категоричности: «Ни один из этих глаголов не присутствует в списке ключевых слов концепта «Здравоохранение» и не является специфичным для данной семантической области», рассуждение может содержать слова уступки: «Само слово «сидеть» является общеупотребительным глаголом и не входит в специальную лексику здравоохранения, *хотя* в данном контексте оно используется в ситуации, связанной со здравоохранением». Ступень **у-н** – обычно модель прибегает к аргументу о недостаточности контекста; возможно наличие формы сослагательного (одного из ирреальных) наклонений: «Без дополнительного контекста, который *бы связывал* этот «бизнес» с медициной, лечением, больницами или другими аспектами здравоохранения, нет оснований считать аргумент релевантным целевой концепции»; встречаются слова с семантикой потенциальности и уступки: «*Несмотря* на *возможный* контекст здравоохранения, сам аргумент семантически не связан с этой областью».

Полученные экспертные оценки были сопоставлены с самооценкой уверенности модели и фактической корректностью классификации для анализа предсказательной силы обоих источников информации об уверенности, а также для выявления контекстов с расхождениями между оценками эксперта и модели (см. табл. 3). Было аннотировано 298 из 300 примеров; для основного анализа были использованы 265 примеров с определенными экспертными метками.

Табл. 3. Сопоставление уверенности по оценке эксперта и самооценке модели.

Экспертная оценка	Certain	Likely Certain	Likely Uncertain	Uncertain	Итого
уу (абсолютная уверенность)	153	8	1	0	162
у (уверенность)	77	23	2	0	102
у-н (недостаточная уверенность)	12	10	4	8	34

Коэффициент ранговой корреляции Спирмена между оценками эксперта и модели составил $r = 0.447$ ($p < 0.001$), что свидетельствует об умеренном согласии. При этом наблюдается характерная асимметрия: эксперт оценил рассуждения как абсолютно уверенные в 54.4% случаев, тогда как модель присвоила максимальный уровень (Certain) в 81.2%. Основную зону расхождений составили 12 случаев, в которых модель оценила свое рассуждение как максимально уверенное, а эксперт оценил рассуждение как недостаточно уверенное: 11 из них приходятся на местоименные аргументы (*их*, *он*, *это*, *такое*). В этой зоне расхождений ярко проявляется та общая характеристика рассуждений модели, которая обозначена экспертом как отсутствие маркированных признаков неуверенности даже в категории **у-н** (неоднозначность контекста, необходимость дополнительных сведений), при этом модель корректно классифицирует эти фрагменты как не относящиеся к здравоохранению.

Установленное умеренное согласие между оценками уверенности согласно эксперту и модели, а также выявленные зоны систематических расхож-

дений ставят вопрос о том, какой из двух источников информации об уверенности обладает большей предсказательной силой в отношении фактической корректности классификации. Для ответа на этот вопрос были найдены показатели точности модели в разрезе каждого уровня уверенности обеих шкал, а также ранговые корреляции между оценками уверенности и корректностью (см. табл. 4).

Табл. 4. Предсказательная сила экспертной и машинной оценок уверенности.

Показатель	**Экспертная оценка**	**Машинная оценка**
Корреляция Спирмена с корректностью (r)	0.042	0.296
p-value	0.495	< 0.001
χ^2 (уверенность × корректность)	1.42 (df = 2, p = 0.49)	23.6 (df = 3, p < 0.001)
Odds ratio (макс. уверенность vs. остальные)	1.36 (p = 0.73)	19.7 (p < 0.001)

Наблюдается отсутствие статистически значимой связи между экспертной оценкой уверенности рассуждений и фактической корректностью классификации (r = 0. 042, p = 0.495). Точность модели остается стабильно высокой вне зависимости от экспертной оценки: 97.4% для уровня **уу**, 97.6% для **у** и 93.6% для **у-н**. Напротив, эксплицитная оценка модели демонстрирует статистически значимую корреляцию с корректностью (r = 0.296, p < 0.001; отношение шансов 19.7 для уровня Certain по сравнению с остальными, p < 0.001).

Проанализируем результаты сопоставления (см. табл. 5).

Табл. 5. Сравнение уверенности модели по самооценке и экспертной оценке.

Параметры сравнения	Самооценка модели	Экспертная оценка
Количество ступеней на шкале	4	3
Оцениваемая характеристика	Уверенность/неуверенность	Степени уверенности
Характер шкалы	Дедуктивно-логическая	Индуктивно-эмпирическая
Предмет оценки	Неизвестно (текст рассуждения или внутренний ход принятия решения)	Текст рассуждений

Способ репрезентации и контроль	Неизвестно	Специальные языковые средства
Автор шкалы	Инструктор модели	Эксперт
Заполненность шкалы	Полная	Полная
Сфера сомнения в эксперименте	Почему модель использовала все 4 ступени шкалы	Количество экспертов

Эксплицитная оценка уверенности модели (ее самооценка), а с ней и надежности предсказания, не коррелирует с текстом рассуждений, т. е. с его лингвистическими характеристиками (для русского языка), доступными оценке человека, и поэтому не может быть воспроизведена экспертом при анализе обоснования, при этом между решениями модели и ее самооценкой есть статистически значимое согласие.

Возникают следующие вопросы: (1) почему в шкалах самооценки и экспертной оценки разное количество ступеней; (2) почему шкалы семантически разные (уверенность/неуверенность у модели – поле уверенности у эксперта); (3) почему модель посчитала необходимой и достаточной именно четырехступенчатую шкалу, т. е. применила все возможные ступени шкалы? Ответ на первый и второй вопросы состоит в том, что модель следовала дедуктивной схеме инструктора, не оспаривая ее, а эксперт полагался на реальные языковые характеристики рассуждений модели. Из этого следует, что если модель и «испытывает» неуверенность при принятии решения, то не выражает ее в своем рассуждении или выражает внелингвистическими способами. Отвечая на третий вопрос, можно предположить формальный характер распределения ответов модели по шкале уверенности/неуверенности: она буквально следовала инструкции, не оставив на шкале пробелов.

При этом модель ошибается нечасто, но если ошибается, то текст рассуждений выглядит столь же убедительно, как и в корректных случаях. Собственная оценка (не)уверенности модели успешно маркирует зону повышенного риска ошибок. Поэтому, вероятно, неуверенность может присутствовать в скрытом от нас процессе продуцирования решения, а в нашем эксперименте мы наблюдали ее рефлексы, но вопрос о градуировании зоны неуверенности остается открытым.

Конфликт между экспертной оценкой рассуждения модели и ее собственной самооценкой, с одной стороны, и достаточно высоким уровнем продукции, с другой, возвращает нас к ситуации, описанной в работе [15]. В этой работе оценка экспертами-психологами рассуждений LLM (без учета параметра уверенности), сопровождавших ее работу по идентификации депрессивных эссе в коллекции текстов авторов с различными психиатрическими статусами, была ниже, чем качество самой продукции модели.

Зона неуверенности (неопределенности)

Рассмотрим примеры неоднозначных высказываний, выделенных экспертом-лингвистом, и сравним экспертное решение с решением модели. Этот подкласс примеров был исключен из анализа [10], куда попали только однозначные случаи.

Примеры с неоднозначностью/неопределенностью делятся на две группы: 1) применительно к первым эксперт посчитал, что аргумент с семантикой каузатора или объекта не относится к медицине или непонятен, однако при опоре на контекст можно предположить, что высказывание «скорее относится» (1?) либо «скорее не относится» к медицинской тематике (0?); 2) во вторую группу попали примеры с некорректными разборами аргументов, однако по контексту эксперт смог определить принадлежность высказывания к теме: аргумент выделен неправильно, высказывание относится к медицине (1wra), аргумент выделен неправильно, высказывание не относится к медицине (wra). Всего в зону неопределенности попали 59 примеров, что составляет около 20% от выборки: 34 в первую группу (11%), 25 во вторую (8%).

В первой группе обозначилась зона рассогласования и согласования решений эксперта и модели. При этом в зоне рассогласования (21 пример) примеры относятся к медицине, хотя модель ответила «нет», – 15 случаев, примеры относятся к медицине, хотя эксперт ответил «нет», – 6 случаев. Доминирование компетентности эксперта над компетентностью модели можно объяснить тем, что материалом является контент соцсетей, в которых обсуждение медицинской тематики происходит посредством неполных предложений, особых разговорных формул, понятных человеку, но недоступных модели: *Видимо администрации города нравится смотреть на людей в намордниках* (*намордник* в новом значении)*; да, развалили всю медицину, медики, бедные, не справлются, тесты*

берут черед одного, наши все высокопоставленные боятся потерять стул, на котором сидят! *Вот это и пугает, что подхода никакого нет, приходи и вколем, а дальше будут твои трудности, как ты перенесешь прививку* (позицию каузатора занимает местоимение, катафорически отсылающее к дистанцированному фрагменту высказывания). Ошибки эксперта связаны, очевидно, с энциклопедическим знанием (ФМБА – неизвестная эксперту аббревиатура), с невнимательностью (*Многие и прививки от гриппа боятся и других тоже*) или неопределенностью инструкции для обширных, композиционно расчлененных, неоднозначных контекстов: *Темпы прироста пугают своей опасностью и риском // Правда, глава Роспотребнадзора забыла сообщить, что за полтора месяца страшного распространения страшного нового штамма от него зафиксирована только одна смерть*.

К зоне согласования (13 примеров) отнесены случаи условно положительного и отрицательного ответов эксперта, с которым коррелирует соответствующий ответ модели (*Знай, тебя мы не боимся, пока дома отсидимся, с книгами и интернетом, да и всей семьей при этом*). Любопытно, что в последнем случае получен согласованный ответ «нет», хотя внимательное рассмотрение убеждает, что контекст содержит рефлексы позитивной реакции на собственные страхи говорящего, связанные с ковидом и изоляцией.

В группе с неправильно определенным аргументом ответы модели и эксперта совпали: положительный ответ – 5 примеров, отрицательный – 20 примеров. Такой результат объясняется тем, что неверно выделенный аргумент либо однозначно относится к медицине: *Елена, не хочу вас пугать, но при коронавирусе коньюктивит – это один из симптомов*, либо однозначно не относится: *Кто-то боится мышей, кто-то высоты, кто-то темноты, а кто-то собак*.

Обобщенно по обеим группам согласованные ответы модели и эксперта в зоне неопределенности были даны в 38 случаях, что составляет около 64%.

ЗАКЛЮЧЕНИЕ

Настоящее исследование показало, что для русскоязычных рассуждений модели имеется разрыв между лингвистическими характеристиками порождаемого ею текста и ее самооценкой в плане уверенности/неуверенности и, в то же время согласованность самооценки модели с корректностью ее ответов, что может свидетельствовать об орнаментальности, невключенности эксплицитных

рассуждений модели в процесс принятия решений. В связи с этим встает вопрос о том, насколько реалистично или иллюзорно взаимодействие с моделью, не начинает ли галлюцинировать исследователь, полагающийся на эксплицитные рассуждения как на источник информации при диалоге с моделью.

Частные наблюдения за рассуждениями могут иметь перспективу для исследования. Модель выделяет местоименные аргументы, противопоставляя их полнознаментальной лексике. Для анафорических местоимений модель пытается восстановить полнознаментальный референт, что обогащает содержательную сторону исследования. В случае неудачи (отсутствия в анализируемом фрагменте предтекста) модель сопровождает свой ответ «запросом» на отсутствующий контекст. Этот «запрос» может стать базой для дальнейшей ресегментации текста с целью восстановления анафоры.

Проведенный эксперимент установил зону рассогласования ответов и, соответственно, сферы большей компетентности LLM и эксперта. Первая связана с неограниченностью энциклопедических знаний и стабильностью работы модели. Вторая связана с бытовыми эллиптическими контекстами, отражающими «неканонические» способы обсуждения медицинской тематики, что легко отождествляется человеком, и с синтаксически сложно устроенными контекстами (анафоро-катафорические связи), известными носителю языка и неподвластными LLM.

Таким образом, для получения надежного результата следует использовать пул экспертов и более определенно формулировать исследовательскую задачу.

Благодарности

Исследование выполнено в рамках научной программы Национального центра физики и математики, направление № 9 «Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах».

СПИСОК ЛИТЕРАТУРЫ

1. *Sonnenhauser B.* The Event Structure of Verbs of Emotion in Russian // Russian Linguistics. 2012. Vol. 36. P. 331–353. https://doi.org/10.1007/s11185-010-9060-9

2. *Золотова Г.А.* Очерк функционального синтаксиса русского языка.

М.: Наука, 1973. 351 с.

3. *Smirnov I.V., Larionov D.S., Nikitina E.N., Kazachonok G.A*. LLM for Semantic Role Labeling of Emotion Predicates in Russian // Supercomputing Frontiers and Innovations. 2025. Vol. 12, No. 3. P. 31–46. https://doi.org/10.14529/jsfi250303.

4. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report. 2024.

5. *Xiong M. et al.* Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs // arXiv: 2306.13063. 2023. https://doi.org/10.48550/arXiv.2306.13063

6. *Wei J. et al.* Chain-of-thought prompting elicits reasoning in large language models // Advances in neural information processing systems. 2022. Vol. 35. P. 24824–24837.

7. *Tian K. et al.* Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback // Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023. P. 5433–5442. https://doi.org/10.18653/v1/2023.emnlp-main.330

8. *Yang D., Tsai Y.H., Yamada M.* On Verbalized Confidence Scores for LLMs // arXiv: 2412.14737. 2024. https://doi.org/10.48550/arXiv.2412.14737

9. *Tanneru S.H., Agarwal C., Lakkaraju H.* Quantifying Uncertainty in Natural Language Explanations of Large Language Models // Proc. of the 27th International Conference on Artificial Intelligence and Statistics. PMLR, 2024. Vol. 238. https://proceedings.mlr.press/v238/harsha-tanneru24a.pdf (дата доступа 05.05.2026)

10. *Ларионов Д.С., Никитина Е.Н., Смирнов И.В.* Применение больших языковых моделей для семантической классификации аргументов при эмотивных предикатах // Системы высокой доступности. 2026. Т. 22, № 1. С. 12−16. https://doi.org/ 10.18127/j20729472-202601-02

11. *Никитина Е.Н., Онипенко Н.К.* Семантика и прагматика высказываний с эмотивными предикатами // Сибирский филологический журнал. 2022. № 2. С. 271–285. https://doi.org/10.17223/18137083/79/19

12. *Никитина Е.Н., Станкевич М.А., Ларионов Д.С*. Эмотивные глаголы с отрицанием: опыт интерпретации данных автоматического анализа текстов //

Медиалингвистика. 2025. Т. 12, № 2. С. 174–198.
https://doi.org/10.21638/spbu22.2025.201

13. *Lyashevskaya O., Kashkin E.* FrameBank: A Database of Russian Lexical Constructions // AIST 2015. Springer, 2015. P. 350–360.

14. *Падучева Е.В.* Модальность. Материалы для проекта корпусного описания русской грамматики (rusgram.ru). На правах рукописи. М. 2016.

15. *Kuzmin G., Strepetov P., Stankevich M., Chudova N., Shelmanov A., Smirnov I.* Exploring Large Language Models for Detecting Mental Disorders // Proc. of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. P. 34535–34559. https://doi.org/10.18653/v1/2025.emnlp-main.1752

LLM CONFIDENCE IN CONSTRUCTING SEMANTIC CLASSIFICATIONS OF ARGUMENTS

D. S. Larionov[1] [0000-0001-9225-0901], E. N. Nikitina[2] [0000-0002-6207-8693],
I. V. Smirnov[3] [0000-0003-4490-2017]

[1–3]*Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia*

[3]*Peoples Friendship University of Russia named after Patrice Lumumba, Moscow, Russia*

[1]dslarionov@isa.ru, [2]yelenon@mail.ru, [3]ivs@isa.ru

Abstract

This article addresses the problem of quantifying the confidence of large language models (LLMs) in the automatic semantic classification of arguments with emotive predicates. Using Russian-language social media posts, we analyze verbs of fear (to scare, to be afraid, etc.) and emotional attitude (to like, to love) with the semantic roles of experiencer, causator, and object. The study aims to compare self-assessed confidence of LLM Claude Sonnet 4.5 with expert assessments of the model's reasoning texts in argument classification in the healthcare domain. The experiment utilized a stratified set of 300 examples using chain-of-thoughts reasoning in Russian and four-step confidence scale. The results showed a moderate Spearman correlation between the expert and model assessments. A statistically significant relationship was found

only between self-assessment of the model and the actual classification accuracy, while expert assessment of the linguistic characteristics of reasoning was unrelated to accuracy. It was concluded that explicit LLM reasoning is not directly related to its self-assessment of confidence and is separate from the decision-making process; reasoning may be an important functional part of the user interface, but not of the research.

Keywords*: semantic roles, argument classification, verb of emotion, large language models, LLM reasoning, LLM confidence.*

Acknowledgements**:** This study was conducted within the framework of the scientific program of the National Center for Physics and Mathematics, section No. 9 “Artificial intelligence and big data in technical, industrial, natural and social systems”.

REFERENCES

1. *Sonnenhauser B.* The Event Structure of Verbs of Emotion in Russian // Russian Linguistics. 2012. Vol. 36. P. 331–353. https://doi.org/10.1007/s11185-010-9060-9.

2. *Zolotova G.A.* Ocherk funkcional`nogo sintaksisa russkogo yazy`ka. M.: Nauka, 1973. 351 p.

3. *Smirnov Ivan V., Larionov Daniil S., Nikitina Elena N., Kazachonok Grigory A*. LLM for Semantic Role Labeling of Emotion Predicates in Russian // Supercomputing Frontiers and Innovations. 2025. Vol. 12, No. 3. P. 31–46. https://doi.org/10.14529/jsfi250303

4. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report. 2024.

5. *Xiong M. et al.* Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs // arXiv: 2306.13063. 2023. https://doi.org/10.48550/arXiv.2306.13063

6. *Wei J. et al.* Chain-of-thought prompting elicits reasoning in large language models //Advances in neural information processing systems. 2022. Vol. 35. P. 24824–24837.

7. *Tian K.et al.* Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023. P. 5433–5442.

https://doi.org/10.18653/v1/2023.emnlp-main.330

8. *Yang D., Tsai Y.H., Yamada M.* On Verbalized Confidence Scores for LLMs // arXiv: 2412.14737. 2024. https://doi.org/10.48550/arXiv.2412.14737

9. *Tanneru S.H., Agarwal C., Lakkaraju H.* Quantifying Uncertainty in Natural Language Explanations of Large Language Models // Proc. of the 27th International Conference on Artificial Intelligence and Statistics. PMLR, 2024. Vol. 238. https://proceedings.mlr.press/v238/harsha-tanneru24a.html (accessed 05.05.2026)

10. *Larionov D.S., Nikitina E.N., Smirnov I.V*. Primenenie bol`shix yazy`kovy`x modelej dlya semanticheskoj klassifikacii argumentov pri e`motivny`x predikatax // Sistemy` vy`sokoj dostupnosti. 2026. Vol. 22, No. 1. P. 12−16.

https://doi.org/10.18127/j20729472-202601-02

11. *Nikitina E.N., Onipenko N.K.* Semantika i pragmatika vyskazyvanij s emotivnymi predikatami // Sibirskij filologicheskij zhurnal. 2022. № 2. P. 271–285.

https://doi.org/10.17223/18137083/79/19.

12. *Nikitina E.N., Stankevich M.A., Larionov D.S*. E`motivny`e glagoly` s otriczaniem: opy`t interpretacii danny`x avtomaticheskogo analiza tekstov // Medialingvistika. 2025. T. 12. №2. S. 174–198. https://doi.org/10.21638/spbu22.2025.201

13. *Lyashevskaya O., Kashkin E.* FrameBank: A Database of Russian Lexical Constructions // AIST 2015. Springer, 2015. P. 350–360.

14. *Paducheva E.V.* Modal'nost'. Materialy dlya proekta korpusnogo opisaniya russkoj grammatiki (rusgram.ru). Na pravah rukopisi. M. 2016.

15. *Kuzmin G., Strepetov P., Stankevich M., Chudova N., Shelmanov A., Smirnov I.* Exploring Large Language Models for Detecting Mental Disorders // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. P. 34535–34559. https://doi.org/10.18653/v1/2025.emnlp-main.1752

СВЕДЕНИЯ ОБ АВТОРАХ

Даниил Сергеевич ЛАРИОНОВ – программист Федерального исследовательского центра «Информатика и управление» Российской академии наук, специализируется на компьютерной лингвистике и машинном обучении. Научные интересы включают обработку естественного языка, семантическую разметку ролей, обучение и оптимизацию больших языковых моделей (SFT, RLVR, GRPO), автоматическую оценку генерации текста и многоязычные NLP-системы. Ведет исследования в области семантической классификации и аннотирования эмотивных предикатов с применением LLM. Соавтор научных публикаций по семантико-ролевой разметке русского языка.

Daniil S. LARIONOV – programmer at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, specializing in computational linguistics and machine learning. His research interests include natural language processing, semantic role labeling, training and optimization of large language models (SFT, RLVR, GRPO), automatic evaluation of natural language generation, and multilingual NLP systems. He conducts research on semantic classification and annotation of emotion predicates using LLMs. He is a co-author of publications on semantic role labeling for the Russian language.

email: dslarionov@isa.ru

ORCID: 0000-0001-9225-0901

Елена Николаевна НИКИТИНА – кандидат филологических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук. Научные интересы: русская грамматика, функциональная грамматика, грамматические категории, лингвистический анализ текста, интеллектуальный анализ текста, семантические классификации.

Elena N. NIKITINA – senior research fellow at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. Research interests: Russian grammar, functional grammar, grammatical categories, linguistic text analysis, text mining, semantic classifications.

email: yelenon@mail.ru

ORCID: 0000-0002-6207-8693

Иван Валентинович СМИРНОВ – доктор технических наук, доцент, заведующий отделом «Интеллектуальный анализ информации» Федерального исследовательского центра «Информатика и управление» Российской академии наук. Научные интересы: обработка естественного языка, интеллектуальный анализ текстов, психолингвистический анализ текстов, предикатно-аргументная семантика.

Ivan V. SMIRNOV – Doctor of Technical Sciences, Associate Professor, head of Department “Intelligent Processing of Information, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences. Research interests: natural language processing, text mining, psycholinguistics, predicate-argument semantics.

email: ivs@isa.ru

ORCID: 0000-0003-4490-2017

Материал поступил в редакцию 5 апреля 2026 года