УДК 004.89

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В ЗАДАЧЕ ОЦЕНКИ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ СЛОВОУПОТРЕБЛЕНИЙ

Д. В. Кокосинский[0000-0001-5642-8725]

Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия

deniskokosss@mail.ru

Аннотация

В работе исследована применимость больших языковых моделей (Large Language Model, LLM) для решения задачи оценки семантической близости значений слова в паре словоупотреблений, известная как Word-in-Context (WiC) с опорой на мультиязычный бенчмарк CoMeDi. Предложены новые подходы к построению автоматических WiC-систем на основе LLM, в частности, конфигурация в которой предсказания LLM корректируются по обучающей выборке, но дообучения LLM при этом не требуется. Выполнено систематическое сравнение пяти различных конфигураций WiC-систем на основе LLM с точки зрения качества и с учетом вычислительных затрат. Результаты на тестовых выборках из семи языков показали, что предложенные подходы позволяют LLM превзойти все существующие специализированные системы, установив новый уровень качества на бенчмарке CoMeDi. Тем не менее достигнутое высокое качество сопряжено со значительным ростом вычислительных затрат: системы на базе LLM требуют на несколько порядков больше вычислений по сравнению с компактными специализированными моделями (такими как XL-DURel). Настоящая работа является одним из шагов к пониманию компромисса между точностью и ресурсоемкостью при использовании современных LLM в задачах лексической семантики.

Ключевые слова*: Word-in-Context, большая языковая модель, обработка естественного языка.*

ВВЕДЕНИЕ

Задача WiC (Word-in-Context, дословный перевод «Слово-в-контексте») заключается в оценке семантической близости значений некоторого целевого

слова в паре контекстов его употребления. К примеру, требуется оценить близость значений слова «пост» в паре словоупотреблений: «на *посту* никого не было» и «эта мысль более подробно раскрыта в *посте*».

Автоматические WiC-системы широко применяются в качестве промежуточного шага для решения различных задач. Например, в задаче выявления значений слов (Word Sense Induction, WSI) для группировки словоупотреблений по значениям целевого слова применяются методы кластеризации поверх предсказаний WiC-системы [1]. В задаче обнаружения семантического сдвига (Lexical Semantic Change Detection, LSCD) [2] предсказания WiC-систем используются для обнаружения слов, значительно изменивших свое значение между двумя временными периодами. Автоматическое решение задачи обнаружения семантического сдвига особенно актуально в рамках исторической лингвистики [2].

В настоящей работе задача WiC рассмотрена в формулировке DURel [3], где задача WiC формализована как задача порядковой классификации (ordinal classification). Автоматической WiC-системе предложено выставить оценку близости по дискретной шкале от 1 (значения полностью различны) до 4 (значения идентичны). Например, для целевого слова «пост» и двух словоупотреблений «на посту никого не было» и «эта мысль более подробно раскрыта в посте» по шкале DURel должна ставиться оценка 1.

Во многих задачах автоматической обработки естественного языка долгое время доминировали модели, обученные для решения конкретной задачи. В последнее время во многих из этих задач большие языковые модели (Large Language Models, LLM) общего назначения превзошли специализированные модели, например в машинном переводе [4] и других задачах понимания естественного языка [5].

При решении задачи WiC до сих пор наилучшее качество демонстрируют узкоспециализированные модели, такие как DeepMistake [6], XL-LEXEME [7] или XL-DURel [8]. Эти модели основаны на предобученных трансформерах-энкодерах BERT [9] или XLM-R [10]. В других работах предложены решения задачи WiC с помощью LLM [11–14]. Однако ни в одной из этих работ не было однозначно показано, что LLM превосходят специализированные модели в решении задачи WiC. В [13] прямое сравнение со специализированными моделями не проводи-

лось. В [14] сравнение выполнено не непосредственно для задачи WiC, а в качестве компоненты систем обнаружения семантического сдвига: на английском языке LLM с автоматической оптимизацией инструкций достигали качества специализированных моделей, однако на испанском качество было хуже. В [12] модель GPT-4 в режиме без настройки (zero-shot) оказалась незначительно хуже XL-LEXEME по качеству для решения задачи WiC для английского языка. В работе [11] LLM показали заметно худшее качество по сравнению со специализированными моделями на мультиязычном бенчмарке CoMeDi (α = 0.499 против α = 0.660 у лучшей узкоспециализированной модели).

С учетом успешного применения LLM в других задачах обработки естественного языка [4, 5] мы предположили, что и задача WiC не должна являться исключением. В настоящей работе предложены различные подходы к построению WiC-систем на основе LLM. Варианты системы исследованы с точки зрения качества и вычислительных затрат. Основные полученные результаты таковы:

– предложена WiC-система на основе LLM, которая превосходит по качеству все существующие решения на мультиязычном WiC-бенчмарке CoMeDi [15];

– оценен вклад различных факторов (количество параметров LLM, режим рассуждения и формат инструкции) в качество решения задачи WiC для семейства моделей Qwen3.5 [16].

БЛИЗКИЕ ПО ТЕМАТИКЕ РАБОТЫ

Для оценки качества работы WiC-систем нами использован бенчмарк CoMeDi [15], ставший основой одноименного соревнования. Бенчмарк опирается на ручные WiC-аннотации с использованием четырехбалльной шкалы DURel [3] и включает две подзадачи: (1) OGWiC – предсказание медианной оценки аннотаторов для пары употреблений; (2) DisWiC – предсказание среднего попарного несогласия аннотаторов. В настоящей работе рассмотрена только подзадача OGWiC. Данные в бенчмарке CoMeDi охватывают семь языков (китайский, английский, немецкий, норвежский, русский, испанский и шведский) и содержат около 48 тыс. обучающих, 8 тыс. валидационных и 15 тыс. тестовых примеров. Один пример – это пара словоупотреблений с разметкой по шкале DURel. Разбивка лексическая: множества целевых лемм в обучающей, валидационной и тестовой выборках не пересекаются [15]. Качество на подзадаче OGWiC оценено

с помощью порядковой версии показателя Криппендорфа α [17], который оценивает степень согласия предсказаний с истинными ответами с поправкой на вероятность их случайного совпадения. Значение α = 1 соответствует идеальному предсказанию, α = 0 означает полностью случайное угадывание, а отрицательные значения α указывают на систематическое несогласие с разметкой. Кроме того, порядковая версия метрики учитывает «расстояние» между классами: она значительно строже штрафует систему за грубые ошибки (например, когда вместо 4 предсказана 1), чем за незначительные промахи (3 вместо 4), что делает ее оптимальной для задач порядковой классификации [15, 18].

На бенчмарке CoMeDi лучшие результаты демонстрируют специализированные модели, основанные на XLM-Roberta-large [10] — предобученном трансформере-энкодере с 350 млн параметров. Система deep-change [19] стала победителем соревнования CoMeDi, среднее значение целевой метрики α = 0.656 по семи тестовым выборкам. Система основана на модели DeepMistake [6]: бинарный WiC-классификатор на базе XLM-R, дообученный на задачу WiC на различных наборах размеченных данных. Авторы deep-change показали, что дообучение модели DeepMistake на 4-классовую классификацию на обучающей выборке CoMeDi дает худшие результаты по сравнению с применением пороговой дискретизации к выходам модели, обученной на 2-классовую классификацию [19]. При дискретизации три порога, разделяющие непрерывный отрезок [0, 1] на четыре интервала, подобраны методом Нелдера – Мида и максимизируют α на валидационной выборке. Другая WiC-система XL-DURel [8] показала лучший результат после соревнования, среднее значение α = 0.67. XL-DURel использует двухдольную архитектуру на базе XLM-R-large, дообученную со специальной функцией потерь AnglE Loss [20]; косинусное сходство предсказаний также дискретизируется с помощью алгоритма Нелдера – Мида на валидационной выборке. Еще одна система ABDN-NLP [11], также участник CoMeDi 2025, применяет LLM GPT-4o-mini с инструкцией с примерами (few-shot, *n* = 20); дополнительно исследован взвешенный выбор примеров по частоте классов или сложности. Результат ABDN-NLP – α = 0.499 – заметно хуже, чем у deep-change (α = 0.656).

Кроме того, существуют другие подходы к WiC на основе LLM, не протестированные в CoMeDi. В [13] предложено ручное итеративное улучшение инструкции для GPT-4 для задачи аннотации семантической близости на DURel-шкале 1–

4. Лучшая конфигурация достигает α = 0.54 на тестовой выборке DWUG EN [21]. Прямого сравнения со специализированными WiC-моделями не проводилось. В [14] представлено систематическое сравнение LLM (Llama 3.1 8B, Mixtral 8x7B, Llama 3.3 70B) и узкоспециализированных WiC-моделей (XL-LEXEME [7], DeepMistake [6]) в задаче обнаружения семантического сдвига на английском и испанском языках. Предложенные WiC-системы на основе LLM использованы в качестве компоненты LSCD-системы. Формат инструкций для LLM автоматически оптимизирован с помощью MIPROv2 (из DSPy [22]). Модель Llama 3.3 70B с оптимизированной инструкцией достигла качества специализированных моделей в задаче LSCD на английском, однако на испанском специализированные модели остаются лучшими. Прямого сравнения со специализированными моделями в задаче WiC не приведено.

Принципильной трудностью в задаче WiC в постановке DURel является размытость границ между оценками семантической близости по шкале от 1 до 4. С этим возникают трудности даже у людей-аннотаторов [3]. В системах-бейзлайнах соревнования CoMeDi [15], deep-change [19] и XL-DURel [8] для «выравнивания» предсказаний моделей со шкалой DURel применялась дискретизация: модель предсказывает вещественное число (вероятность положительного класса или косинусное сходство), на валидационной выборке методом Нелдера – Мида подбираются три порога, максимизирующих α, и по этим порогам непрерывное значение преобразуется в оценку 1–4. В работах, связанных с LLM [11, 13, 14], задача решалась через небольшое количество примеров в инструкции (in-context learning), показывающих LLM-примеры пар словоупотреблений и соответствующие им оценки.

Наши эксперименты в основном полагаются на LLM с открытыми весами из семейства Qwen3.5 [16]. Семейство Qwen3.5 включает модели различных размеров, есть варианты как с плотной (dense) архитектурой, так и с архитектурой MoE (Mixture-of-Experts, «смесь экспертов»). В случае плотной архитектуры для генерации каждого выходного токена в вычислениях участвуют все параметры модели (обозначение Qwen3.5-4b означает, что модель имеет 4 млрд параметров), а в случае MoE для каждого выходного токена использовано лишь некоторое подмножество параметров (обозначение Qwen3.5-122b-A10b означает, что

всего в модели 122 млрд параметров, на каждый токен используется только 10 млрд). На различных бенчмарках модели семейства Qwen3.5 показали наилучшие результаты среди моделей с открытыми весами схожего размера [16]. Модели Qwen3.5 поддерживают режим рассуждения. Это означает, что перед финальным ответом такие модели генерируют текстовую цепочку рассуждений. Режим рассуждения, как правило, улучшает метрики качества, но значительно увеличивает вычислительные затраты [16]. В моделях Qwen3.5 этот режим можно отключить.

WiC-СИСТЕМЫ НА ОСНОВЕ LLM

Рассмотрим пять основных конфигураций системы, решающих задачу WiC в постановке DURel с помощью LLM. В каждой из этих конфигураций может применяться любая LLM. Единственное ограничение состоит в том, что некоторые конфигурации требуют наличия в модели режима рассуждения. Важно отметить, что формат инструкций (prompt) всегда дается на английском языке, примеры в инструкции также не переводятся. В предварительных экспериментах перевод инструкций на язык бенчмарка не оказал влияния на качество.

Базовая конфигурация. В ней LLM оценивает семантическую близость от 1 до 4 напрямую, без примеров. Режим рассуждений отключен. Инструкция выглядит следующим образом[1]:

```
Определи, имеет ли '{target}' одинаковое значение в следующих предложениях. Являются ли они Одинаковыми (4), разными, но Родственными (3), отдаленно Связанными (2) или не связанными, Различными значениями слова (1)? Ответь одним числом.
Предложение 1: {sent1}
Предложение 2: {sent2},
```

где `target` – целевая лемма, `sent1` – первое словоупотребление целевой леммы, `sent2` – второе.

[1] В тексте приведены переводы инструкций с английского языка на русский, в самих методах использованы англоязычные инструкции вне зависимости от языка целевого слова и словоупотреблений. Оригиналы инструкций представлены в Приложении 1.

«Базовая + Рассуждения». Инструкция остается такой же, как в базовой конфигурации, но режим рассуждения не отключен. В некоторых ранних работах [13, 14] использованы инструкции вида «цепочка рассуждений», где LLM подталкивают «думать по шагам перед ответом», однако мы от них отказались в пользу встроенного режима рассуждения.

«С примерами». Использована инструкция из [13] без изменений. Значения от 1 до 4 «объяснены» модели через четыре специально подобранных показательных примера, по одному на каждое из четырех значений шкалы. Приведены примеры для целевых слов *bank* (банк/берег), *tree* (дерево-растение/дерево-структура данных), *cell* (ячейка/клетка) и *horse* (лошадь). Полный текст инструкции приведен в Приложении 1.

«Со шкалой 0–10». В этой конфигурации мы адаптируем идею дискретизации из систем deep-change [19] и XL-DURel [8] для использования с LLM. В инструкции вместо шкалы от 1 до 4 использована шкала от 0 до 10:

```
Насколько близки значения слова '{target}' в следующих предложениях. Оцени по шкале от 0 (Абсолютно несвязанные значения) до 10 (Идентичные значения). Ответь одним числом.
Предложение 1: {sent1}
Предложение 2: {sent2}
```

Чтобы перевести предсказания из шкалы от 0 до 10 в шкалу от 1 до 4, в этой конфигурации использована обучающая выборка CoMeDi на соответствующем языке. Сначала собираются 11-балльные предсказания от выбранной LLM для каждого из примеров в обучающей выборке, разметка в которой 4-балльная. Далее выбираются три различных пороговых значения, определяющие правило преобразования 10-балльной шкалы в 4-балльную. Для выбора пороговых значений производится полный перебор и выбирается тот набор, который при применении к предсказаниям LLM дает наибольшее значение показателя α на обучающей выборке. Иными словами, в данной конфигурации LLM делает предсказания в промежуточной 11-балльной шкале, которые затем преобразуются в 4-балльные, причем использовано оптимальное правило преобразования, найденное на обучающей выборке. Таким образом, шкала предсказания LLM «выравнивается» со шкалой, получаемой в результате работы аннотаторов.

«Со шкалой 0–10 + рассуждения». Этот подход повторяет предыдущий, но использован режим с рассуждениями как для обучающей, так и для тестовой выборок. В конфигурации «со шкалой 0–10» режим с рассуждениями отключен на всех выборках.

Отдельно отметим алгоритмы автоматической оптимизации формата инструкции. Как показали авторы работы [14], автоматическая оптимизация инструкции через MIPROv2 DSPy [22] заметно увеличила качество WiC для модели Llama 3.3 70B. Однако для моделей Qwen3.5-4b и Qwen3.5-9b оптимизатор MIPROv2 на той же обучающей выборке не позволил улучшить качество даже на самой этой выборке[2]. По этой причине подобные конфигурации в дальнейшем не рассматриваются.

Кратко рассмотрим предложенные конфигурации с точки зрения вычислительных расходов. Базовая конфигурация и конфигурация «с примерами» требуют один проход через LLM для каждого входного примера. У узкоспециализированных моделей также требуется только один проход, но у LLM заметно больше параметров (например, 4 млрд у Qwen3.5-4b против 350 млн у XLM-R-large). Для конфигурации «со шкалой 0–10» дополнительно требуется обработать каждый пример из обучающей выборки (на этапе подбора порогов). Режим рассуждения, как уже было отмечено, значительно увеличивает вычислительные затраты.

ЭКСПЕРИМЕНТЫ НА ВАЛИДАЦИОННЫХ ВЫБОРКАХ

Для экспериментов были взяты валидационные выборки бенчмарка CoMeDi. Мы рассмотрели выборки на английском и русском языках из-за ограниченности вычислительных ресурсов. Целевым показателем качества Криппендорфа является α. Исследованы пять конфигураций, описанных в предыдущем разделе.

Эксперименты были проведены на моделях из семейства Qwen3.5 различных размеров. Были выбраны модели с плотной архитектурой: 0.8b, 4b, 9b, а также с архитектурой «смесь экспертов» (MoE): 122b-A10b, 397b-A17b. Для мо-

[2] Использованы оригинальный код и данные обучения из работы [14].

делей с архитектурой MoE из-за значительных вычислительных затрат эксперименты с включенным режимом рассуждения не проводились. Кроме того, модель Qwen3.5-0.8b во всех конфигурациях оказалась хуже решений-бейзлайнов соревнования CoMeDi, поэтому далее она не рассматривается.

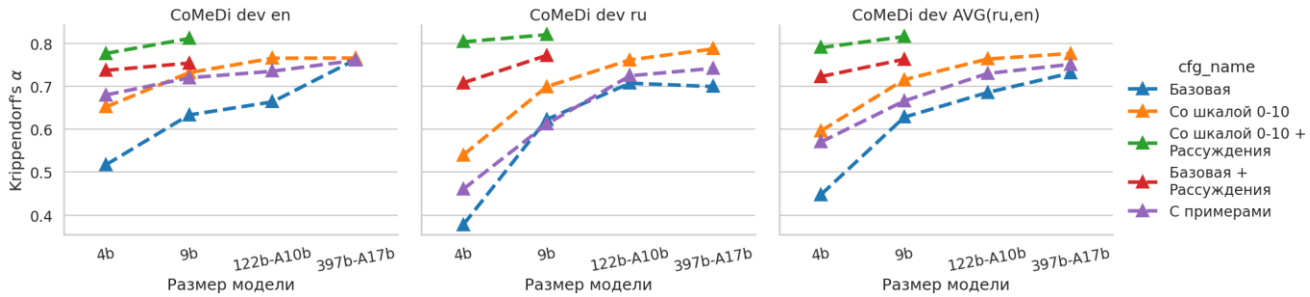

Рис. 1. Качество (α) на бенчмарке CoMeDi для 5 конфигураций в зависимости от размера модели. Слева – бенчмарк на русском языке, посередине – на английском. Справа – среднее по двум бенчмаркам. Конфигурации с рассуждениями для 122b-A10b, 397b-A17b отсутствуют по причине высоких вычислительных затрат.

Результаты экспериментов представлены на рис. 1. Для всех размеров модели проведено ранжирование конфигураций по качеству. В порядке возрастания α оно таково: «базовая», «с примерами», «со шкалой 0–10», «базовая + рассуждения» и «со шкалой 0–10 + рассуждения», с наилучшим качеством. Наблюдается также стабильный прирост качества при увеличении размера модели с сохранением конфигурации. Однако в рассмотренных конфигурациях наблюдается эффект «убывающей отдачи»: переход с 4b на 9b приносит значительно более заметный прирост качества, чем переход с 122b-A10b на 397b-A17b.

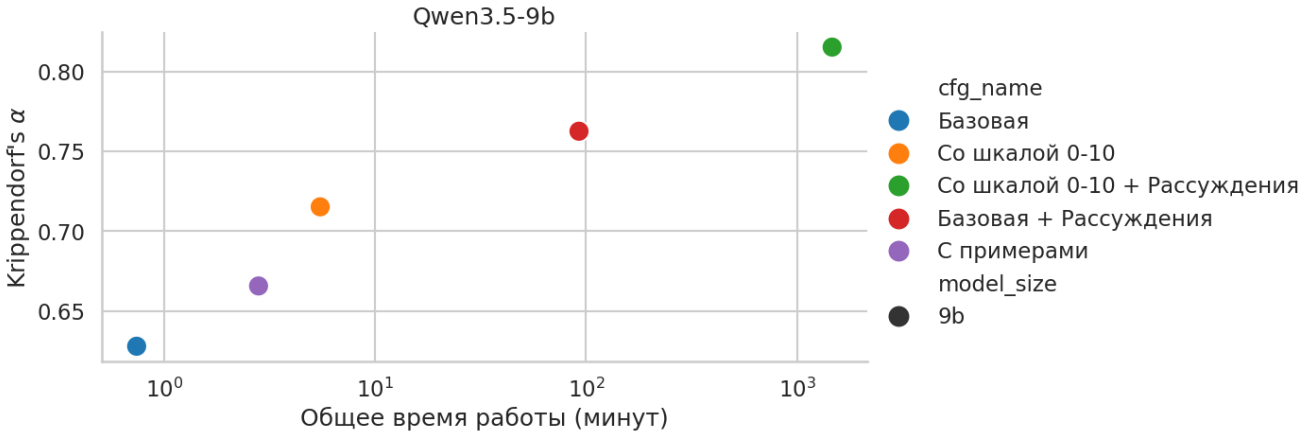

Рис. 2. Связь общего времени работы систем в различных конфигурациях и качества на валидационных бенчмарках для модели Qwen3.5-9b. Ось абсцисс логарифмическая. Использованы оптимизированный фреймворк LLM инференса vLLM [25] и один графический ускоритель RTX 5090 32GB.

Отметим, что ранжирование конфигураций по вычислительным затратам обратно ранжированию по качеству. Это продемонстрировано для модели Qwen3.5-9b на рис. 2. По времени работы конфигурации различаются на порядки. Так, время работы конфигурации «со шкалой 0–10 + рассуждения» превосходит время работы базовой конфигурации более чем в тысячу раз при одинаковом размере модели. В целом использование режима рассуждений оказалось определяющим фактором в вычислительных затратах, увеличивая их более чем в сто раз по сравнению с аналогичными конфигурациями с отключенным режимом.

РЕЗУЛЬТАТЫ НА ТЕСТОВЫХ ВЫБОРКАХ

Исследовав различные конфигурации на двух валидационных выборках, для финальных экспериментов на тестовых выборках для семи языков мы выбрали несколько вариантов WiC-систем на основе LLM. Были использованы конфигурации: «базовая», «с примерами» и «со шкалой 0–10» с моделью Qwen3.5-397b-A17b, а также конфигурации «базовая + рассуждения» и «со шкалой 0–10 + рассуждения» с моделью Qwen3.5-9b. Такой выбор продиктован объемом вычислительных затрат. Дополнительно были протестированы модели с закрытыми весами, точнее, использованы конфигурации «с примерами» и «со шкалой 0–10» и моделью GPT-5.3-Chat, моделью без режима рассуждения. Для оптими-

зации порогов в конфигурации GPT-5.3-Chat «со шкалой 0–10» были использованы только 10% обучающей выборки. Даже ограниченные эксперименты показали, что существенного влияния на качество такое сужение обучающей выборки не оказывает.

Результаты работы указанных WiC-систем по сравнению с существующими решениями представлены в табл. 1. Конфигурации GPT-5.3-Chat «с примерами», GPT-5.3-Chat «со шкалой 0–10» и Qwen3.5-9b «со шкалой 0–10 + рассуждения» превзошли по качеству все существующие решения по среднему значению на семи бенчмарках. Варианты системы с GPT-5.3-Chat опередили все предыдущие решения на всех бенчмарках, кроме испанского языка. Система GPT-5.3-Chat «со шкалой 0–10» установила новый уровень качества (SOTA) по среднему α по семи бенчмаркам.

Табл. 1. Результаты на тестовых выборках бенчмарка CoMeDi. Лучший результат из прошлых работ выделен жирным шрифтом.

WiC-система	EN	RU	ZH	DE	NO	ES	SV	Среднее
Верхняя граница	0.97	0.96	1.00	0.88	0.94	0.96	0.95	0.95
Бейзлайн 1 [15]	0.10	0.11	0.06	0.27	0.12	0.18	0.02	0.12
Бейзлайн 2 [15]	0.65	0.55	0.38	0.73	0.52	0.66	0.60	0.58
Deep-change [19]	**0.73**	0.62	0.42	0.72	**0.67**	0.75	0.68	0.66
XL-Durel [8]	**0.73**	**0.66**	**0.44**	**0.74**	0.65	**0.76**	**0.69**	**0.67**
ABDN-NLP [11]	0.62	0.48	0.21	0.67	0.43	0.57	0.51	0.50
Qwen3.5-397b-A10b «Базовая»	0.68	0.61	0.27	0.71	0.49	0.63	0.58	0.57
Qwen3.5-397b-A10b «С примерами»	0.70	0.68	0.41	0.66	0.64	0.66	0.53	0.61
Qwen3.5-397b-A10b «Со шкалой 0–10»	0.74	0.72	0.37	0.78	0.59	0.65	0.71	0.65
Qwen3.5-9b «Со шкалой 0–10 + Рассуждения»	**0.81**	0.76	0.43	0.80	0.65	0.75	0.76	0.71
GPT-5.3-chat «С примерами»	**0.81**	0.75	**0.47**	0.75	**0.72**	0.69	0.79	0.71
GPT-5.3-chat «Со шкалой 0–10»	**0.81**	**0.78**	0.44	**0.82**	0.70	0.74	**0.80**	**0.73**

Лучший результат по всем системам подчеркнут.

Здесь можно отметить прогресс LLM как таковых. В работе [13] результаты конфигурации «с примерами» с использованием модели GPT-4 признаны «неудовлетворительными», а в наших экспериментах та же конфигурация, но

с моделью GPT-5.3-Chat, уже заметно превзошла все системы на основе специализированных WiC-моделей. Можно также отметить эффективность использования процедуры подбора порога на обучающих выборках в конфигурациях «со шкалой 0–10»: она дала прирост α на +0.08 для Qwen3.5-397b-A17b и на +0.02 для GPT-5.3-Chat.

Однако следует отметить и существенный недостаток предлагаемых систем, уже упомянутый в предыдущем разделе, — это вычислительные издержки. По приблизительной оценке, работа, например, системы XL-DURel [8] требует приблизительно в сто тысяч раз меньше вычислений по сравнению с конфигурацией Qwen3.5-9B «со шкалой 0–10 + рассуждения». В случае моделей с закрытыми весами большие вычислительные затраты выражаются в высокой суммарной стоимости запросов к API. Мы полагаем, что дообучение LLM на задачу WiC подобно системам XL-DURel, DeepMistake или XL-LEXEME может позволить снизить вычислительные издержки, сохранив при этом высокое качество, но оставляем проверку этого предположения на будущее.

АНАЛИЗ ОШИБОК

Дополнительно разберем ошибки WiC-системы GPT-5.3-chat «со шкалой 0–10» на русскоязычной части бенчмарка CoMeDi-ru. Рассмотрим «грубые» ошибки, т. е. ошибки с аннотацией 1 и предсказанием 4 или, наоборот, с аннотацией 4 и предсказанием 1. Таких ошибок на CoMeDi-ru двадцать семь, а общая точность на примерах с аннотацией 1 или 4 составила 87%.

В 11 из 27 примеров спорными, на наш взгляд, выглядят как аннотация, так и предсказания модели, например, в паре словоупотреблений «*Хор* похвал и восхищений...» и «Ой, дубинушка, охнем – устало подхватывал *хор*» выставлены аннотация 4 и предсказание 1 или в паре «*Загрузка* активной зоны ядерного реактора: 66 т металлического урана» и «рациональный коэффициент *загрузки* оборудования», выставлены аннотация 4 и предсказание 1. Для этих примеров более адекватными представляются промежуточные оценки 2 или 3.

В других примерах заметна неспособность WiC-системы уловить переносный смысл в устойчивом выражении. Например, в выражении «за кулисами»: «за кулисами официальной повестки» и «я всегда приходила за кулисы смотреть» системой выставлена оценка 4. С выражением «за кулисами» связаны три

«грубые» ошибки, еще три связаны с другими устойчивыми выражениями. Кроме того, можно отметить излишнее внимание WiC-системы к общему стилю или тематике предложений, например, в паре «ее гнетут любопытные взгляды барынь, рассматривающих ее как оригинальный *монстр*» и «спросил я, наблюдая за мечущимися по экрану *монстрами*» была выставлена оценка 1 при аннотации 4. В будущем подобные случаи можно адресовать отдельно, например, дополнив инструкцию LLM.

ЗАКЛЮЧЕНИЕ

Проведено систематическое исследование применимости современных больших языковых моделей (LLM) для решения задачи Word-in-Context (WiC) на мультиязычном бенчмарке CoMeDi. Показано, что при правильном подборе конфигурации LLM способны превзойти специализированные модели (такие как deep-change и XL-DURel), которые ранее считались лучшими для этой задачи.

Установлено, что использование непрерывной шкалы оценки (от 0 до 10) с последующим подбором порогов дискретизации на обучающей выборке, а также включение режима рассуждений вносят наибольший вклад в повышение качества. Конфигурация Qwen3.5-9b «со шкалой 0–10» и рассуждениями показала наилучшие результаты среди моделей с открытыми весами. Модель с закрытыми весами GPT-5.3-Chat продемонстрировала еще более высокое качество даже без режима рассуждения. Тем не менее использование LLM сопряжено с многократно возросшими вычислительными затратами, что делает их применение дорогим по сравнению с компактными специализированными энкодерами. Дальнейшие исследования могут быть направлены на дообучение LLM для сохранения баланса между качеством и вычислительной эффективностью.

СПИСОК ЛИТЕРАТУРЫ

1. *Fedorova M., Mickus T., Partanen N., Siewert J., Spaziani E., Kutuzov A.* AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling // Proc. of the 5th Workshop on Computational Approaches to Historical Language Change. Bangkok, Thailand, August 2024. Association for Computational Linguistics, 2024. P. 72–91. https://doi.org/10.18653/v1/2024.lchange-1.8

2. *Tahmasebi N., Borin L., Jatowt A., Xu Y., Hengchen S.* Computational Approaches to Semantic Change. Berlin: Language Science Press, 2021. https://doi.org/10.5281/zenodo.5040241

3. *Hendy A. et al.* How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation // arXiv:2302.09210 [cs.CL], 2023. https://doi.org/10.48550/arXiv.2302.09210

4. *Liang P. et al.* Holistic Evaluation of Language Models // arXiv: 2211.09110 [cs.CL], 2022. https://doi.org/10.48550/arXiv.2211.09110

5. *Schlechtweg D., Schulte im Walde S., Eckmann S.* Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 2. New Orleans, Louisiana, June 2018. Association for Computational Linguistics, 2018. P. 169–174. https://doi.org/10.18653/v1/N18-2027.

6. *Arefyev N., Fedoseev M., Protasov V., Homskiy D., Davletov A., Panchenko A.* DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model // Computational Linguistics and Intellectual Technologies (Dialog 2021). Russian Federation, 2021. 20. P. 16–30.

7. *Cassotti P., Siciliani L., DeGemmis M., Semeraro G., Basile P.* XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Vol. 2. Toronto, Canada, July 2023. Association for Computational Linguistics, 2023. P. 1577–1585. https://doi.org/10.18653/v1/2023.acl-short.135.

8. *Yadav S., Schlechtweg D.* XL-DURel: Finetuning Sentence Transformers for Ordinal Word-in-Context Classification // arXiv:2507.14578 [cs.CL], 2025. https://doi.org/10.48550/arXiv.2507.14578

9. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805 [cs.CL], 2018. https://doi.org/10.48550/arXiv.1810.04805

10. *Conneau A. et al.* Unsupervised Cross-lingual Representation Learning at Scale // arXiv:1911.02116 [cs.CL], 2019. https://doi.org/10.48550/arXiv.1911.02116

11. *Loke Y.X., Schlechtweg D., Zhao W.* ABDN-NLP at CoMeDi Shared Task: Predicting the Aggregated Human Judgment via Weighted Few-Shot Prompting // Proc.

of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 122–128.

12. *Periti F., Tahmasebi N.* A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change // Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. Mexico City, Mexico, June 2024. Association for Computational Linguistics, 2024. P. 4262–4282.
https://doi.org/10.18653/v1/2024.naacl-long.240

13. *Yadav S., Choppa T., Schlechtweg D.* Towards Automating Text Annotation: A Case Study on Semantic Proximity Annotation using GPT-4 // arXiv:2407.04130 [cs.CL], 2024. https://doi.org/10.48550/arXiv.2407.04130

14. *Zamora-Reina F.D., Bravo-Marquez F., Schlechtweg D., Arefyev N.* Can Large Language Models Compete with Specialized Models in Lexical Semantic Change Detection? // ECAI 2025. Frontiers in Artificial Intelligence and Applications. Vol. 413. IOS Press, 2025. P. 4201–4208. https://doi.org/10.3233/FAIA251313.

15. *Schlechtweg D., Choppa T., Zhao W., Roth M.* CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments // Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 33–47.

16. *Yang A. et al.* Qwen3 Technical Report // arXiv:2505.09388 [cs.CL], 2025. https://doi.org/10.48550/arXiv.2505.09388

17. *Krippendorff K.* Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks, CA: SAGE Publications, 2018.

18. *Sakai T.* Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1. Online, August 2021. Association for Computational Linguistics, 2021. P. 175–187. https://doi.org/10.18653/v1/2021.acl-long.15

19. *Kuklin M., Arefyev N.* Deep-change at CoMeDi: The Cross-Entropy Loss is not All You Need // Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 48–64.

20. *Li X., Li J.* AnglE-optimized Text Embeddings // arXiv: 2309.12871. 2023. https://doi.org/10.48550/arXiv.2309.12871

21. *Schlechtweg D., Tahmasebi N., Hengchen S., Dubossarsky H., McGillivray B.* DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages // Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics, 2021. P. 7079–7091.
https://doi.org/10.18653/v1/2021.emnlp-main.567

22. *Khattab O. et al.* DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines // arXiv: 2310.03714 [cs.CL], 2023.
https://doi.org/10.48550/arXiv.2310.03714

23. *Shao Z. et al.* DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models // arXiv:2402.03300 [cs.CL], 2024.
https://doi.org/10.48550/arXiv.2402.03300

24. *DeepSeek-AI, Guo D. et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning // arXiv:2501.12948 [cs.CL].
https://doi.org/10.1038/s41586-025-09422-z

25. *Woosuk K. et al.* Efficient Memory Management for Large Language Model Serving with PagedAttention // Proc. of the ACM SIGOPS 29th Symposium on Operating Systems Principles. 2023. https://doi.org/10.1145/3600006.3613165

ПРИЛОЖЕНИЕ 1. ФОРМАТЫ ИСПОЛЬЗОВАННЫХ ИНСТРУКЦИЙ.

Базовая конфигурация

```
Determine whether '{target}' has the same meaning in the following sentences. Do they refer to roughly the Same (4), different but closely Related (3), distant/figuratively Linked (2) or unrelated Distinct word meanings (1)? Answer with a single number.
Sentence 1: {sent1}
Sentence 2: {sent2}
```

С примерами

```
You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic similarity of the target word in these sentences 1 and 2. Consider only the objects/concepts the word forms refer to: ignore any common etymology and metaphorical similarity! Ignore case! Ignore number (cat/Cats = identical meaning). Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is unrelated; 2 is distantly related; 3 is closely related; 4 is identical meaning. Your response should align with a human’s succinct judgment.
Example:
target word: bank
first sentence: His parents had left a lot of money in the bank and now it was all Measle's, but a judge had said the Measle was too young to get it.
second sentence: Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner.
Score: 1, let's think step by step, the annotation is 1 because in the first sentence "bank" is related to financial institutions, where people can deposit or withdraw their money, while in the second sentence "bank" refers to a bench or seat, typically found in parks or public spaces, where people can sit and rest.
Example:
target word: tree
first sentence: A binary tree is a data structure for storing organized data.
second sentence: The trees in that forest are rich in aroma and colors.
Score: 2, let's think step by step, The annotation is 2 because in the first sentence "tree" means a data structure in computing, and in the second sentence "trees" means the plant that has leaves. The meaning is loosely related because they share the shape of a tree, but in one case it is a data structure and in the second sentence, it is a tree.
Example:
target word: cell
first sentence: With this well-known program, we will be able to manipulate the data through its cells.
second sentence: As a result of this analysis, we will obtain a different map, a map where each visible cell comes with a number of 0 or 180 degrees.
Score: 3, let's think step by step, the annotation is 3 because in the first sentence the meaning of "cell" is related to grids where you can input information, and in the second sentence the meaning is also related to grids
```

```
where you can gather information. It is not 4 because the meanings differ in
that in one sentence they are used in a program and in the other sentence they
are used in a map.
Example:
target word: horse
first sentence: Marti's white horse gallops to victory.
second sentence: When horses are in a herd, they protect each other.
Score: 4, let's think step by step, in this case, the annotation is 4 because in
both sentences, the meaning of horse y horses is related to the animal. Please
note that horse is singular and horses is plural do not affect the annotation.
Given the following sentence pair score according the explained, give
only a numeric value as answer and do not explain why the chosen score:
target word: {target}
first sentence: {sent1}
second sentence: {sent2}
score:
```

Со шкалой 0–10

```
How close are the meanings of word with the lemma '{target}' in the
following sentences? Rate on a score of 0 (Completely unrelated meanings) to 10
(Identical meanings). Answer with a single number.
Sentence 1: {sent1}
Sentence 2: {sent2}
```

LARGE LANGUAGE MODELS FOR WORD-IN-CONTEXT

D. V. Kokosinskii[0000-0001-5642-8725]

Lomonosov Moscow State University, Moscow, Russia

deniskokosss@mail.ru

Abstract

Task-specific models have long dominated natural language processing tasks. However, general-purpose large language models (LLMs) have recently begun to successfully compete with highly specialized solutions across various NLP domains. In this paper, we investigate the applicability of LLMs to the task of estimating the semantic proximity of a word's meanings in a pair of usages, known as Word-in-Context (WiC). Drawing on the multilingual CoMeDi benchmark, we propose novel approaches to building automated WiC-systems based on LLMs. We conduct a systematic comparison of five different configurations in terms of quality and computational costs. In particular, we propose a configuration where LLM predictions are adjusted using a training set, without the need to fine-tune the LLM itself. Results

on test sets across seven languages show that our proposed approaches enable LLMs to outperform all existing specialized systems, establishing a new state-of-the-art (SOTA) on the CoMeDi benchmark. Nevertheless, the achieved high quality comes with a significant increase in computational costs: LLM-based systems require several orders of magnitude more computations compared to compact specialized models (such as XL-DURel). This work represents a step towards understanding the trade-off between accuracy and resource efficiency when using modern LLMs in lexical semantics tasks.

Keywords*: Word-in-Context, large language models, natural language processing.*

REFERENCES

1. *Fedorova M., Mickus T., Partanen N., Siewert J., Spaziani E., Kutuzov A.* AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling // Proc. of the 5th Workshop on Computational Approaches to Historical Language Change. Bangkok, Thailand, August 2024. Association for Computational Linguistics, 2024. P. 72–91. https://doi.org/10.18653/v1/2024.lchange-1.8

2. *Tahmasebi N., Borin L., Jatowt A., Xu Y., Hengchen S.* Computational Approaches to Semantic Change. Berlin: Language Science Press, 2021. https://doi.org/10.5281/zenodo.5040241

3. *Hendy A. et al.* How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation // arXiv:2302.09210 [cs.CL], 2023. https://doi.org/10.48550/arXiv.2302.09210

4. *Liang P. et al.* Holistic Evaluation of Language Models // arXiv: 2211.09110 [cs.CL], 2022. https://doi.org/10.48550/arXiv.2211.09110

5. *Schlechtweg D., Schulte im Walde S., Eckmann S.* Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 2. New Orleans, Louisiana, June 2018. Association for Computational Linguistics, 2018. P. 169–174. https://doi.org/10.18653/v1/N18-2027.

6. *Arefyev N., Fedoseev M., Protasov V., Homskiy D., Davletov A., Panchenko A.* DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model

// Computational Linguistics and Intellectual Technologies (Dialog 2021). Russian Federation, 2021. 20. P. 16–30.

7. *Cassotti P., Siciliani L., DeGemmis M., Semeraro G., Basile P.* XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Vol. 2. Toronto, Canada, July 2023. Association for Computational Linguistics, 2023. C. 1577–1585. https://doi.org/10.18653/v1/2023.acl-short.135

8. *Yadav S., Schlechtweg D.* XL-DURel: Finetuning Sentence Transformers for Ordinal Word-in-Context Classification // arXiv:2507.14578 [cs.CL], 2025. https://doi.org/10.48550/arXiv.2507.14578

9. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805 [cs.CL], 2018. https://doi.org/10.48550/arXiv.1810.04805

10. *Conneau A. et al.* Unsupervised Cross-lingual Representation Learning at Scale // arXiv:1911.02116 [cs.CL], 2019. https://doi.org/10.48550/arXiv.1911.02116

11. *Loke Y. X., Schlechtweg D., Zhao W.* ABDN-NLP at CoMeDi Shared Task: Predicting the Aggregated Human Judgment via Weighted Few-Shot Prompting // Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 122–128.

12. *Periti F., Tahmasebi N.* A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change // Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. Mexico City, Mexico, June 2024. Association for Computational Linguistics, 2024. P. 4262–4282. https://doi.org/10.18653/v1/2024.naacl-long.240

13. *Yadav S., Choppa T., Schlechtweg D.* Towards Automating Text Annotation: A Case Study on Semantic Proximity Annotation using GPT-4 // arXiv:2407.04130 [cs.CL], 2024. https://doi.org/10.48550/arXiv.2407.04130

14. *Zamora-Reina F. D., Bravo-Marquez F., Schlechtweg D., Arefyev N.* Can Large Language Models Compete with Specialized Models in Lexical Semantic Change Detection? // ECAI 2025. Frontiers in Artificial Intelligence and Applications. Vol. 413. IOS Press, 2025. P. 4201–4208. https://doi.org/10.3233/FAIA251313.

15. *Schlechtweg D., Choppa T., Zhao W., Roth M.* CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments // Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 33–47.

16. *Yang A. et al.* Qwen3 Technical Report // arXiv:2505.09388 [cs.CL], 2025. https://doi.org/10.48550/arXiv.2505.09388

17. *Krippendorff K.* Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks, CA: SAGE Publications, 2018.

18. *Sakai T.* Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 1. Online, August 2021. Association for Computational Linguistics, 2021. P. 175–187. https://doi.org/10.18653/v1/2021.acl-long.15

19. *Kuklin M., Arefyev N.* Deep-change at CoMeDi: The Cross-Entropy Loss is not All You Need // Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics, 2025. P. 48–64.

20. *Li X., Li J.* AnglE-optimized Text Embeddings // arXiv: 2309.12871. 2023. https://doi.org/10.48550/arXiv.2309.12871

21. *Schlechtweg D., Tahmasebi N., Hengchen S., Dubossarsky H., McGillivray B.* DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages // Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics, 2021. P. 7079–7091.
https://doi.org/10.18653/v1/2021.emnlp-main.567

22. *Khattab O. et al.* DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines // arXiv: 2310.03714 [cs.CL], 2023.
https://doi.org/10.48550/arXiv.2310.03714

23. *Shao Z. et al.* DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models // arXiv:2402.03300 [cs.CL], 2024.
https://doi.org/10.48550/arXiv.2402.03300

24. *DeepSeek-AI, Guo D. et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning // arXiv:2501.12948 [cs.CL]. https://doi.org/10.1038/s41586-025-09422-z

25. *Woosuk K. et al.* Efficient Memory Management for Large Language Model Serving with PagedAttention // Proc. of the ACM SIGOPS 29th Symposium on Operating Systems Principles. 2023. https://doi.org/10.1145/3600006.3613165

СВЕДЕНИЯ ОБ АВТОРЕ

КОКОСИНСКИЙ Денис Владиславович – аспирант факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

Denis V. KOKOSINSKII – PhD student at the Department of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

email: deniskokosss@mail.ru

ORCID: 0000-0001-5642-8725

Материал поступил в редакцию 4 апреля 2026 года