

УДК 004.8

МЕТОДЫ СНИЖЕНИЯ ОБЪЁМА ЭКСПЕРТНОЙ ВЕРИФИКАЦИИ ПРИ ПОСТРОЕНИИ ЭТАЛОННЫХ КОРПУСОВ АНАЛИТИЧЕСКИЙ ОБЗОР

*Кременчуцкий Анатолий Анатольевич,
Мардонов Шахриёр Зиёдулло угли*

Аннотация. Формирование эталонной (gold-standard) разметки корпусов знаний остаётся самым дорогим этапом создания интеллектуальных систем, поскольку каждое решение должно быть подтверждено экспертом предметной области. В статье систематизированы методы, позволяющие сократить объём экспертного подтверждения: активное обучение, слабая (программная) разметка, предразметка моделью с выборочной верификацией, краудсорсинг с вероятностной агрегацией и использование больших языковых моделей (LLM) в роли разметчиков. Для каждого класса методов определены механизм экономии, типичные метрики эффективности и точки, в которых теряется качество эталона. Показано, что существующие подходы оптимизируют либо стоимость, либо качество обучаемой модели, но редко дают гарантию качества самого эталона при заданном бюджете экспертных проверок. Сформулированы открытые задачи, в том числе для малоресурсных языков, к которым относится узбекский.

Ключевые слова: эталонная разметка, корпус знаний, экспертная верификация, активное обучение, слабая разметка, большие языковые модели, человек в контуре.

Abstract. Building gold-standard annotation of knowledge corpora is costly because every label must be confirmed by a domain expert. This paper reviews methods that reduce the volume of expert confirmation: active learning, weak (programmatic) supervision, model pre-annotation with selective verification, crowdsourcing with probabilistic aggregation, and LLM-based annotation. For each class we identify the source of savings, typical efficiency metrics, and the points where gold-standard quality is lost, and we outline open problems, including those of low-resource languages such as Uzbek.

Keywords: gold-standard annotation, knowledge corpus, expert verification, active learning, weak supervision, large language models, human-in-the-loop.

1. Введение

Качество систем извлечения знаний, классификации и поиска ограничено качеством эталонной разметки, на которой они обучаются и проверяются. Под эталонной разметкой здесь понимается множество пар «элемент корпуса — метка», в котором каждая метка признана верной компетентным экспертом и может служить мерой истины при оценке других разметчиков и моделей. Именно требование экспертного подтверждения отличает эталон от обычной обучающей выборки.

Экспертное подтверждение дорого по трём причинам. Во-первых, эксперты предметной области (медицина, право, педагогика, техника) малочисленны и загружены основной работой. Во-вторых, для надёжного эталона обычно требуется несколько независимых суждений и разрешение расхождений, что умножает затраты. В-третьих, для малоресурсных языков, к которым относится узбекский, дефицит экспертов, владеющих одновременно языком и предметной областью, особенно выражен.

За последние два десятилетия сложилось несколько направлений, сокращающих эту нагрузку: активное обучение [1, 2], слабая и программная разметка [3, 4], краудсорсинг с

вероятностной агрегацией меток [5, 6], а с 2021–2023 годов — использование больших языковых моделей как автоматических разметчиков [7–9]. Эти направления развивались в разных сообществах и оцениваются разными метриками, поэтому их трудно сопоставить именно с позиции задачи построения эталона.

Цель статьи — систематизировать методы сокращения объёма экспертного подтверждения по единому основанию: какой механизм даёт экономию, как она измеряется и где возникает риск потери качества эталона. Результат обзора служит основой для постановки задачи разработки алгоритмического и программного обеспечения, минимизирующего число экспертных проверок при гарантированном качестве эталонной разметки.

2. Постановка задачи

Пусть корпус содержит N элементов $x_1, \dots, x_N$, а $y^*(x)$ — истинная метка, которую дал бы эксперт. Автоматическая процедура (модель, набор правил, ансамбль, LLM) выдаёт предварительную метку $\hat{y}(x)$ и, возможно, оценку уверенности $c(x)$. Стратегия верификации S выбирает подмножество $V \subseteq \{1, \dots, N\}$ элементов, отправляемых эксперту; для них метка заменяется экспертной, для остальных принимается автоматическая.

Качество итогового эталона естественно измерять долей ошибок в нём:

$$\varepsilon(S) = (1/N) \cdot \sum_{i \notin V} \mathbb{I}[\hat{y}(x_i) \neq y^*(x_i)] \quad (1)$$

Тогда задача сокращения экспертного подтверждения формулируется как задача условной оптимизации: минимизировать число проверок $|V|$ при условии, что ошибка эталона с вероятностью не менее $1 - \delta$ не превышает допустимого уровня ε_0 :

$$\min_S |V| \text{ при } P(\varepsilon(S) \leq \varepsilon_0) \geq 1 - \delta \quad (2)$$

Двойственная постановка — максимизировать качество эталона при фиксированном бюджете B экспертных проверок ($|V| \leq B$). Принципиальная трудность в том, что $\varepsilon(S)$ неизвестна: она зависит от истинных меток непроверенных элементов. Поэтому любой практический метод должен либо оценивать её статистически по проверенной части, либо опираться на калиброванную уверенность $c(x)$.

Для сравнения методов в обзоре используются три величины: **доля сэкономленных проверок** $1 - |V|/N$; **остаточная ошибка эталона** $\varepsilon(S)$; **наличие гарантии** — даёт ли метод доказуемую или хотя бы эмпирически проверяемую оценку $\varepsilon(S)$, а не только качество обученной на эталоне модели.

3. Классы методов сокращения экспертной нагрузки

3.1. Активное обучение

Активное обучение сокращает число размечаемых экспертом элементов за счёт того, что модель сама выбирает наиболее информативные примеры [1]. Основные стратегии отбора — по неопределённости (энтропия, наименьший отрыв между двумя лучшими классами), по разногласию комитета моделей (query-by-committee) и по ожидаемому изменению модели; в глубоком обучении к ним добавились стратегии, учитывающие разнообразие и представительность выборки [2].

Ограничение для задачи построения эталона принципиальное: активное обучение отбирает примеры, полезные для **обучения модели**, а не для **контроля качества эталона**. Размеченная таким способом выборка смещена в сторону трудных случаев и не является репрезентативной, поэтому по ней нельзя напрямую оценить качество всего корпуса.

3.2. Слабая и программная разметка

В парадигме слабой разметки эксперт вместо меток отдельных элементов формулирует правила (функции разметки): ключевые слова, регулярные выражения, обращения к базам знаний, выходы других моделей. Генеративная модель оценивает точность и корреляции этих

источников без истинных меток и выдаёт вероятностную разметку; наиболее известная реализация — система Snorkel [3], обзор направления дан в [4].

Здесь экономия возникает за счёт переноса усилий эксперта с уровня элементов на уровень правил: одно правило размечает тысячи элементов. Однако результат — «зашумлённые» метки, пригодные для обучения, но не являющиеся эталоном без дополнительной проверки.

3.3. Предразметка моделью с выборочной верификацией

Модель размечает весь корпус, а эксперт проверяет только часть меток, отобранную по уровню уверенности или иному критерию. Этот подход наиболее близок к постановке из раздела 2, поскольку эксперт не создаёт метку, а подтверждает или исправляет готовую. Для поиска вероятных ошибок в уже размеченных данных разработаны методы «уверенного обучения» (confident learning), оценивающие совместное распределение наблюдаемых и истинных меток [14].

Эффективность подхода целиком зависит от калибровки уверенности: если модель уверена в ошибочных метках, они пройдут без проверки. Современные нейросети склонны к переуверенности, что требует дополнительной калибровки (например, температурного масштабирования) [18].

3.4. Краудсорсинг с вероятностной агрегацией

Экспертное суждение заменяется несколькими суждениями неспециалистов, которые затем агрегируются с учётом надёжности каждого исполнителя. Классическая модель Дэвида — Скина оценивает матрицы ошибок разметчиков методом ЕМ-алгоритма [5]; было показано, что для ряда лингвистических задач агрегированные метки нескольких неэкспертов сопоставимы с экспертными [6]. Согласованность разметчиков традиционно измеряется коэффициентами каппа и альфа Кrippendorфа [15].

Для корпусов знаний в специализированных областях метод ограниченно применим: неспециалисты систематически ошибаются в одних и тех же трудных случаях, и агрегация не устраняет коррелированную ошибку.

3.5. Большие языковые модели как разметчики

С появлением GPT-3 и последующих моделей LLM стали использоваться как автоматические разметчики в режиме нулевого или малого числа примеров. Уже в 2021 году было показано, что разметка с помощью GPT-3 при повторной разметке человеком только наименее уверенных элементов снижает стоимость разметки на 50–96 % при сопоставимом качестве [7]; позднее сообщалось, что для ряда задач классификации текстов ChatGPT превосходит краудсорсинговых исполнителей [8]. Систематический обзор применения LLM для разметки приведён в [9].

Практические системы оформляют это в цикл «разметка LLM → выборочная проверка человеком», где человек проверяет метки с низкой уверенностью; пример — инструмент MEGAnno+ [10]. В работе [11] LLM дополнительно генерирует объяснение метки, а отдельная модель-верификатор используется для приоритизации проверки. Вместе с тем установлено, что качество LLM-разметки сильно варьирует между задачами и требует обязательной валидации на размеченной экспертом выборке [17].

3.6. Статистические гибридные схемы

Наиболее близкое к построению эталона направление — методы, дающие **статистически обоснованные** выводы при комбинации дешёвых автоматических меток и небольшого числа экспертных. Prediction-powered inference использует малую экспертную выборку для коррекции смещения автоматических предсказаний и построения корректных

доверительных интервалов [13]. Метод confidence-driven inference развивает эту идею: уверенность LLM используется для выбора того, какие именно элементы отдать человеку, при сохранении корректности доверительных интервалов [12].

Эти методы, однако, нацелены на оценку агрегированных величин (доли класса, средние), а не на получение верной метки для каждого элемента корпуса, что требуется от эталона.

4. Сравнительный анализ

Ни один из рассмотренных классов не решает задачу из раздела 2 целиком: методы с высокой экономией не дают гарантии качества эталона, а методы с гарантией оценивают агрегаты, а не поэлементные метки. Таблица 1 сводит методы по единому основанию.

Таблица 1. Методы сокращения экспертного подтверждения

Класс методов	Механизм экономии	Роль эксперта	Типичные метрики в литературе	Гарантия качества эталона	Главное ограничение
Активное обучение [1, 2]	Отбор информативных примеров	Размечает отобранные элементы	Качество модели от числа меток (кривая обучения)	Нет	Смещённая выборка, не годится для оценки корпуса
Слабая разметка [3, 4]	Правила вместо меток	Пишет и отлаживает правила	Качество модели, покрытие правил	Нет	Шумные вероятностные метки
Предразметка + выборочная проверка [10, 14]	Проверка вместо создания метки, отбор по уверенности	Подтверждает или исправляет	Доля проверенных, точность после проверки	Частичная, зависит от калибровки	Переуверенность модели
Краудсорсинг + агрегация [5, 6]	Замена эксперта неспециалистами	Разрешает спорные случаи	Согласие с экспертом, каппа, альфа	Частичная	Коррелированные ошибки в трудных случаях
LLM-разметка [7–9, 11]	Автоматическая разметка всего корпуса	Проверяет низкоуверенные метки	Согласие с человеком, стоимость за метку	Нет без валидации	Нестабильность между задачами и языками
Статистические гибриды [12, 13]	Малая экспертная выборка корректирует автоматику	Размечает выборку	Ширина и покрытие доверительных интервалов	Да, для агрегатов	Не даёт верных поэлементных меток

Из таблицы следуют два наблюдения. Первое: метрики в литературе в основном относятся к качеству обученной модели или к стоимости, а остаточная ошибка самого эталона $\epsilon(S)$ почти не измеряется. Второе: доказуемые гарантии существуют только для агрегированных оценок, тогда как построение эталона требует контроля поэлементной ошибки.

5. Где теряется качество эталона

Анализ работ позволяет выделить пять источников потери качества при сокращении экспертного подтверждения.

1. **Уверенные ошибки.** Элементы, на которых автоматический разметчик ошибается с высокой уверенностью, не попадают на проверку при отборе по порогу. Именно они формируют остаточную ошибку эталона, и их доля растёт при плохой калибровке [18].

2. **Эффект якоря.** Эксперт, которому показана готовая метка, склонен с ней соглашаться. Установлено, что показ разметчикам меток LLM повышает их субъективную уверенность и смещает распределение меток в сторону предложенных, не ускоряя при этом работу [16].

3. **Коррелированные ошибки источников.** Агрегация нескольких разметчиков (людей, правил или моделей) повышает качество, только если их ошибки независимы. Модели, обученные на близких данных, и неспециалисты систематически ошибаются в одних и тех же трудных случаях, и согласие между ними создаёт ложную уверенность.

4. **Смещение проверяемой выборки.** Если эксперт проверяет только трудные или низкоуверенные элементы, по его исправлениям нельзя оценить ошибку в непроверенной, «лёгкой» части корпуса. Для такой оценки нужна дополнительная случайная контрольная выборка.

5. **Перенос между языками и доменами.** Качество LLM-разметки неравномерно между задачами [17] и, как правило, ниже на неанглоязычных данных. Пороги уверенности, подобранные на одном корпусе, не переносятся на другой без повторной калибровки.

Общий вывод: экономия экспертных проверок без механизма оценки остаточной ошибки превращает эталон в обычную обучающую выборку неизвестного качества.

6. Открытые проблемы и направления исследований

Выявленные пробелы определяют направления разработки алгоритмического и программного обеспечения.

– **Поэлементная гарантия качества эталона.** Требуются алгоритмы выбора множества V , обеспечивающие оценку $\epsilon(S)$ с заданной доверительной вероятностью не для агрегатов, а для корпуса в целом. Перспективно сочетание отбора по уверенности с небольшой случайной контрольной выборкой, по которой оценивается ошибка непроверенной части.

– **Согласие независимых разметчиков как сигнал качества.** Вместо самооценки уверенности одной модели можно использовать согласие нескольких разнородных моделей (разные архитектуры, разные LLM, правила). Открытый вопрос — как количественно учесть корреляцию их ошибок, например с помощью информационно-теоретических мер.

– **Снижение эффекта якоря в интерфейсе верификации.** Варианты: показ эксперту двух конкурирующих меток вместо одной, скрытие метки до вынесения собственного суждения на части элементов, журналирование времени и исправлений для оценки внимательности проверки.

– **Перенос экспертного решения на похожие элементы.** Если эксперт подтвердил представителя кластера семантически близких элементов, решение можно распространить на кластер с контролем риска. Нужны критерии однородности кластера, при которых такой перенос допустим.

– **Единая метрика эффективности.** Сопоставление методов требует метрики, связывающей долю сэкономленных проверок и остаточную ошибку эталона, аналогично кривым «покрытие — риск» в селективной классификации.

– **Малоресурсные языки.** Для узбекского языка характерны ограниченное число предобученных моделей, два алфавита (латиница и кириллица), богатая агглютинативная морфология и дефицит экспертов. Здесь калибровка и пороги, полученные на англоязычных данных, ненадёжны, а сокращение экспертной нагрузки наиболее востребовано. Отдельная

задача — построение небольших эталонных корпусов на узбекском языке для проверки самих методов.

7. Заключение

В статье систематизированы шесть классов методов сокращения объёма экспертного подтверждения при формировании эталонной разметки: активное обучение, слабая разметка, предразметка с выборочной проверкой, краудсорсинг с агрегацией, разметка с помощью LLM и статистические гибридные схемы. Задача сформулирована как минимизация числа экспертных проверок при ограничении на остаточную ошибку эталона.

Сравнительный анализ показал, что наибольшую экономию дают предразметка моделью и LLM-разметка с выборочной проверкой, но они не обеспечивают гарантии качества эталона, а статистически обоснованные методы гарантируют качество лишь агрегированных оценок. Выделены пять источников потери качества: уверенные ошибки, эффект якоря, коррелированные ошибки источников, смещение проверяемой выборки и перенос между языками.

В качестве направлений дальнейших исследований предложены алгоритмы отбора элементов для экспертной проверки с поэлементной гарантией качества, использование согласия независимых разметчиков, перенос экспертных решений на кластеры похожих элементов и адаптация методов к малоресурсным языкам, прежде всего узбекскому.

Список литературы

1. Settles B. Active Learning Literature Survey: Computer Sciences Technical Report 1648. Madison: University of Wisconsin–Madison, 2009.
2. Ren P., Xiao Y., Chang X. et al. A Survey of Deep Active Learning // ACM Computing Surveys. 2021. Vol. 54, No. 9. Art. 180. P. 1–40. DOI: 10.1145/3472291.
3. Ratner A., Bach S. H., Ehrenberg H., Fries J., Wu S., Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision // Proceedings of the VLDB Endowment. 2017. Vol. 11, No. 3. P. 269–282. DOI: 10.14778/3157794.3157797.
4. Zhang J., Hsieh C.-Y., Yu Y., Zhang C., Ratner A. A Survey on Programmatic Weak Supervision // arXiv preprint arXiv:2202.05433. 2022.
5. Dawid A. P., Skene A. M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm // Journal of the Royal Statistical Society. Series C (Applied Statistics). 1979. Vol. 28, No. 1. P. 20–28. DOI: 10.2307/2346806.
6. Snow R., O'Connor B., Jurafsky D., Ng A. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu, 2008. P. 254–263.
7. Wang S., Liu Y., Xu Y., Zhu C., Zeng M. Want To Reduce Labeling Cost? GPT-3 Can Help // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, 2021. P. 4195–4205. DOI: 10.18653/v1/2021.findings-emnlp.354.
8. Gilardi F., Alizadeh M., Kubli M. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks // Proceedings of the National Academy of Sciences. 2023. Vol. 120, No. 30. Art. e2305016120. DOI: 10.1073/pnas.2305016120.
9. Tan Z., Li D., Wang S. et al. Large Language Models for Data Annotation and Synthesis: A Survey // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2024. arXiv:2402.13446.
10. Kim H., Mitra K., Li Chen R., Rahman S., Zhang D. MEGAnno+: A Human-LLM Collaborative Annotation System // Proceedings of the 18th Conference of the European Chapter of the

ACL: System Demonstrations (EACL). St. Julians, 2024. P. 168–176. DOI: 10.18653/v1/2024.eacl-demo.18.

11. Wang X., Kim H. et al. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels // Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, 2024. DOI: 10.1145/3613904.3641960.

12. Gligorić K., Zrnic T., Lee C., Candès E., Jurafsky D. Can Unconfident LLM Annotations Be Used for Confident Conclusions? // Proceedings of NAACL-HLT 2025 (Vol. 1: Long Papers). Albuquerque, 2025. P. 3514–3533. DOI: 10.18653/v1/2025.naacl-long.179.

13. Angelopoulos A. N., Bates S., Fannjiang C., Jordan M. I., Zrnic T. Prediction-Powered Inference // Science. 2023. Vol. 382, No. 6671. P. 669–674.

14. Northcutt C. G., Jiang L., Chuang I. L. Confident Learning: Estimating Uncertainty in Dataset Labels // Journal of Artificial Intelligence Research. 2021. Vol. 70. P. 1373–1411. DOI: 10.1613/jair.1.12125.

15. Artstein R., Poesio M. Inter-Coder Agreement for Computational Linguistics // Computational Linguistics. 2008. Vol. 34, No. 4. P. 555–596. DOI: 10.1162/coli.07-034-R2.

16. Schroeder H., Roy D., Kabbara J. Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks // Findings of the Association for Computational Linguistics: ACL 2025. Vienna, 2025. P. 25771–25795. DOI: 10.18653/v1/2025.findings-acl.1323.

17. Pangakis N., Wolken S., Fasching N. Automated Annotation with Generative AI Requires Validation // arXiv preprint arXiv:2306.00176. 2023.

18. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR, 2017. Vol. 70. P. 1321–1330.