Оценка стоимости применения для выбора модели структурированной разметки текста

О. Ю. Седельникова*, Н. А. Кукузей‡, И. А. Цыбанов§, О. Л. Ринк¶, А. И. Майсурадзе‖, Р. В. Ищенко**

Выбор модели для практического внедрения определяется балансом между функциональной точностью и стоимостью применения. Предлагается оценка стоимости применения, адаптированная для моделей сложноструктурированной разметки текстов. Интегральная оценка учитывает полный процесс автоматической разметки и возвращает стоимость разметки 1000 текстов. На 21 наборе данных сопоставлены модели разной архитектуры и размера.

Ключевые слова: структурированная разметка текста; выделение и связывание фрагментов; вложенные фрагменты; оценка стоимости применения; выбор модели; операционные затраты.

1. Постановка задачи

Универсальная модель разметки текста требует сложноструктурированной модели данных, позволяющей обрабатывать разные типы задач в едином представлении [1]: фрагменты с тегами объединяются в элементы, описывающие сущности, отношения и кореферентные группы. Цель

* *Седельникова Ольга Юрьевна* — Лаборант, Институт ИИ МГУ, e-mail: sedelnikovaoy@my.msu.ru, ORCID: 0009-0007-4923-3701.

Sedelnikova Olga Iuryevna — Laboratory Assistant, IAI MSU.

‡ *Кукузей Никита Александрович* — Лаборант, Институт ИИ МГУ, e-mail: kukuzeyna@my.msu.ru, ORCID: 0009-0004-0420-9685.

Kukuzei Nikita Alexandrovich — Laboratory Assistant, IAI MSU.

§ *Цыбанов Илья Антонович* — Стажер-исследователь, Институт ИИ МГУ, e-mail: tsybanovia@my.msu.ru, ORCID: 0009-0000-1845-464X.

Tsybanov Ilya Antonovich — Research Intern, IAI MSU.

¶ *Ринк Ольга Леонидовна* — Научный сотрудник, Институт ИИ МГУ, e-mail: rinkol@my.msu.ru, ORCID: 0000-0001-5977-6741.

Rink Olga Leonidovna — Researcher, IAI MSU.

‖ *Майсурадзе Арчил Ивериевич* — к. ф.-м. н., в. н. с. Института ИИ МГУ, e-mail: maysuradze@cs.msu.ru, ORCID: 0000-0002-1757-556X.

Maysuradze Archil Iverievich — Cand. of Phys. and Math. Sc., Principal Researcher, IAI MSU.

** *Ищенко Роман Валерьевич* — Научный сотрудник, Институт ИИ МГУ, e-mail: rois@cs.msu.ru, ORCID: 0009-0008-9720-8209.

Ischenko Roman Valeryevich — Researcher, IAI MSU.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

работы — предложить оценку стоимости применения (операционных затрат) для моделей сложноструктурированной разметки. Универсальность относится к общей схеме обработки, допускающей настройку под разные корпуса и типы задач. Оценивается структурная разметка, включая вложенные фрагменты; генерация дополнительных текстов не рассматривается.

2. Интегральная оценка стоимости применения

В работах [2, 3] рассматриваются вычислительная эффективность и экономические аспекты применения моделей. Здесь предлагается оценка стоимости полного конвейера структурной разметки на 1000 исходных текстов. Интегральность означает учёт всех этапов конвейера после загрузки и прогрева модели. При постоянной конфигурации и линейной тарификации

$$C_{1000} = \frac{1000pT}{3600N} = \frac{1000p}{3600q}, \qquad q = \frac{N}{T}. \tag{1}$$

Здесь N — число исходных текстов до разбиения на окна; T — фактическое время выполнения конвейера, с; p — тариф конфигурации, USD/ч; q — пропускная способность, текстов/с. Результат измеряется в USD за 1000 текстов. Окна и повторы увеличивают T, но не N; при параллельной обработке учитывается фактическое время работы конфигурации, а не сумма длительностей запросов.

Таймер запускается после загрузки и прогрева. Учитываются предобработка, применение модели, разбор ответов, объединение окон и предусмотренные протоколом исправления; при прямом замере GPU синхронизируются. Тариф включает GPU, CPU и RAM без двойного учёта; фиксируются провайдер, дата, регион и тип аренды. Пиковые RAM процессов и VRAM каждого GPU определяют допустимость конфигурации.

Для воспроизводимости фиксируются оборудование, версии ПО, размер пакета, параллелизм, точность вычислений, длины входов; для LLM — также промпт, параметры генерации и предел повторов. Учитываются ошибки, режим мониторинга памяти, число прогонов, медиана и диапазон стоимости. По нескольким корпусам суммируются затраты и объёмы текстов. Обучение, подготовка инструкции, запуск, простой и экспертная проверка не входят в стоимость обработки.

3. Схема эксперимента

Бенчмарк включает 21 набор данных. Для каждой модели вычисляются оценки стоимости применения и показатели функциональной точности. Интегральное качество $Q = \sum_{i=1}^{5} \alpha_i C_i$, $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, объединяет полноту сопоставления фрагментов и элементов (C_1, C_4), совпадение границ (C_2) и тегов (C_3, C_5). Веса α_i зависят от типа задачи; оценки усредняются по экспертным вариантам и тестовым данным. Разбиение 8:1:1 (обучение/валидация/тест) зафиксировано; качество оценивается на тесте, few-shot примеры берутся из обучения.

Сравниваются `mamba-1.4b-ru` (модель пространства состояний), `spanbert-base-cased` (кодировщик с предобучением на фрагментах), `mDeBERTa-v3-base` (многоязычный трансформерный кодировщик) и генеративная `Qwen3.5-397B-A17B`. Для первых трёх моделей двухуровневые головы выделяют фрагменты по парам токенов, затем формируют элементы и теги по парам фрагментов. Qwen выдаёт JSON, из которого программа восстанавливает смещения и проверяет ссылки между объектами.

Для mDeBERTa, SpanBERT и Mamba предобученная основа заморожена; для каждого набора данных обучаются только классификационные головы. Представления фрагментов получают максимальным пулингом токенных представлений; расстояние между парными токенами ограничено 64, пороги тегов подбираются на валидации. Qwen3.5-397B-A17B используется в FP8 без градиентной адаптации; промпт включает примеры определений классов из таксономии и примеры разметки из обучающей выборки. Из 397 млрд параметров языковой части 17 млрд активируются на токен, однако для размещения необходима память под все веса и состояние выполнения. Всего сравниваются четыре конфигурации.

4. Результаты и выводы

Получены пары $(Q; C_{1000})$ (стоимость в USD за 1000 текстов): mDeBERTa $(0{,}44; 0{,}113)$, SpanBERT $(0{,}40; 0{,}146)$, Mamba $(0{,}47; 0{,}520)$, Qwen $(0{,}69; 2{,}172)$. По неокруглённым значениям Qwen дороже mDeBERTa, SpanBERT и Mamba соответственно в 19,2, 14,9 и 4,2 раза при приросте Q примерно на 0,25, 0,29 и 0,22. Методика позволяет выбрать конфигурацию с минимальным C_{1000} среди поддерживающих нужные типы разметки и удовлетворяющих порогу качества. При одинаковом охвате задач и порогах $Q_{\min} = 0{,}40$, 0,45 и 0,50 выбираются соответственно mDeBERTa, Mamba и Qwen. Примеры основаны на сводных

оценках бенчмарка; для отдельных задач нужны собственные оценки, а для других условий — повторные измерения.

Список литературы

[1] A. I. Maysuradze et al., “Generalized Text Markup Structures to Organize Knowledge in AI Applications for Humanities and Social Sciences”, *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2861–2874. DOI: 10.1134/S1995080226619302.

[2] D. Narayanan et al., “Cheaply estimating inference efficiency metrics for autoregressive transformer models”, *NeurIPS*, **36**, 2023, 66518–66538. DOI: 10.52202/075280-2904.

[3] E. Erdil, “Inference economics of language models”, arXiv: 2506.04645.

Статья поступила 23 октября 2026 г.

Inference Cost Estimation for Selecting a Structural Text Annotation Model

O. I. Sedelnikova, N. A. Kukuzei, I. A. Tsybanov, O. L. Rink, A. I. Maysuradze, R. V. Ischenko

Model selection for practical deployment is governed by the trade-off between functional accuracy and cost of use. A cost-of-use estimation procedure is tailored to models for structurally complex text annotation. The aggregate estimate accounts for the full automatic annotation pipeline and returns the cost of annotating 1,000 texts. Models of differing architecture and size were compared across 21 datasets.

Keywords: structured text annotation; span detection and linking; nested spans; inference costs; model selection; operational costs.

References

[1] A. I. Maysuradze et al., “Generalized Text Markup Structures to Organize Knowledge in AI Applications for Humanities and Social Sciences”, *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2861–2874. DOI: 10.1134/S1995080226619302.

[2] D. Narayanan et al., “Cheaply estimating inference efficiency metrics for autoregressive transformer models”, *NeurIPS*, **36**, 2023, 66518–66538. DOI: 10.52202/075280-2904.

[3] E. Erdil, “Inference economics of language models”, arXiv: 2506.04645.

Received on October 23, 2026