УДК 004.852

АДАПТИВНОЕ СМЕШИВАНИЕ ДАННЫХ С КОЭФФИЦИЕНТОМ ЗОЛОТОГО СЕЧЕНИЯ И МОНИТОРИНГОМ SURPLEXITY ДЛЯ ПРЕДОТВРАЩЕНИЯ КОЛЛАПСА ГЕНЕРАТИВНЫХ МОДЕЛЕЙ

Боброва Е.В. (НИЯУ МИФИ), **Петровская Я.В.** (НИЯУ МИФИ), **Зайцев К.С.** (НИЯУ МИФИ)

Научный руководитель – доктор технических наук, профессор Зайцев К.С. (НИЯУ МИФИ)

Введение. Коллапс моделей представляет критическую проблему при обучении больших языковых моделей на смеси реальных и синтетических данных [4]. При рекурсивном обучении на собственных генерациях модель постепенно теряет разнообразие выходов, что приводит к деградации качества и повторяющимся предсказаниям. Современные подходы к предотвращению коллапса основаны либо на фиксированных пропорциях смешивания данных, либо на статических метриках качества. Основная проблема заключается в отсутствии динамической адаптации соотношения реальных и синтетических данных в процессе обучения на основе ранних индикаторов коллапса. Метрика surplexity, определяющая «неожиданность» документа для модели как экспоненту средней отрицательной логарифмической вероятности токенов, показала эффективность в выявлении признаков коллапса [2]. Исследования демонстрируют, что оптимальный вес реальных данных при фиксированных объёмах выборок асимптотически стремится к обратной величине золотого сечения (≈0.618) [1].

Основная часть. Работа направлена на разработку адаптивной системы обучения LLM, динамически корректирующей пропорцию смешивания real/synthetic данных на основе мониторинга surplexity и риска коллапса. Система включает три ключевых компонента: (1) LLMSurplexityMonitor — вычисляет surplexity как комбинацию perplexity, diversity (distinct-1/2 метрики) и surprise (KL-дивергенция распределений токенов относительно референсного набора) [3]; (2) детектор риска коллапса на основе анализа тренда, волатильности и абсолютных значений surplexity в скользящем окне; (3) AdaptiveController, корректирующий mixing ratio от начального значения GOLDEN_RATIO=1/φ≈0.618 в диапазоне [0.3, 0.9] с шагом α=0.1 при превышении порога риска. При collapse_score>0.5 доля реальных данных увеличивается, при collapse_score<0.2 и отрицательном тренде surplexity — уменьшается. Метод опирается на теоретический результат He et al. о золотом сечении как оптимальном весе и на surplexity-метрику Gambetta et al. для идентификации документов, минимизирующих риск коллапса [1,2]. Система интегрирована в pipeline обучения с периодической регенерацией синтетического датасета и пересчётом метрик после каждой итерации.

Выводы. Предложенный метод адаптивного управления mixing ratio позволяет динамически балансировать использование синтетических данных, предотвращая коллапс модели при сохранении преимуществ data augmentation. Ожидается снижение collapse_score с начальных значений >0.5 до уровня <0.2 при стабилизации surplexity на уровне ≤5.0 после 10–15 итераций обучения. Применение порога для веса синтетических данных согласуется с теоретическими результатами о существовании оптимального соотношения, обеспечивающего повышение эффективности оценки параметров. Целевые показатели включают поддержание distinct-1 и distinct-2 метрик на уровне >0.4 и >0.3 соответственно при росте mixing ratio в безопасном диапазоне. Достижимость целей подтверждается экспериментами на wikitext-2 и openwebtext датасетах с моделями семейства GPT-2. Дальнейшее развитие включает интеграцию методов обучения с подкреплением для

оптимизации стратегии выбора документов на основе surplexity.

Список использованных источников:

1. He H., Xu S., Cheng G. Golden Ratio Weighting Prevents Model Collapse // arXiv. – 2025. – arXiv:2502.18049v4.
2. Gambetta D., Gezici G., Giannotti F., Pedreschi D., Knott A., Pappalardo L. Surplexity for Mitigating Model Collapse in Generative AI // arXiv. – 2024. – arXiv:2410.12341v3.
3. Garg S., Tsai Y.-H., Salakhutdinov R., Morency L.-P. Learning to Optimize Data Usage for Fine-Tuning Large Language Models // Proceedings of EMNLP. – 2024. – P. 14130–14151
4. Shumailov I., Shumaylov Z., Zhao Y., Papernot N., Anderson R., Gal Y. AI models collapse when trained on recursively generated data // Nature. – 2024. – Vol. 631. – P. 755–759.