Построение неавторегрессионного Seq2Seq-пайплайна на основе скрытых возможностей LLM для одношаговой генерации

Общая гипотеза: подход на основе "прото-токенов" [1] позволяет построить неавторегрессионную модель генерации, которая не уступает по качеству авторегрессионным LLM, обладает значительно более высокой скоростью вывода, способностью оценивать собственную неопределенность и устойчивостью к распределительным сдвигам.

План экспериментов

Фаза 1: Исследование природы сжатых представлений

Цель: Определить, какую лингвистическую информацию (синтаксис, семантика, лексика) кодируют прото-токены *e* и *m*.

Эксперименты:

- ***синтаксис***: реконструкция текстов на синтетическом псевдоязыке [2];
- ***семантика***: реконструкция списков фактов, определений, где синтаксис минимален;
- ***как оценивать***: точность пословной реконструкции, анализ ошибок (служебные vs. смысловые слова), перплексия.

Фаза 2: Адаптация метода для построения инструкционной LLM

Цель: Настроить метод для эффективного сжатия и восстановления ответов в инструкционном датасете.

Эксперименты:

- ***базовый подход***: обучение уникальных *e* и *m* для ответа (response) в каждом примере «эталонно размеченного» инструкционного датасета

типа **`databricks/databricks-dolly-15k`** на основе некоторой предобученной LLM, например, семейства Pythia [3];

- ***абляционные исследования***:
  - сравнение схем m_{shared} (один на все), $m_{\text{per\_category}}$ (по категориям датасета Dolly) и $m_{\text{per\_example}}$.
  - проверка гипотезы о "механистической" роли m (замена на фиксированный вектор).
- ***масштабирование***: исследование зависимости качества реконструкции от размера модели-декодера (на семействе Pythia).

Фаза 3: Создание и обучение энкодера для оценки *e*, *m* и *length*

Цель: Обучить модель, которая по инструкции (входному текстуу) предсказывает оптимальные параметры e и m для генерации ответа (response) и длину этого ответа (в токенах) *length*.

Эксперименты:

- ***архитектура***: выбор энкодера (например, RoBERTa), проекционных голов для e, m и регрессионной головы для *length*.
- ***функции потерь***:
  - для e и m: сравнение контрастивных функций потерь distance-based logistic loss [4], SigLIP [5] и других;
  - для *length*: функция потерь на основе минимизации отрицательного логарифмического правдоподобия для гауссовского распределения ошибок оценивания целевой переменной [6].
- ***валидация***: Оценка конечного пайплайна (энкодер → e, m, *length* → декодер → ответ) на качестве реконструкции Dolly-15k.

Фаза 4: Всестороннее сравнение с авторегрессионными LLM

Цель: Доказать, что неавторегрессионный seq2seq-пайплайн конкурентоспособен.

Эксперименты:

- ***Качество***: Сравнение неавторегрессионного seq2seq-пайплайна с сопоставимыми по размеру авторегрессионными моделями (декодерными типа GPT и seq2seq типа FLAN-T5) на бенчмарках понимания языка (MMLU, MMLU-Pro, GSM8K и др.).
- ***Устойчивость***: Построение F1-retention curve [7] для оценки устойчивости к сдвигам распределения.
- ***Производительность***: замеры latency (время отправкой входного промпта и получением сгенерированного ответа) и throughput (количество генерируемых токенов в секунду) на различных длинах ответов в сравнении с авторегрессионным бейзлайном, а также бейзлайном на основе stable diffusion типа LLaDA [8].
- ***Полезность оценки неопределенности***: Проверка, что высокая предсказанная неопределенность коррелирует с фактически низким качеством ответа, позволяя модели "воздерживаться" от ответа.

Список литературы

1. *Gleb Mezentsev*, *Ivan Oseledets*. **Exploring the Latent Capacity of LLMs for One-Step Text Generation** // arXiv preprint arXiv:2505.21189. – 2025.

2. *Taisiya Gorbacheva*, *Ivan Bondarenko*. **Safe Pretraining of Deep Language Models in a Synthetic Pseudo-Language** // Dokl. Math. 108 (Suppl 2). – 2023.

3. *Stella Biderman*, *Hailey Schoelkopf* et al. **Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling** // Proceedings of the 40th International Conference on Machine Learning (ICML 2023). – 2023.

4. *Nam N. Vo*, *James Hays*. **Localizing and Orienting Street Views Using Overhead Imagery** // Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). – 2016.

5. *Xiaohua Zhai*, *Lucas Beyer* et al. **Sigmoid Loss for Language Image Pre-Training** // Proceedings of the International Conference on Computer Vision (ICCV 2023). – 2023.

6. *D. A. Nix*, *A. S. Weigend*. **Estimating the mean and variance of the target probability distribution** // In IEEE International Conference on Neural Networks. – 1994.

7. *Andrey Malinin* et al. **Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks** // Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). – 2021.

8. *Fengqi Zhu*, *Rongzhen Wang* et al. **LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models** // arXiv preprint arXiv:2505.19223. – 2025.