

# Анализ неопределенности и внутренних представлений языковых моделей для поиска сгенерированных документов

Анастасия Евгеньевна Вознюк  
Научный руководитель: к.ф.-м.н. А. В. Грабовой

Кафедра интеллектуальных систем ФПМИ МФТИ  
Специализация: Интеллектуальный анализ данных  
Направление: 01.03.02 Прикладные математика и информатика

21 декабря 2024

# Цель работы

Исследуется проблема определения природы текста.

## **Цель работы**

Построение методов поиска сгенерированных фрагментов с помощью подсчет неопределенности в тексте, а также с помощью анализа внутренних представлений языковых моделей.

## Общая постановка задачи

Определим документ как конечную последовательность символов из заданного алфавита  $\mathbf{W}$ . Пространство документов:

$$\mathbb{D} = \left\{ \left[ t_j \right]_{j=1}^n \mid t_j \in \mathbf{W}, n \in \mathbb{N} \right\}.$$

Дан набор из  $N$  документов

$$\mathbf{D} = \bigcup_{i=1}^N D^i, D^i \in \mathbb{D}.$$

Определим множество авторов, тексты которых встречаются в наборе  $\mathbf{D}$ :

$$\mathbf{C} = \{0, \dots, k-1\}.$$

Тогда задача классификации автора документа записывается как:

$$\phi : \mathbb{D} \rightarrow \mathbf{C} \tag{1}$$

## Анализ неопределенности

$\theta$  - параметры нашей модели  $f$ .

$$y_k = f(D, y_1, \dots, y_{k-1} \mid \theta)$$

$$\mathbf{y} = [y_1, \dots, y_n]^T$$

### Maximum Sequence Probability

$$\text{MSP}(\mathbf{y} \mid D, \theta) = 1 - P(\mathbf{y} \mid D, \theta) \quad (2)$$

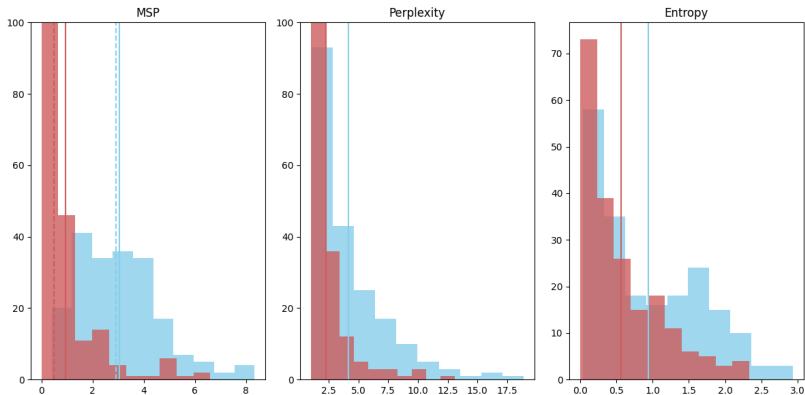
### Perplexity

$$P(\mathbf{y} \mid D, \theta) = \exp\left\{-\frac{1}{|D|} \log P(\mathbf{y} \mid D, \theta)\right\} \quad (3)$$

### Mean Token Entropy

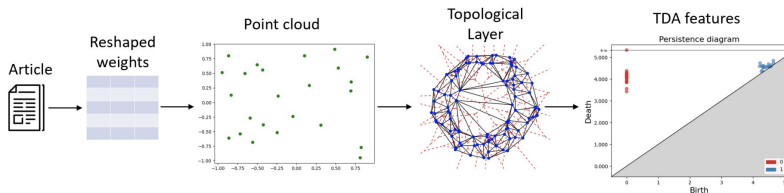
$$\mathcal{H}_T(\mathbf{y}, D; \theta) = \frac{1}{|D|} \sum_{l=1}^{|D|} \mathcal{H}(y_l \mid \mathbf{y}_{<l}, D, \theta) \quad (4)$$

# Вычислительный эксперимент



Было взято 1000 человеческих текстов и 1000 текстов и на них были посчитаны функции, описанные на предыдущем слайде.

# Использование топологических признаков текста



Основной подход основан на использовании понятий персисентной гомологии конечного набора точек  $\mathcal{M}$  в метрическом пространстве с метрикой  $d$ .

$$E_t = \{(v, u) : v, u \in \mathcal{M}, \quad d(v, u) \leq t\} \quad \forall t \in (0, \infty)$$

Персисентная гомология  $\text{PH}_i$  определена набором признаков размерности  $i$ , так  $\text{PH}_0$  задается компонентами связности,  $\text{PH}_1$  - циклами, и т.д.

## Внутренняя размерность

У каждого признака есть своя *продолжительность жизни*, это пара  $(t_{\text{birth}}, t_{\text{death}})$ , когда данный признак появляется, и когда исчезает.

Введем  $\alpha$ -взвешенную сумму,  $I(\gamma)$  - продолжительность жизни признака  $\gamma$ .

$$E_{\alpha}^i(X) := \sum_{\gamma \in \text{PH}_i(X)} |I(\gamma)|^{\alpha} \quad (5)$$

$$E_{\alpha}^0(X) \sim Cn^{\frac{d-\alpha}{d}}, n \rightarrow \infty \Leftrightarrow \alpha < d \quad (6)$$

$$\dim_{\text{PH}}(\mathcal{M}) = \inf \left\{ d \mid \exists C \quad E_d^0(X) \leq C \quad \forall X \subset \mathcal{M}, |X| < \infty \right\}. \quad (7)$$

### Гипотеза

Внутренняя размерность  $\dim_{\text{PH}}(\mathcal{M})$  показывает количество степеней свободы у точки в  $\mathcal{M}$ .

**Персисентная гомологическая размерность (PHD)** - метрика, основанная на внутренней размерности текстов, показала себя статистически значимой метрикой для разделения текстов разной природы для первых языковых моделей<sup>1</sup>. Однако для более новых моделей разделимость уже не такая хорошая.

#### Гипотеза

Можно адаптировать PHD для новых моделей, так что она все еще будет статистической метрикой разделимости текстов.

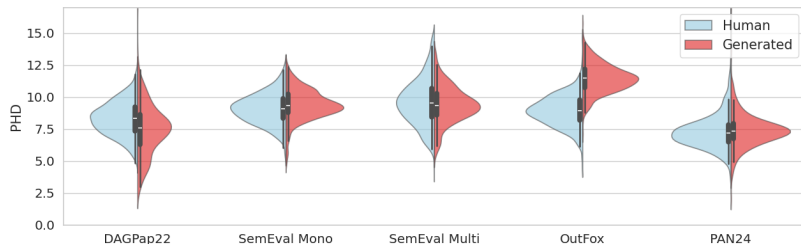
---

<sup>1</sup>Tulchinskii et al. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts, NeurIPS 2023



# Вычислительный эксперимент

Dataset	PHD <sub>human</sub>	PHD <sub>machine</sub>
OutFox	$8.96 \pm 1.21$	$11.48 \pm 1.13$
SemEval24 Mono	$9.11 \pm 1.19$	$9.41 \pm 1.2$
SemEval24 Multi	$9.65 \pm 1.81$	$9.42 \pm 1.44$
DAGPap22	$8.35 \pm 1.33$	$7.48 \pm 2.01$
PAN24	$9.4 \pm 1.05$	$8.52 \pm 1.59$
MGT-1 Mono	$9.19 \pm 1.75$	$8.96 \pm 2.24$
MGT-1 Multi	$8.76 \pm 1.85$	$8.6 \pm 2.29$



## Новый способ подсчитывать внимание

Пусть  $N$  - длина текстовой последовательности. Выделим в тексте "якоря"  $d_1, \dots, d_n$ , в которых содержится основной смысл текста. Рассмотрим голову внимания  $h$  со слоя  $l$  модели  $M$ . Определим  $QK$ -score  $S_{QK}^{(l,h)}(d_i)$  и  $Attention$ -score  $S_{Att}^{(l,h)}(d_i)$ :

$$S_{QK}^{(l,h)}(d_i) = q_N^{(l,h)\top} k_{t_i}^{(l,h)}, \quad S_{Att}^{(l,h)}(d_i) = a_{N,t_i}^{(l,h)}, \quad i \in \{1, 2, \dots, n\} \quad (8)$$

### Гипотеза

$QK$ -score на некоторых задачах может лучше определять паттерны и решать задачу чем стандартный подсчет внимания.

## Вычислительный эксперимент

Данные два подхода подсчета внимания сравнивались на задаче ответов на тестовые вопросы MMLU.

Method	LLaMA...			
	2-13B	2-70B	3-8B	3-70B
Baseline, Acc	47.4	57.7	60.5	<b>78.2</b>
Baseline, PA	34.6	45.9	47.7	<b>70.1</b>
QK-score, Acc	<b>49.7</b>	<b>58.9</b>	<b>63.0</b>	77.9
QK-score, PA	<b>38.3</b>	<b>47.1</b>	<b>49.3</b>	67.9

**Таблица:** Comparison of different base models in zero-shot setup on various Q&A datasets. Reported metrics are Accuracy (Acc) and Permutation Accuracy (PA). Best results are highlighted in **bold**.

# Итоги НИР за семестр и планы на следующий семестр

## Результаты

1. Опубликована 1 работа на A\* конференции, еще 1 принята к публикации, еще 2 в состоянии препринта и ожидают оценки ревьюеров;
2. Получены первые результаты с анализом PHD для сгенерированных текстов и с оценкой неопределенности на них;

## Планы

1. Необходимо модифицировать PHD чтобы добиться разделимости по текстам от более новых моделей;
2. Применить QK-Score для новых задач;

# Список работ автора по теме НИР

## Публикации

1. Listening to the Wise Few: Select-and-Copy Attention Heads for Multiple-Choice QA. *arXiv preprint:2410.02343*
2. Advachek at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking, *Proceedings of Workshop on Detecting AI Generated Content at COLING 2025*
3. Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts, *arXiv preprint:2410.14677*

## Выступления с докладом

1. DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning.// *EMNLP, Miami, Florida*