January 15, 2025

Comparing Large Language Models for supervised analysis of students' lab notes

Введение

Тема исследования: Сравнение моделей машинного обучения для анализа лабораторных заметок студентов.
Цель: Исследование эффективности моделей в классификации предложений по экспериментальным навыкам.

- Rebeckah K. Fussell, Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA.
- Megan Flynn, Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA; Department of Computer Science, Cornell University, Ithaca, New York 14853, USA.
- Anil Damle, Department of Computer Science, Cornell University, Ithaca, New York 14853, USA.
- Michael F.J. Fox, Department of Physics, Imperial College London, London, UK.
- N.G. Holmes, Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA.

Методы исследовани

Данные: Лабораторные заметки студентов физического факультета (2019 и 2022 годы).
873 заметки, 58,369 предложений.

Кодировка навыков:

- **QC Code (Количественное сравнение):**
  - **Описание:** Предложения, где студенты сравнивают данные, линии на графиках, предсказания и т.д.
  - **Требования для включения:** Должна быть явная ссылка на использование инструментов анализа данных и сравнение двух величин.
- **PI Code (Предложение итерации):**
  - **Описание:** Предложения, где студенты предлагают улучшения или дальнейшие эксперименты.
  - **Требования для включения:** Наличие слов, связанных с планами, улучшениями, или повторными измерениями.

Методы исследовани

Модели:

- Bag of Words (BoW)
- BERT (110M параметров)
- LLaMA (8B параметров, с и без подсказок)
- Zero-shot LLaMA

Метрики: AUC, Accuracy, Balanced Accuracy.

Сравнение производительности

QC Code:

- LLaMA (с и без подсказок) показала наилучшие результаты.
- BERT уступает LLaMA, но превосходит BoW.
- Zero-shot LLaMA значительно уступает остальным.

PI Code:

- Все модели показывают схожие тренды, но LLaMA снова лидирует.
- Bag of Words оптимизирует Balanced Accuracy за счет чувствительности.

Итог: LLaMA ближе к результатам межэкспертной надежности.

Ресурсы и затраты

Временные затраты на обучение:

- BoW: 2 секунды (CPU).
- BERT: 7 минут 16 секунд (GPU A6000).
- LLaMA: 6+ часов (GPU A6000).

Человеческое участие:

- Простота: BoW > BERT > LLaMA.
- LLaMA требует больше опыта с кодом и настройкой.

Рекомендации: Выбор модели зависит от доступных ресурсов и цели исследования.

Исследовательские результаты

QC и PI тренды:

- Все модели определяют схожие тренды изменений навыков в лабораторных заметках.
- Различия в абсолютных значениях требуют учета систематических и статистических неопределенностей.

Тенденции:

- Максимальная частота навыков наблюдается в сеансах с явной инструкцией (например, L1b).
- LLaMA показывает более точное соответствие статистическим трендам.

Выводы

Рекомендации для PER:

- LLM (например, LLaMA) улучшает качество классификации по сравнению с традиционными методами.
- Даже ресурсоемкие модели оправданы при необходимости высокой точности.

Будущие исследования:

- Оценка влияния гиперпараметров и оптимизация подсказок.
- Адаптация более крупных LLM (e.g., GPT-4).