CS324 - Large Language Models

Data, Legality

HUMANE Lab

김태균

2025.01.10

Overview

- **Data**
  - Examples of several datasets
  - The importance of documentation
  - Research on data from various perspectives

- **Legality**
  - Legal considerations for LLM development and deployment

Data – (1) Behind LLM

- LLM are trained on "**raw text**" that should span a broad range of domains, genres, languages, etc.
- A natural place to look for such text is the **web**

Data – (1) Behind LLM

- **Common Crawl**
  - Crawls the web data and provides snapshots that are free to the public
  - To train such as T5, GPT-3 and Gopher

- **WebText**
  - Large-scale text dataset collected by OpenAI for training GPT-2
  - Only use data that has received more than three upvotes on Reddit
  - **OpenWebText**
  - RealToxicityPrompts indicates the presence of a certain level of toxicity

Data – (1) Behind LLM

- **Colossal Clean Crawled Corpus (C4)**
  - Large-scale text dataset collected by Google Research for training T5
  - Composed of text collected from Common Crawl data

- **The Pile**
  - A high-quality dataset(academic + professional sources) built by EleutherAI
  - Contains a lot of information that’s not well covered by GPT-3’s dataset

Data – (1) Behind LLM

- Despite the richness of web data, large-scale data still has **bias**
  - GPT-2's training data is based on Reddit, which 67% Reddit users in the US are men
  - Filtering "bad words" could further marginalize certain populations (e.g. LGBT+)

- Benchmark data **contamination**
  - If benchmark data appears in the training data, benchmark performance will become biased
  - In the case of LLMs, since both data sets are derived from the Internet, separating them is difficult

Data – (1) Behind LLM

- **Representational harms**
  - May include biased expressions toward specific races, ethnicities, genders, or other groups
- **Allocational harms**
  - Data related to specific groups (e.g. LGBT+) may be filtered out

Data – (1) Behind LLM

Therefore, the massive amount of data collected from the web should be used through proper filtering and curation processes

Data – (2) Documentation

- Examples from other fields
  - Electronics industry has a datasheet with operating characteristics, test results, and usage
  - The FDA mandates that food be labeled with their nutrition content

Data – (2) Documentation

Two purposes:

1. **Dataset creators** : Reflect on decisions, potential harms
2. **Dataset consumers** : Know when the dataset can and can’t be used

Data – (2) Documentation

- **Dataset lifecycle**
  - Motivation
  - Composition
  - Collection process
  - Preprocessing/cleaning/labeling
  - Uses
  - Distribution
  - Maintenance

Data – (2) Documentation

- **Data statements** (specialized to NLP datasets)
  - Mitigate bias-related issues in language technology
  - Improve the precision of claims regarding NLP research

Data – (3) Ecosystems

- **Data management**
  - In ML research, we tend to think of datasets as fixed objects
  - In DB community, there is whole subfield thinking about the ecosystem in which data comes to be and is used

- **Data dignity**
  - Individually, data doesn't have value, but collectively, it has a lot of value
  - Think about data as labor rather than property rights

Legality

- Whenever a new powerful technology emerges, it raises many questions about whether existing laws still apply or make sense
  - e.g. Internet ⇒ intellectual property law, privacy law

- Internet clearly has its own unique challenges
  - Laws usually had clear jurisdiction, but the Internet is not geographically bound
  - Anonymous on the Internet
  - Anyone can post a piece of content that in principle can get be viewed by anyone

Legality

- The difference between laws and ethics
  - **Laws is enforceable** by government
  - **Ethics is not enforceable** and can be created by any organization

Legality

- Types of law
  - **Common law (Judiciary)** : Based on judges referencing previous similar cases and making a ruling
  - **Statutory law (Legislature)** : Produced by government agencies through the legislative process
  - **Regulatory law (Executive)** : Created by the executive branch of government, often focusing on procedures

Legality

Two main areas where the law intersects the LLMs lifecycle

1. Data

- All ML relies on data
- LLM rely on a lot of data, especially other people's data, and often scraped without consent

2. Applications

- LLMs can be used for a wide range of downstream tasks
- Technologies can be used intentionally for harm (e.g. spam)
- They could be deployed in various high-stakes settings (e.g. education)

Legality – Copyright Law

- Intellectual property law
  - Motivation : Encourage the creation of a wide variety of intellectual goods
  - Types : Copyright, patents, trademarks, trade secrets

- **Copyright law**
  - *"Original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device"*
  - Registration is not required for copyright protection (in contrast with patents)
  - Last for 75 years, and then the copyright expires and it becomes part of the public domain

Legality – Copyright Law

Two ways of using a copyrighted work

1. Get a **license**
2. Appeal to the **fair use** clause

Legality – Copyright Law

- **Licenses**
  - A license from contract law is granted by a licensor to a licensee
  - “Promise not to sue”

Legality – Copyright Law

- **Fair use**
  - A legal doctrine that allows limited use of copyrighted material without requiring permission
  - Four factors to determine whether fair use applies
    - the purpose and character of the use
    - the nature of the copyrighted work
    - the amount and substantiality of the portion of the original work used
    - the effect of the use upon the market for the original work

  ex) Watch a movie, write a summary of it

Legality – Copyright Law

- Terms of service
  - Rules that must be agreed to in order to use the service

Legality – Copyright Law

- Arguments for LLM as fair use
  - Broad access to training data makes better systems for society
  - Produce new value
- Arguments against LLM as fair use
  - Argue that LLM don’t produce a creative “end product” but just make money
  - Problems with LLM (e.g. spread disinformation)

⇒ The future of copyright and LLM is very much open

Legality – Copyright Law

When training LLMs, we may face issues related to copyright and fair use, so careful usage is required