Секция «1.1 Цифровая трансформация и искусственный интеллект в государственном администрировании 3.0: от автоматизации к принятию интеллектуальных решений»

Галлюцинации в генеративных нейросетевых моделях: причины, классификация, риски и методы снижения

Научный руководитель – Прончев Геннадий Борисович

Андронов А.М.[1], *Гаранин Д.К.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Высшая школа государственного администрирования (факультет), Москва, Россия, *E-mail: andronovam@my.msu.ru*; 2 - Московский государственный университет имени М.В.Ломоносова, Высшая школа государственного администрирования (факультет), Москва, Россия, *E-mail: garanindk@my.msu.ru*

Статья посвящена "галлюцинациям" в генеративных нейросетевых моделях, преимущественно больших языковых моделях (англ. : LLM Large Language Model) для обработки естественного языка (англ. : NLP Natural Language Processing).

"Галлюцинация" определяется как генерация утверждений, не подтверждаемых проверяемыми источниками или установленными фактами, при сохранении грамматической корректности и семантической связности ответа.

Рассматривается место генеративных моделей в структуре ИИ (машинное обучение, глубокое обучение) и специфика их обучения на масштабных корпусах данных.

Систематизируются причины возникновения "галлюцинаций":

- вероятностный характер генерации
- недостаточная репрезентативность и шум обучающих данных
- противоречивость источников
- сдвиг распределения между обучением и применением
- ограничения контекстного представления и особенности "дообучения" на инструкциях.

Предлагается классификация "галлюцинаций" (фактические, атрибутивные, контекстные, логико-каузальные, вычислительные) и анализируются последствия для достоверности результатов, принятия решений и информационной безопасности.

Обобщаются методы снижения и обнаружения галлюцинаций:

- улучшение качества данных, внешняя верификация знаний (retrieval-подходы и базы знаний)
- обучение отказу при недостатке информации
- калибровка уверенности
- регламенты использования и процедуры оценивания качества генерации.

Полученные выводы применимы для проектирования и эксплуатации нейросетевых систем в задачах, требующих проверяемости и надежности, в данном случае систем государственного управления и администрирования.

Метод представленный в данной статье не является прямым решением проблемы "галлюцинаций", так как от данного феномена на сегодняшний день не выходит избавиться в силу природы его возникновения, а точнее ошибок, которые зачастую наблюдаются на этапе глубокого обучения модели и построения связей между её параметрами. Следовательно, метод нацелен на минимизацию случаев появления "галлюцинаций". Суть метода заключается в создании алгоритма для определения достоверности источника информации (и/или в создании стандарта оценки достоверности источника по сетевым маркерам), который будет решать проблему "шума", "противоречивости" и "репрезентативности" источников, при интеграции нейросетевых моделей во вспомогательные системы для государственного администрирования.

Алгоритм для определения достоверности источника информации реализуется на базе имеющихся маркеров в порядке убывания их достоверности:

1. Рецензируемые научные статьи (peer-reviewed journals)
2. Монографии, академические книги
3. Отчёты международных организаций (WHO, OECD и т.д.)
4. Официальная статистика
5. СМИ
6. Блоги, форумы

Источники и литература

1) New sources of inaccuracy? A conceptual framework for studying AI hallucinations | HKS Misinformation Review, DOI: https://doi.org/10.37016/mr-2020-182.
2) (PDF) Comprehensive Review of AI Hallucinations: Impacts and Mitigation Strategies for Financial and Business Applications, DOI: 10.7753/IJCATR1406.1003
3) Н.В. Макарова. Информатика. https://library.cbr.ru/catalog/lib/books/284393/
4) Brian D. Ripley. Pattern Recognition and Neural Networks https://books.google.com /bo oks?id=m12UR8QmLqoC