**Большие языковые модели для предсказания выходов реакций: скрытые возможности и перспективы**

***Орлова А.А., Дмитренко А.В., Дмитренко А.В., Виноградов В.В.***

*Аспирант, 1 год обучения*

*Национальный исследовательский университет ИТМО, Санкт-Петербург, Россия*

*E-mail:* *orlova@scamt-itmo.ru*

В последние годы развитие искусственного интеллекта и больших языковых моделей (LLM) привело к значительным прорывам в различных областях науки, включая химию [1]. Одной из важнейших задач в химической информатике остается предсказание выхода химических реакций. Несмотря на успехи машинного обучения в этой области, проблема точного прогнозирования выхода реакций остается нерешённой из-за ограниченности доступных данных и сложности химических процессов [2]. В данной работе исследуются возможности использования LLM для предсказания выхода химических реакций и анализируются их скрытые возможности в данной области.

Основной целью исследования является оценка эффективности LLM в задаче предсказания выхода химических реакций. Для этого были сформированы четыре различных набора данных, содержащие реакции из открытых баз USPTO и ORD. Были отобраны ведущие универсальные языковые модели и проведён их систематический анализ в сравнении с классическими методами машинного обучения и специализированными химическими моделями. Дополнительно изучались возможности использования эмбеддингов LLM для дальнейшего обучения предсказательных моделей.

Анализ показал, что модели Mistral Small и Claude 3 Haiku, несмотря на их меньшую популярность в сравнении с новейшими LLM, продемонстрировали превосходную точность в условиях обучения с малым числом примеров (few-shot learning). Они превзошли базовые модели на 3% по точности и F1-метрике. Кроме того, исследование выявило, что модели машинного обучения, обученные на эмбеддингах LLM, демонстрируют повышение точности до 2% по сравнению с традиционными методами. Было также отмечено, что в некоторых случаях универсальные языковые модели превосходят специализированные химические модели, что указывает на наличие в их эмбеддингах значимой информации о химических реакциях.

Полученные результаты свидетельствуют о значительном потенциале LLM в предсказательной химии. Универсальные языковые модели могут не только успешно справляться с задачами классификации выхода реакций, но и предоставлять ценные эмбеддинги для последующего машинного обучения. Данное исследование открывает перспективы для дальнейшего изучения скрытых возможностей LLM в химии и их интеграции в существующие методы предсказательного анализа.



Рис. 1. Сравнение LLM в задаче бинарной классификации выходов реакций

**Литература**

1. Guo T. et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks // NeurIPS. – 2023. – Т. 36. – С. 59662-59688.
2. Schwaller P. et al. Prediction of chemical reaction yields using deep learning // Mach. Learn.: Sci. Technol. – 2021. – Т. 2. – №. 1. – С. 015016.