

УДК 004.89

ПОВЫШЕНИЕ ДОВЕРИЯ К ПРЕДОБУЧЕННЫМ МОДЕЛЯМ ЧЕРЕЗ АДАПТИВНУЮ НАСТРОЙКУ

Нефедов В.В., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

С ростом масштабов и сложности предобученных моделей, таких как трансформеры в языковых задачах или глубокие сверточные сети в компьютерном зрении, становится всё более очевидной проблема несоответствия между универсальностью модели и спецификой прикладной задачи.

Предобученные модели обладают высокими обобщающими способностями и часто демонстрируют отличные результаты на разнообразных наборах данных, однако при переносе в конкретные области применения – например, в медицину, юриспруденцию, промышленную диагностику или образование – возникает разрыв между потенциалом модели и доверием к ней со стороны пользователей, разработчиков и регуляторов. Этот разрыв обусловлен не только техническими ограничениями, но и отсутствием гибкости, интерпретируемости, объяснимости и предсказуемости поведения модели в новых контекстах. В этой связи важнейшим направлением становится повышение доверия к предобученным моделям через адаптивную настройку, то есть согласование поведения модели с контекстом, требованиями и ожиданиями конечной среды применения [1].

Адаптивная настройка (adaptive fine-tuning) – это обобщённое понятие, включающее в себя не только классическое дообучение модели на новых данных, но и более сложные стратегии, такие как модульная адаптация, настройка через параметры prompt-интерфейсов, использование обучаемых адаптеров, многозадачное дообучение, регуляризация с контролем сдвига и динамическое перенастроение поведения модели во время её использования. Основной целью таких стратегий является обеспечение соответствия модели не только формальным метрикам производительности, но и параметрам доверия – устойчивости к нетипичным входам, прозрачности логики вывода, снижению предвзятости, способности сообщать об ошибке или неопределённости.

Одним из наиболее значимых подходов в этом направлении является использование адаптивных слоёв или адаптеров, которые встраиваются между замороженными слоями предобученной модели и минимально изменяют её поведение. Это позволяет сохранить глобальную структуру знаний, накопленных

при предобучении, и одновременно адаптировать поведение к особенностям конкретной задачи.

Адаптеры могут быть обучены на ограниченном объёме целевых данных и при этом обеспечивать высокую устойчивость и интерпретируемость вывода. Более того, такая модульность позволяет встраивать дополнительные механизмы контроля: например, слой, анализирующий уверенность предсказания или блок, фильтрующий токсичность текста. Таким образом, адаптация становится точкой входа для внедрения доверительных механизмов в архитектуру модели, без необходимости её полной переработки.

Другим важным аспектом адаптивной настройки является контроль за смещением модели при дообучении. Известно, что при агрессивном fine-tuning возможно разрушение исходных полезных свойств модели: ухудшается способность к обобщению, растёт переобучение, нарушаются изначально заложенные симметрии и свойства, например, справедливость или сбалансированность. Для предотвращения этого разрабатываются методы регуляризации с контролем дрейфа – такие как elastic weight consolidation, KL-регуляризация, адаптивное смещение параметров (weight interpolation) и сравнение латентных пространств до и после адаптации. Эти методы позволяют адаптировать модель, сохраняя при этом свойства, критичные для доверия: например, корректность при редких входах, предсказуемость поведения, устойчивость к атакующим шумам [2].

Особое значение имеет адаптация в условиях ограниченного доступа к целевым данным. В задачах с высокими требованиями к приватности или при невозможности доступа к исходным данным (например, в здравоохранении) используются методы federated fine-tuning и privacy-preserving adaptation, позволяющие обучать адаптационные компоненты на локальных данных без их передачи на центральный сервер. Такие подходы повышают доверие не только к поведению модели, но и к самой инфраструктуре, в которой она функционирует [3].

Дополнительно адаптация может осуществляться через динамическую настройку поведения модели в зависимости от контекста. Современные языковые модели всё чаще используют prompt-интерфейсы, в которых контекст и инструкция задаются во входном тексте. В этом случае настройка осуществляется не через изменение параметров модели, а через управление её вниманием и генерацией с помощью специально сконструированных запросов. Такие подходы позволяют формировать модели, которые адаптируются «на лету», и повышают прозрачность их поведения, поскольку инструкция или ограничение явно указаны. Более того, динамическая адаптация может быть управляемой – например, система может изменять поведение модели в зависимости от профиля

пользователя, чувствительности данных, текущей задачи или уровня допустимого риска.

Важно отметить, что доверие формируется не только на уровне архитектуры, но и через восприятие модели пользователем, что означает необходимость адаптации интерфейса и объяснений. Адаптивная настройка должна включать в себя и адаптацию методов интерпретации, объясняющих решения модели в понятной и целевой для пользователя форме. Это может быть, например, использование различных форм визуализации в задачах компьютерного зрения или генерация текстовых обоснований в языковых системах. При этом важно, чтобы интерпретации тоже адаптировались к конкретной задаче, учитывали уровень экспертности пользователя и не вводили в заблуждение [4].

Наконец, адаптивная настройка может сопровождаться мониторингом поведения модели в реальной среде и обратной связью от пользователя, что позволяет формировать замкнутый цикл доверия. На основе анализа ошибок, отклонений, пользовательских жалоб или повторных запросов можно автоматически или полуавтоматически перенастраивать модель, корректировать адаптеры, усиливать регуляризацию или обновлять стратегии генерации. Такой подход делает возможной эволюцию доверия – модель со временем становится более предсказуемой, прозрачной и согласованной с ожиданиями конкретной организации или прикладного сообщества [5].

Таким образом, адаптивная настройка – это не просто технический механизм повышения производительности, а полноценный инструмент формирования и укрепления доверия к предобученным моделям. Её применение позволяет согласовать универсальные знания, полученные в ходе предобучения, с конкретной логикой, ограничениями и рисками прикладной задачи [6].

В перспективе такие адаптивные и доверительные системы станут основой зрелой ИИ-инфраструктуры, где каждый шаг модели – от генерации вывода до его обоснования и обновления – будет понятен, контролируем и верифицируем, соответствуя как инженерным, так и социальным требованиям.

Список литературы:

1. Howard, J. Universal Language Model Fine-Tuning for Text Classification / J. Howard, S. Ruder // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2018. – P. 328–339. – DOI 10.18653/v1/P18-1031.
2. Devlin, J. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of NAACL-HLT 2019. – Minneapolis : ACL, 2019. – P. 4171–4186. – DOI 10.18653/v1/N19-1423.

3. Houlsby, N. Parameter-Efficient Transfer Learning for NLP / N. Houlsby, A. Giurgiu, S. Jastrzebski [et al.] // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 2790–2799.

4. Hu, E. J. LoRA : Low-Rank Adaptation of Large Language Models / E. J. Hu, Y. Shen, P. Wallis [et al.] // International Conference on Learning Representations. – 2022.

5. Li, X. L. Prefix-Tuning : Optimizing Continuous Prompts for Generation / X. L. Li, P. Liang // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2021. – P. 4582–4597. – DOI 10.18653/v1/2021.acl-long.353.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2025664977 Российская Федерация. Программа для компенсации эффекта взрывающегося градиента в нейронных сетях пользователя : заявл. 14.05.2025 : опубл. 09.06.2025 / Е. А. Николаева, П. А. Пылов, Р. В. Майтак ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Кузбасский государственный технический университет имени Т.Ф. Горбачева». – EDN KHFZJQ.