УДК 004.89

РАЗРАБОТКА МЕТРИК ДОВЕРИЯ К ПРЕДОБУЧЕННЫМ ЯЗЫКОВЫМ И ВИЗУАЛЬНЫМ МОДЕЛЯМ

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

С развитием крупных предобученных моделей – как языковых, так и визуальных – встает всё более острый вопрос обоснованного доверия к их решениям [1].

Эти модели, обладая колоссальными параметрическими масштабами и высокой обобщающей способностью, часто используются как универсальные инструменты в задачах генерации текста, классификации изображений, распознавания объектов и обработки мультимодальных сигналов. Однако эффективность таких моделей не гарантирует их надежности в прикладном применении, особенно если отсутствует понимание границ применимости и уверенности в корректности конкретного вывода. Это делает необходимым не только интуитивное или эвристическое оценивание качества, но и разработку формальных метрик доверия, которые позволят количественно и воспроизводимо оценивать, насколько предсказание модели заслуживает доверия в конкретном контексте.

Метрика доверия – это числовой или категориальный показатель, отражающий степень уверенности, прозрачности, устойчивости и предсказуемости поведения модели в отношении конкретного входного примера или класса задач. В отличие от традиционных метрик, таких как точность, precision, recall или BLEU, метрика доверия не измеряет априорную производительность модели на тестовом наборе, а описывает способность модели вести себя надёжно на конкретном примере или в определённой ситуации.

Таким образом, метрика доверия служит механизмом локальной оценки и может применяться в реальном времени, формируя динамические сценарии обработки – например, маршрутизацию к человеку, отказ от ответа, или запуск дополнительных проверок [2].

В языковых моделях доверие часто ассоциируется с уровнем вероятности предсказания следующего токена или с логперплексией текста. Однако эти значения могут быть обманчивыми: высокая уверенность модели вовсе не гарантирует её правоту, особенно если входной текст содержит неочевидные, абсурдные или противоречивые формулировки. Поэтому современные подходы

стремятся дополнить вероятность токенов более структурированными метриками – например, энтропией распределения по выходным токенам, стабильностью при переформулировке запроса, чувствительностью к небольшим сдвигам формулировки и когерентностью ответа на уровне документа.

Используются и методы анализа внутренних слоев модели, позволяющие выявить, насколько активации модели согласованы с известными структурами языка или с обучающим распределением.

В визуальных моделях – особенно в архитектурах типа CNN и ViT – доверие к выводу также сложно вывести из одной лишь вероятности класса. Здесь на первый план выходят метрики пространственной уверенности, такие как плотность карты внимания, степень фокусировки на семантически релевантных регионах, устойчивость к геометрическим и фотометрическим преобразованиям изображения [3].

Популярны также градиентные методы оценки важности входов (например, Grad-CAM), позволяющие сопоставить активные регионы модели с ожидаемыми признаками объекта. Однако даже при визуальной интерпретации важно учитывать, что высокая активность в «правильной» области может возникать по статистическим причинам и не всегда отражает истинную причинность предсказания.

Разработка метрик доверия требует выработки общих критериев того, что считать "надёжным" предсказанием. Одним из таких критериев является калиброванность: хорошо откалиброванная модель должна с заданной вероятностью быть права, если она указывает на эту вероятность. Например, среди всех предсказаний с уверенностью 80%, правильными должно быть примерно 80%.

Разрабатываются специальные методы температурной коррекции и многомодельной ансамблизации для достижения калибровки, и метрики вроде Expected Calibration Error (ECE) становятся стандартом в оценке доверия. Визуальные и языковые модели адаптируют эти подходы к своим специфическим выходам, иногда вводя локальные меры калиброванности, зависящие от типа входа [4].

Другим важным направлением является анализ стабильности предсказаний. Если при малейших изменениях входа (например, при переформулировке фразы или вращении изображения) предсказания модели резко меняются, это является сигналом низкого доверия. Разрабатываются метрики устойчивости (robustness scores), измеряющие амплитуду флуктуаций предсказаний в окрестности входа, и метрики консистентности, оценивающие совпадение выводов при перетестировании с близкими данными [5].

В дополнение к этим подходам предлагаются метрики контекстной релевантности и правдоподобия. В языковых задачах они могут учитывать не только статистические, но и семантические критерии – например, соответствие

ответа запросу, логическую связность, отсутствие противоречий и наличие обоснований. На практике это реализуется через модели оценки правдоподобия (naturalness scoring), дискурсивной связности и фактологической достоверности, что особенно важно при применении LLM в сфере генерации знаний.

Одним из перспективных направлений становится интеграция метрик доверия с механизмами самооценки модели, когда сама модель предсказывает вероятность того, что её ответ ошибочен, некорректен или выходит за рамки области применимости.

Такие механизмы активно разрабатываются в трансформерах нового поколения, где реализуются уровни self-verification, re-ranking, rerouting и deferred decision-making. Метрики доверия в этом случае становятся не только инструментом внешнего аудита, но и частью архитектуры модели, влияя на её поведение в реальном времени.

Кроме чисто технических аспектов, метрики доверия должны быть пригодны к регуляторному применению и сертификации. Это означает, что они должны быть интерпретируемыми, формализуемыми и воспроизводимыми.

В этом контексте важно создавать методологические рамки оценки доверия, включающие в себя процедуры тестирования, пороговые значения, процедуры отклонения вывода, документирование случаев риска и обоснование допустимости вывода. Метрики доверия становятся основой для построения систем цифровой ответственности и для внедрения ИИ в критически важные системы, где требуется доказуемое соответствие модели требованиям безопасности, надёжности и социальной приемлемости.

Таким образом, разработка метрик доверия к предобученным языковым и визуальным моделям представляет собой ключевое направление развития зрелой инфраструктуры ИИ. Такие метрики необходимы не только для технической отладки и отбора моделей, но и для формирования доверительных интерфейсов взаимодействия с человеком, обеспечения нормативной прозрачности и построения устойчивых систем принятия решений.

По мере того, как ИИ-системы становятся всё более автономными, метрики доверия превращаются из факультативного инструмента в обязательный компонент ответственного и контролируемого машинного обучения.

Список литературы:

1. Brier, G. W. Verification of Forecasts Expressed in Terms of Probability / G. W. Brier // Monthly Weather Review. – 1950. – Vol. 78, № 1. – P. 1–3.

2. Murphy, A. H. A New Vector Partition of the Probability Score / A. H. Murphy // Journal of Applied Meteorology. – 1973. – Vol. 12, № 4. – P. 595–600.

3. Naeini, M. P. Obtaining Well Calibrated Probabilities Using Bayesian Binning / M. P. Naeini, G. Cooper, M. Hauskrecht // Proceedings of the Twenty-Ninth

AAAI Conference on Artificial Intelligence. – Austin : AAAI Press, 2015. – P. 2901–2907.

4. Kull, M. Beyond Temperature Scaling : Obtaining Well-Calibrated Multi-Class Probabilities with Dirichlet Calibration / M. Kull, M. Perello-Nieto, M. Kängsepp [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

5. Corbiere, C. Addressing Failure Prediction by Learning Model Confidence / C. Corbiere, N. Thome, A. Bar-Hen [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.