Галлюцинации как примеры вне распределения в больших языковых моделях: подход к обнаружению с теоретическими гарантиями

Рындин Максим Алексеевич

ИСП РАН
ИЦ ДИИ

OOD

Задача OOD – отказ от решения задачи, если задача поменялась (сдвиг концепций, сдвиг признаков).

- Головная боль разработчика.
- Для классики понятно, что такое сдвиг концепций.
- Для LLM – можно расценивать как ответ на неизвестные вопросы.

Галлюцинации – OOD для LLM.

LLM OOD постановка

Как и для OOD хотим

1. **White-box**: Доступ только к внутренним скрытым состояниям (hidden states) модели.
2. **Post-hoc**: Детекция выполняется после генерации (без вмешательства в процесс декодирования).
3. **Knowledge-free**: Без использования внешних баз знаний (RAG и т.д.).
4. **Unsupervised**: Без ручной разметки человеком (используются только псевдометки).

SOTA

IRIS удовлетворяет всем 4 критериям (анализ Chain-of-Thought эмбеддингов **последнего слоя** LLM).

Фундаментальные уязвимости IRIS:

- **Катастрофическая деградация при сдвиге домена (Domain Shift):** При переносе на новые данные (Cross-Domain) IRIS сохраняет лишь 76.1% своей точности. Точность падает практически до случайного угадывания.
- **Отсутствие гарантий надежности:** Точечные предсказания не имеют distribution-free меры уверенности. Модель не знает, когда ей следует "воздержаться"от ответа.

Направления развития Создать детектор, устойчивый к доменному сдвигу (Domain Shift) и теоретически обоснованный.

Обзор архитектуры PLLS и PLLS-CP

Предлагаемый фреймворк **PLLS** (Pseudo-Label Layer Selection) состоит из 5 этапов:

1. **Зондирование слоев (Layer Probing):** Обучение легковесных проб на псевдометках для каждого слоя $\mathcal{L}$.
2. **Top-K Selection:** Выбор наиболее информативных слоев на основе композитного скора.
3. **Классификатор:** Обучение классификатора на выбранных слоях.
4. **Ансамблирование (Ensemble):** Выпуклая комбинация предсказаний полученного классификатора и базового IRIS.
5. **Конформная калибровка (MCP):** Оборачивание ансамбля в Mondrian Conformal Prediction для получения гарантий.

Для кросс-доменного переноса (PLLS-CP) обновляется только конформный порог $\hat{q}$ на целевой калибровочной выборке.

Этап 1: Извлечение псевдоразметки

Проблема: Нам запрещено использовать разметку человеком. Откуда взять Ground Truth?

Решение (Протокол IRIS): Использование *вербализованной неуверенности* (verbalized confidence) самой LLM.

- LLM генерирует рассуждение (Chain-of-Thought) по утверждению x.
- LLM выдает финальную оценку своей уверенности $c(x) \in [0, 1]$ или бинарный ответ.
- Псевдометка: $\tilde{y}(x) = \mathbb{I}[c(x) < \tau]$.

Все обучение (выбор слоев, веса ансамбля, пороги калибровки) происходит **исключительно на этих псевдометках**. Истинные метки (Ground Truth) используются только для финальной тестовой оценки.

Этап 2: Выбор информативного слоя (PLLS)

IRIS использует только эмбеддинги **последнего слоя** $h^{(L-1)}$. Однако для разных доменов наиболее богатые представления могут лежать на разных глубинах.

Обучаем логистическую регрессию для каждого кандидата $l \in \{0, 4, 8, 12, 16, 20, 24, 28\}$ и оцениваем их на валидации.

Критерий отбора (Композитный скор):

$$score_l = \lambda \cdot \mathrm{AUC}_l + (1 - \lambda) \cdot (1 - \mathrm{ECE}_l)$$

- AUC_l — разделительная способность.
- ECE_l (Expected Calibration Error) — штраф за плохую калибровку (важно для конформного слоя!).
- Эмпирически $\lambda = 0.7$ дает наилучшую стабильность.
- Мы выбираем $K = 1$ лучших слоев (эксперименты показали, что агрегация многих слоев добавляет шум).

Этапы 3 и 4: Классификатор и Ансамблирование

На выбранном слое обучается MLP-классификатор, выдающий вероятность $\hat{p}_{\text{PLLS}}(y|x)$.

Ансамблирование с IRIS: Чтобы обеспечить стабильность и исключить деградацию по сравнению с SOTA, строим выпуклую комбинацию:

$$\hat{p}_{\text{ens}}(y|x;\alpha) = \alpha \cdot \hat{p}_{\text{PLLS}}(y|x) + (1-\alpha) \cdot \hat{p}_{\text{IRIS}}(y|x)$$

Где $\alpha \in \{0, 0.1, \ldots, 1.0\}$.

Оптимальный параметр $\hat{\alpha}$ выбирается на валидационной выборке путем минимизации эмпирической функции потерь (Cross-Entropy/Brier score).

Теорема: Гарантия не-ухудшения ансамбля

Теорема (Ensemble non-degradation)

Пусть $L(p, y) \in [0, M]$ — ограниченная функция потерь. Пусть $p_{ens}(\alpha) = \alpha p_{PLLS} + (1 - \alpha) p_{IRIS}$, где α выбирается из конечной сетки $\mathcal{A} \subset [0, 1]$ размера $|\mathcal{A}|$, содержащей 0, по правилу:

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \hat{L}_{val}(\alpha)$$

на валидационной выборке размера N_{val}.
Тогда с вероятностью не менее $1 - \gamma$:

$$L_{ens}(\hat{\alpha}) \leq L_{IRIS} + 2M\sqrt{\frac{\ln(2|\mathcal{A}|/\gamma)}{2N_{val}}}$$

Пояснение к Теореме

Физический смысл: Поскольку в сетку параметров перебора включен $\alpha = 0$ (что соответствует чистому IRIS), наш ансамбль на обучающей выборке всегда не хуже базового метода.

Теорема формализует это для тестовых данных:

- Наш ансамбль может проиграть IRIS не более чем на величину $\mathcal{O}\left(1/\sqrt{N_{val}}\right)$.
- При достаточном размере N_{val} метод **гарантированно не деградирует** относительно SOTA-бейзлайна.
- Это критически важно, так как конформному слою на следующем этапе нужен максимально стабильный скор.

Этап 5: Mondrian Conformal Prediction (MCP)

Вместо точечного предсказания (0 или 1) мы переходим к предсказывающим множествам $C(x)$.

Вводятся **функции неконформности** (nonconformity scores): $s(x, y) = 1 - \hat{p}_{ens}(y|x)$.

По калибровочной выборке $\mathcal{D}_{calib}$ вычисляются поклассовые квантили $\hat{q}_y$:

$$\hat{q}_y = \text{Quantile}\left(\{s_i\}_{y_i=y}, \frac{\lceil (n_y + 1)(1 - \epsilon) \rceil}{n_y}\right)$$

Предиктивное множество: $C(x) = \{y : s(x, y) \leq \hat{q}_y\}$.

Оно может содержать:

- Один класс (уверенное предсказание).
- Пустое множество (аномалия).
- **Оба класса (Воздержание от ответа / Abstention)** — система признает свою некомпетентность на сложных примерах.

Кросс-доменный перенос (Cross-Domain Transfer)

Главная проблема развертывания: при смене домена распределение $P(X, Y)$ меняется $\rightarrow$ меняются распределения вероятностей классификатора $\rightarrow$ точность рушится.

Предлагаемая парадигма PLLS-CP:

1. Мы **НЕ** переобучаем классификатор ($\hat{f}_{PLLS}$ и $\hat{\alpha}$ заморожены).
2. Берем небольшую калибровочную выборку ($n \approx 100 - 200$) **из нового домена**.
3. Генерируем для нее псевдометки.
4. Пересчитываем **только конформный порог** $\hat{q}$ на новой выборке.

Достаточно ли изменения одного лишь порога для получения гарантий на новом домене?

Теорема: Кросс-доменное покрытие

Теорема (Cross-domain coverage under recalibration)

Пусть функция неконформности $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ фиксирована **до** наблюдения целевых данных (обучена на source-домене).
Пусть калибровочная выборка целевого домена $Z_{cal} = \{z_1, \ldots, z_n\}$ и новый тестовый объект z_{n+1} — **взаимозаменяемы** (exchangeable) из P_{target}.
Определим $\hat{q}$ как эмпирический квантиль уровня $\frac{\lceil (n+1)(1-\epsilon) \rceil}{n}$ по Z_{cal}, и $C(x_{n+1}) = \{y : s(x_{n+1}, y) \leq \hat{q}\}$.
Тогда **маржинальное покрытие** на новом домене гарантировано:

$$\mathbb{P}[y_{n+1} \in C(x_{n+1})] \geq 1 - \epsilon$$

Пояснение к Теореме

Физический смысл: Даже если детектор обучался на совершенно другом (source) распределении и выдает смещенные или плохо откалиброванные скоры на target-домене, **свойство exchangeability (взаимозаменяемости) сохраняется**.

Поскольку параметры функции $s(\cdot)$ заморожены заранее, скалярные скоры $S_i = s(x_i, y_i)$ на калибровочной и тестовой выборке нового домена остаются независимыми и одинаково распределенными (i.i.d.).

Итог: Конформный классификатор можно безопасно перенести на новый домен, просто обновив порог на псевдоразмеченных примерах.

Теорема: Концентрация зазора покрытия (Coverage gap)

Маржинальное покрытие — это ожидание по всем возможным калибровочным выборкам. А что происходит при *конкретной, фиксированной* выборке $\mathcal{D}_{calib}$ размера n?

Теорема (Coverage gap concentration)

В условиях Теоремы о покрытии, условное покрытие $P_{cond} = \mathbb{P}[Y \in C(X)|\mathcal{D}_{calib}]$ имеет Бета-распределение:
$P_{cond} \sim Beta(k, n-k+1)$, где $k = \lceil (n+1)(1-\epsilon) \rceil$.
Следовательно, для любого $\delta > \frac{1}{n+1}$, отклонение покрытия ограничено экспоненциально:

$$\mathbb{P}\left[|P_{cond} - (1-\epsilon)| \geq \delta\right] \leq 2\exp\left(-2n\left(\delta - \frac{1}{n+1}\right)^2\right)$$

Пояснение к Теореме

Практический смысл: Насколько мы можем промахнуться мимо целевого покрытия $1 - \epsilon$ (например, 0.90) из-за конечного размера выборки?

При $n = 200$ примеров и допуске $\delta = 0.1$:

- $\frac{1}{n+1} \approx 0.005$.
- Вероятность того, что реальное покрытие будет отличаться от целевого более чем на 10%, оценивается как ≤ 0.054 (менее 5.4%).

Это доказывает, что кросс-доменная перекалибровка не только асимптотически верна, но и **практически надежна** на очень малых калибровочных выборках.

Экспериментальная установка

Модель: LLaMA-3.1-8B-Instruct (32 трансформерных слоя).

Бенчмарки (18 доменов, 324 пары переноса):

- **TrueFalse** (7 доменов: животные, города, компании, факты и др.)
- **HaluEval** (5 доменов: Bio-Medical, Education, Finance, Science и др.)
- **HELM** (6 модельно-специфичных доменов: falcon40b, gptj7b, opt7b и др.)

Метрики:

- *Accuracy* — точечная точность $\arg\max$.
- *Coverage (Покрытие)* — доля тестов, где истинный класс попал в $C(x)$.
- *Abstention rate* — доля отказов от ответа (размер $C(x) = 2$).
- *Selective accuracy* — точность только на тех примерах, где дан конкретный ответ.

Результаты: Точечные метрики внутри домена

Таблица усредненных метрик по 18 доменам (до конформного слоя):

Группа	**IRIS**	**PLLS** (наш)	$>$ **IRIS**	**Средний** Δ
Все 18 доменов	0.723	**0.758**	10/18	+0.035
Информативные (12)	0.787	**0.827**	9/11	**+0.040**
HELM (6)	0.594	**0.620**	1/6	+0.026

- На 17 из 18 доменов PLLS демонстрирует поведение не хуже базового.
- Наибольшие приросты на сложных доменах: Bio-Medical (+0.198), Science (+0.185).

Результаты: Кросс-доменный перенос (PLLS-CP)

Матрица переноса 18×18 (усреднение по 324 парам *source* → *target*).

	Accuracy		Selective Accuracy	
	IRIS	**PLLS-CP**	IRIS	**PLLS-CP**
Diagonal mean (In-domain)	0.723	0.758	0.652	0.683
Off-diagonal mean	0.545	**0.669**	0.606	**0.651**
Retention (Удержание)	0.754	**0.883**	0.929	**0.953**

Главное достижение: Прирост сохранения точности при переносе (Retention) составил **+12.9 п.п.**
На "информативных"доменах при переносе метод теряет менее **2.5%** выборочной точности.

Эмпирическая проверка конформных метрик

Внутри домена (целевое покрытие $1 - \epsilon = 0.90$):

	Coverage	**MAD(gap)**	**Abstention**	**Sel. Acc.**
IRIS + MCP	0.909	0.033	0.508	0.664
PLLS-CP	**0.920**	0.041	**0.502**	**0.675**

Кросс-доменное распределение зазора покрытия (Coverage gap):

- Стандартное отклонение зазора по 324 парам составило всего **0.029**.
- Максимальный зазор **0.073**.
- Это **строго согласуется с теорией**, предсказывающей экспоненциально низкую вероятность отклонений ≥ 0.1.

Абляции: Почему такая архитектура?

Вариант	**Non-worse / > / <**	**Mean Δ acc**
Мульти-слойная агрегация (без селекции)	13 / 12 / 5	+0.048
PLLS, $K = 4$, Композитный скор	15 / 9 / 3	+0.027
PLLS, $K = 1$, Только AUC-скор	14 / 6 / 4	+0.015
PLLS, $K = 1$, Композитный скор (наш)	**17 / 10 / 1**	**+0.035**

- Наивная агрегация всех слоев (Multi-layer) дает высокий средний прирост, но **нестабильна** (5 жестких просадок).
- Исключение члена калибровки ($1 - \mathrm{ECE}$) из скора снижает точность в 2 раза и ведет к регрессиям, так как конформный слой получает плохие вероятности.
- Выбор одного лучшего слоя ($K = 1$) оказался самым надежным.

Границы применимости: Проблема HELM

На бенчмарке HELM все методы (IRIS и PLLS) показывают точность около 0.5-0.6 (почти случайность), а Abstention rate (доля отказов) доходит до 0.79-0.95.

Почему это происходит?

- В TrueFalse / HaluEval утверждения оцениваются **самой моделью** (LLaMA-3). Когда модель генерирует CoT о факте, в котором она "сомневается", это отражается в ее скрытых состояниях.
- В HELM утверждения сгенерированы **другими моделями** (falcon40b, gptj7b), а наша модель выступает лишь как "пассивный читатель".
- При чтении чужого текста нет внутреннего конфликта генерации. Сигнал в эмбеддингах **отсутствует алгоритмически**.

Вывод: Методы на основе эмбеддингов могут детектировать только галлюцинации самой модели.

Заключение

Теоретический вклад:

1. Доказана экспоненциальная концентрация условного конформного покрытия и корректность переноса через перекалибровку.
2. Предложен механизм ансамблирования с гарантией не-деградации.

Практические результаты:

1. Разработан алгоритм PLLS для селекции слоев без учителя и пайплайн PLLS-CP для кроссдоменного переноса.
2. Удержание точности (Accuracy Retention) при переносе на новые домены увеличено с **0.754 до 0.883**, а удержание селективной точности (Selective Accuracy Retention) — с **0.929 до 0.953**.
3. Применение на новом домене требует только $\sim 100 - 200$ **примеров** для перекалибровки порога без затратной тонкой настройки параметров нейросети.