Применение больших языковых моделей для классификации табличных данных

Applying Large Language Models to Tabular Data Classification

The 9th International Conference in Deep Learning in Computational Physics

6-8 июля 2026, НИИЯФ МГУ, Москва, Россия

Факультет компьютерных наук, НИУ ВШЭ, Москва, Россия

Ижицкий Т. М., Ким М., Филатов Е., Матвеев Д., Ларионов А., Гущин М. И.

Докладывает: Ижицкий Тимофей

Бустинги сильны на таблицах, но не используют смысл признаков

Что видит бустинг

`42, 3, 1, 1506, …`

числа и целочисленные коды категорий

Из данных можно извлечь больше

age = **42**, job = **management**,

marital = **married**, balance = **1506**

имена и значения несут смысл

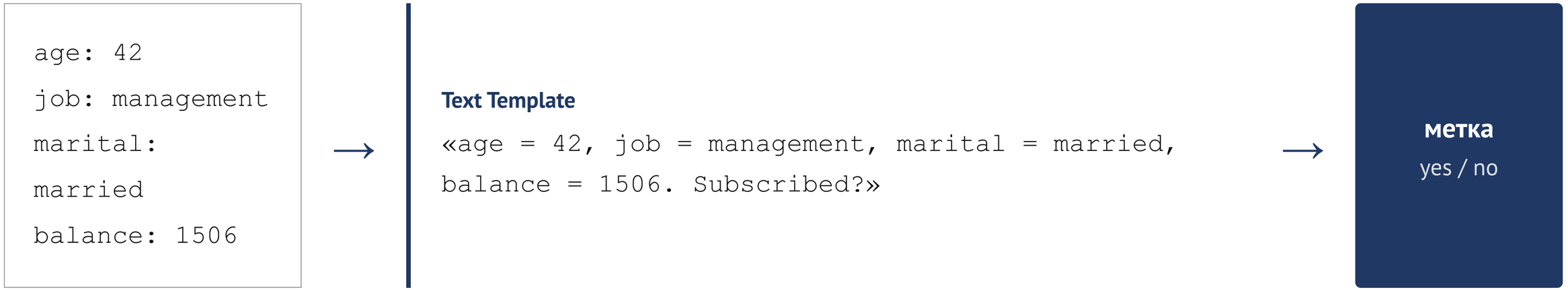
Может ли LLM догнать бустинг и стать foundation-моделью?

Проверяем три гипотезы

1	в режимах ICL или дообучения LLM конкурентны с CatBoost / LightGBM
2	дообученная LLM решает одновременно много задач без существенной потери качества
3	LLM могут хорошо работать при экстремальном количестве пропусков

Идея: строку таблицы сериализуем в текст и дообучаем LLM

Строка → текстовый шаблон → метка



- сериализация включает имена признаков, модель учитывает их смысл
- дообучаем через LoRA предсказывать правильную метку

Четыре способа использования LLM

01

Zero-shot

инструкция и один объект, веса не меняются

02

Few-shot

дополнительно 64 примера в промпте

03

Single fine-tune

LoRA под каждый датасет

04

Multitask

одна LoRA сразу на нескольких датасетах

меньше адаптации — больше адаптации

План экспериментов

Пайплайн (единый для всех моделей)

Данные - 9 датасетов, train/test 80/20

Сериализация - Text Template

Обучение - 4 режима, LoRA

Оценка - ROC-AUC, bootstrap ×1000

Почему именно эти модели

открытые модели 4-8B помещаются на одну GPU

разные модели: Qwen, Gemma, LLaMA, Phi

сопоставимы по размеру

Отдельно проверяем устойчивость: в датасетах случайно удаляем 20 / 50 / 90 % значений

Девять датасетов разных доменов

Датасет	Строк	Признаков	Классов	Область
Bank	45 211	16	2	Финансы
Blood	748	4	2	Медицина
California	20 640	8	2	Недвижимость
Car	1 728	6	4	Автомобили
Credit-G	1 000	20	2	Финансы
Diabetes	768	8	2	Медицина
Heart	918	11	2	Медицина
Income	48 842	12	2	Социология
Jungle	44 819	6	3	Игры

Без дообучения LLM заметно слабее классики

ROC-AUC по 9 датасетам в режимах zero-shot и few-shot (без изменения весов)

Режим	Модель	Bank	Blood	Calif.	Car	Credit	Diab.	Heart	Income	Jungle
Zero-shot	Qwen 3-4B	0,54	0,57	0,74	0,42	0,51	0,80	0,82	0,82	0,39
	LLaMA 3.1-8B	0,56	0,56	0,46	0,51	0,48	0,77	0,56	0,82	0,37
	Gemma 3-4B	0,61	0,61	0,47	0,46	0,45	0,71	0,56	0,67	0,40
	Phi-4-mini	0,57	0,72	0,59	0,69	0,48	0,82	0,50	0,80	0,60
Few-shot	Qwen 3-4B	0,58	0,71	0,74	0,79	0,60	0,75	0,89	0,84	0,46
	LLaMA 3.1-8B	0,64	0,65	0,79	0,87	0,50	0,76	0,90	0,84	0,49
	Gemma 3-4B	0,63	0,57	0,78	0,63	0,59	0,75	0,85	0,79	0,50
	Phi-4-mini	0,50	0,63	0,50	0,84	0,50	0,50	0,50	0,50	0,61
Справка	**CatBoost**	**0,94**	**0,71**	**0,97**	**1,00**	**0,84**	**0,83**	**0,96**	**0,93**	**0,97**

Дообучение решает: fine-tune » few- и zero-shot

Средний ROC-AUC по 9 датасетам

Модель	Zero-shot	Few-shot	Single FT	Multitask
Qwen 3-4B	0,62	0,71	**0,89**	**0,87**
Gemma 3-4B	0,55	0,68	**0,88**	**0,87**
LLaMA 3.1-8B	0,54	0,72	**0,87**	**0,88**
Phi-4-mini	0,64	0,56	**0,79**	**0,83**

Полное сравнение моделей: ROC-AUC по 9 датасетам

Модель	Bank	Blood	Calif.	Car	Credit	Diab.	Heart	Income	Jungle
Log. Regression	0,91	0,76	0,91	0,99	**0,86**	**0,87**	0,94	0,91	0,81
CatBoost	0,93	0,72	**0,97**	1,00	0,75	0,80	0,94	0,93	0,97
LightGBM	**0,94**	0,70	**0,97**	1,00	0,83	0,81	0,94	0,93	**0,98**
TABMLP	**0,94**	0,76	0,93	1,00	0,81	0,86	**0,97**	0,91	0,97
LSTM	0,90	**0,78**	0,93	1,00	0,76	0,71	0,92	0,91	**0,98**
Transformer	0,91	0,74	0,95	1,00	0,81	0,78	0,94	0,92	**0,98**
BERT	0,79	0,72	0,96	1,00	0,83	0,77	0,94	0,93	-
Qwen 3-4B	0,92	**0,77**	0,95	1,00	0,75	0,80	0,92	0,93	**0,98**
LLaMA 3.1-8B	0,91	0,74	0,96	1,00	0,73	0,80	0,94	0,93	0,83
Gemma 3-4B	0,92	0,76	0,95	1,00	0,74	0,78	0,92	0,93	0,90
Phi-4-mini	0,88	0,56	0,92	0,99	0,70	0,79	0,50	0,91	0,86
T0 11B	0,92	0,70	0,95	1,00	0,70	0,81	0,94	0,92	0,96
LLaMA-2 7B	**0,94**	0,72	**0,97**	1,00	0,76		0,93	0,93	**1,00**

Дообученные LLM в среднем не уступают классике

Средний ROC-AUC по 9 датасетам, по убыванию качества

Модель	ROC-AUC
TABMLP	0,91
LightGBM	0,90
CatBoost	0,89
Qwen 3-4B	0,89
Transformer	0,89

Модель	ROC-AUC
Gemma 3-4B	0,88
Log. Regression	0,88
LSTM	0,88
LLaMA 3.1-8B	0,87
BERT	0,87
Phi-4-mini	0,79

Multitask почти не отличается от single-task

ROC-AUC по датасетам, среднее по Qwen, LLaMA, Gemma

Режим	Bank	Blood	Calif.	Car	Credit	Diab.	Heart	Income	Jungle
Single	0,92	0,76	0,95	1,00	0,74	0,79	0,93	0,93	0,90
Multitask	0,92	0,75	0,92	1,00	0,66	0,80	0,92	0,93	0,97
Δ	0,00	−0,01	−0,03	0,00	**−0,08**	+0,01	−0,01	0,00	**+0,07**

При 90% пропусков LLM устойчивее бустинга

ROC-AUC по 8 датасетам, все модели, при росте доли пропущенных значений

%	Модель	Bank	Blood	Calif.	Car	Credit	Diab.	Heart	Income
	CatBoost	0,94	0,71	0,97	1,00	0,84	0,83	0,96	0,93
	BERT	0,79	0,72	0,96	1,00	0,83	0,77	0,94	0,93
0%	Qwen 3-4B	0,92	0,77	0,95	1,00	0,73	0,80	0,94	0,93
	Gemma 3-4B	0,90	0,75	0,96	1,00	0,71	0,78	0,92	0,92
	Phi-4-mini	0,88	0,56	0,92	0,99	0,70	0,79	0,50	0,91
	CatBoost	0,90	0,65	0,94	0,94	0,78	0,78	0,92	0,92
	BERT	0,78	0,73	0,92	0,95	0,78	0,69	0,89	0,92
20%	Qwen 3-4B	0,89	0,78	0,93	0,95	0,70	0,75	0,91	0,92
	Gemma 3-4B	0,88	0,74	0,94	0,96	0,63	0,74	0,90	0,91
	Phi-4-mini	0,84	0,56	0,87	0,95	0,72	0,74	0,50	0,90
	CatBoost	0,83	0,68	0,86	0,82	0,66	0,64	0,86	0,87
	BERT	0,75	0,70	0,84	0,85	0,70	0,65	0,85	0,87
50%	Qwen 3-4B	0,83	0,69	0,86	0,80	0,60	0,78	0,86	0,88
	Gemma 3-4B	0,81	0,66	0,83	0,81	0,61	0,71	0,85	0,85
	Phi-4-mini	0,69	0,59	0,79	0,85	0,56	0,69	0,50	0,86
	CatBoost	0,63	0,58	0,65	0,53	0,52	0,40	0,59	0,70
	BERT	0,64	0,57	0,61	0,56	0,59	0,57	0,56	0,69
90%	Qwen 3-4B	**0,69**	**0,64**	**0,68**	**0,66**	**0,51**	**0,62**	**0,77**	**0,76**
	Gemma 3-4B	0,66	0,60	0,62	0,59	0,50	0,56	0,62	0,71
	Phi-4-mini	0,58	0,57	0,66	0,69	0,48	0,60	0,50	0,73

До 50% бустинг сопоставим. При **90%** Qwen теряет меньше всех: **−0,21** против CatBoost **−0,32**

Выводы

1. после дообучения LLM имеет такое же качество, как у бустинга. ICL-режимы работают хуже
2. дообученную на всех датасетах LLM можно использовать как foundation-модель
3. при значительном количестве пропусков LLM работают лучше классики

Спасибо за внимание

Ижицкий Т. М., НИУ ВШЭ, Факультет компьютерных наук

Детали методики - в резервных слайдах

Гиперпараметры и инференс

Конфигурация LoRA	
rank r	**16**
alpha α	**32**
target modules	**q, k, v, o**
learning rate	**2×10^{-4}**
dropout	**0,05**
precision	**bf16**

Инференс

ответ задачи, для ROC-AUC - softmax по логитам токенов-меток

Обучение

cross-entropy только по токенам правильной метки

Метрики

ROC-AUC с bootstrap ×1000

Few-shot дороже в ~33 раза по длине промпта

Режим	Токенов	Относительно zero-shot
Zero-shot	168	×1,0
Single fine-tune	175	×1,0
Multitask	175	×1,0
Few-shot	**5600**	**×33**

64 примера в промпте делают few-shot **на порядок дороже** при инференсе

Полная таблица экспериментов с пропусками

%	Модель	Bank	Blood	Calif.	Car	Credit	Diab.	Heart	Income
0%	CatBoost	0,94	0,71	0,97	1,00	0,84	0,83	0,96	0,93
	BERT	0,79	0,72	0,96	1,00	0,83	0,77	0,94	0,93
	T0 3B (Full)	0,73	-	0,89	1,00	-	-	-	-
	Qwen 3-4B	0,92	0,77	0,95	1,00	0,73	0,80	0,94	0,93
	Gemma 3-4B	0,90	0,75	0,96	1,00	0,71	0,78	0,92	0,92
	Phi-4-mini	0,88	0,56	0,92	0,99	0,70	0,79	0,50	0,91
20%	CatBoost	0,90	0,65	0,94	0,94	0,78	0,78	0,92	0,92
	BERT	0,78	0,73	0,92	0,95	0,78	0,69	0,89	0,92
	T0 3B (Full)	0,59	-	0,80	0,76	-	-	-	-
	Qwen 3-4B	0,89	0,78	0,93	0,95	0,70	0,75	0,91	0,92
	Gemma 3-4B	0,88	0,74	0,94	0,96	0,63	0,74	0,90	0,91
	Phi-4-mini	0,84	0,56	0,87	0,95	0,72	0,74	0,50	0,90
50%	CatBoost	0,83	0,68	0,86	0,82	0,66	0,64	0,86	0,87
	BERT	0,75	0,70	0,84	0,85	0,70	0,65	0,85	0,87
	T0 3B (Full)	0,56	-	0,71	0,65	-	-	-	-
	Qwen 3-4B	0,83	0,69	0,86	0,80	0,60	0,78	0,86	0,88
	Gemma 3-4B	0,81	0,66	0,83	0,81	0,61	0,71	0,85	0,85
	Phi-4-mini	0,69	0,59	0,79	0,85	0,56	0,69	0,50	0,86
90%	CatBoost	0,63	0,58	0,65	0,53	0,52	0,40	0,59	0,70
	BERT	0,64	0,57	0,61	0,56	0,59	0,57	0,56	0,69
	T0 3B (Full)	0,50	-	0,54	0,52	-	-	-	-
	Qwen 3-4B	**0,69**	**0,64**	**0,68**	**0,66**	0,51	**0,62**	**0,77**	**0,76**
	Gemma 3-4B	0,66	0,60	0,62	0,59	0,50	0,56	0,62	0,71
	Phi-4-mini	0,58	0,57	0,66	0,69	0,48	0,60	0,50	0,73

При 90% пропусков **Qwen** держится выше остальных почти на всех датасетах. Прочерк - эксперимент не проводился