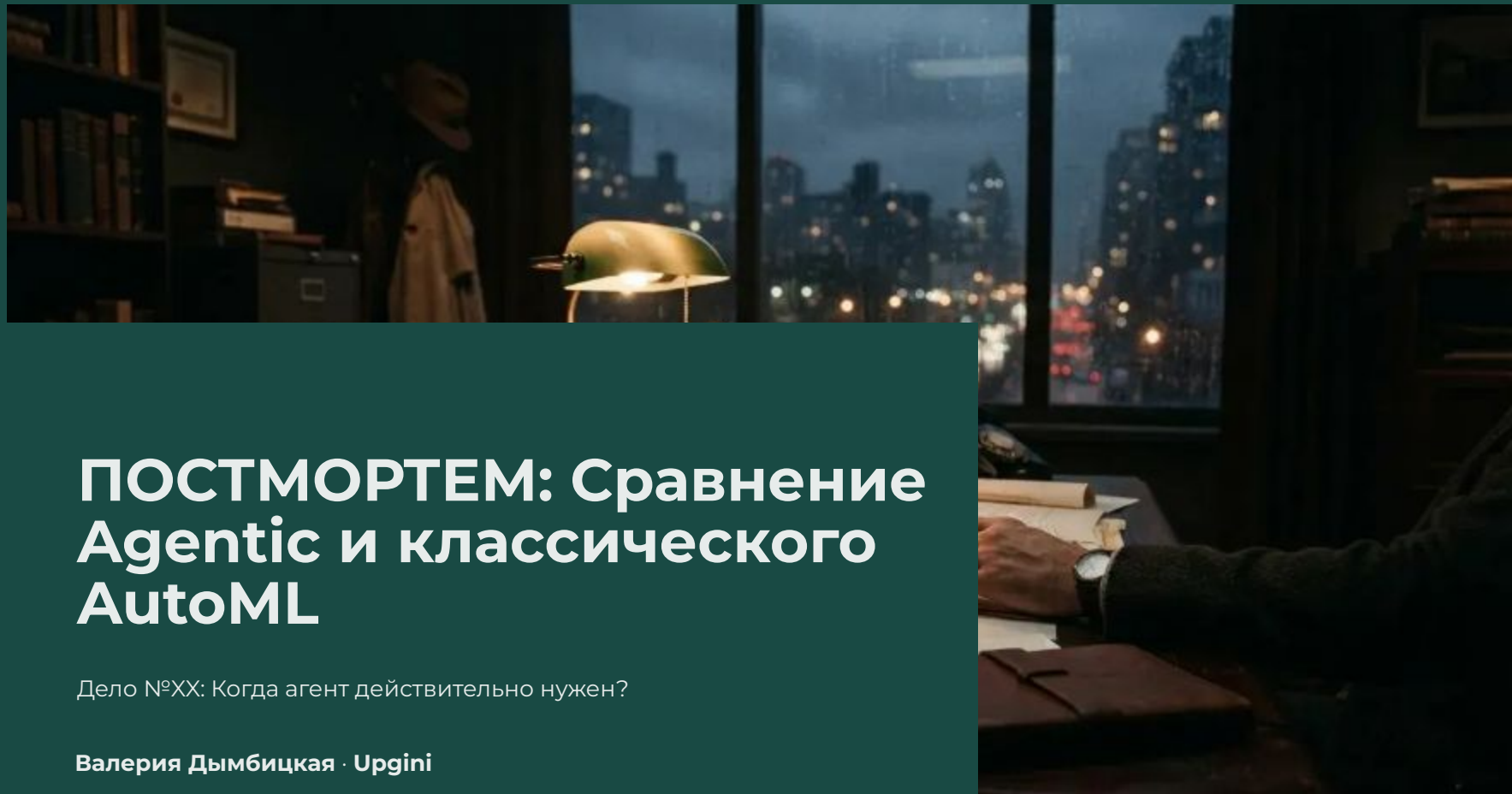




ПОСТМОРТЕМ: Сравнение Agentic и классического AutoML

Дело №XX: Когда агент действительно нужен?

Валерия Дымбицкая · Upsilon

Мы — Upgini: enrichment как часть (Auto)ML-пайплайна

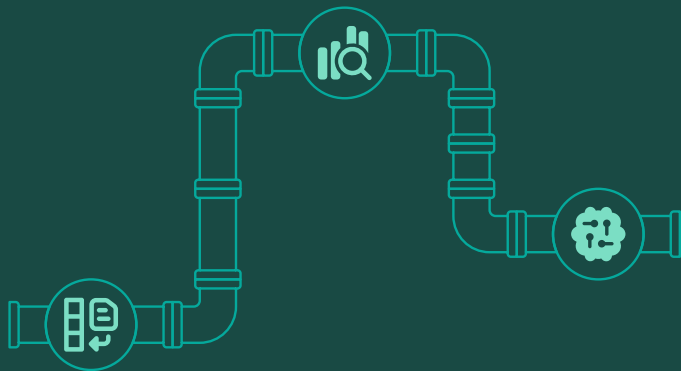


Мы — Upgini: enrichment как часть (Auto)ML-пайплайна

```
%pip install upgini
```

```
df = pd.read_csv(df_paths)
train_features = df.drop(["avg_salary"], axis=1)
train_target = df.avg_salary
```

```
enricher = FeaturesEnricher(
    search_keys={
        "country": SearchKey.COUNTRY,
        "Postal_code": SearchKey.POSTAL_CODE,
    }
)
enricher.fit(train_features, train_target,
             scoring="mean_absolute_error")
```



AutoML начинает становиться агентским



NVIDIA

3 LLM-агента
сгенерировали
600к+ строк кода,
провели 850
экспериментов и
заняли 1-е место в
churn prediction



Kaggle

Запуск
Autonomous Agent
Prediction (Beta)
для полностью
автономного
обучения и
сабмитов



Benchmarks

Появление
специализирован
ных фреймворков
для оценки data-
science агентов

AutoML начинает становиться агентским



NVIDIA

3 LLM-агента
сгенерировали
600к+ строк кода,
провели 850
экспериментов и
заняли 1-е место в
churn prediction



Kaggle

Запуск
Autonomous Agent
Prediction (Beta)
для полностью
автономного
обучения и
сабмитов



Benchmarks

Появление
специализирован
ных фреймворков
для оценки data-
science агентов

Что если встроить нашу библиотеку как
инструмент?

Два мира AutoML

Classic AutoML

Линейная
последовательность
шагов



Шаги predeterminedены



Отлаженный и
надёжный

Agentic AutoML



Цикл гипотеза-
проверка-оценка

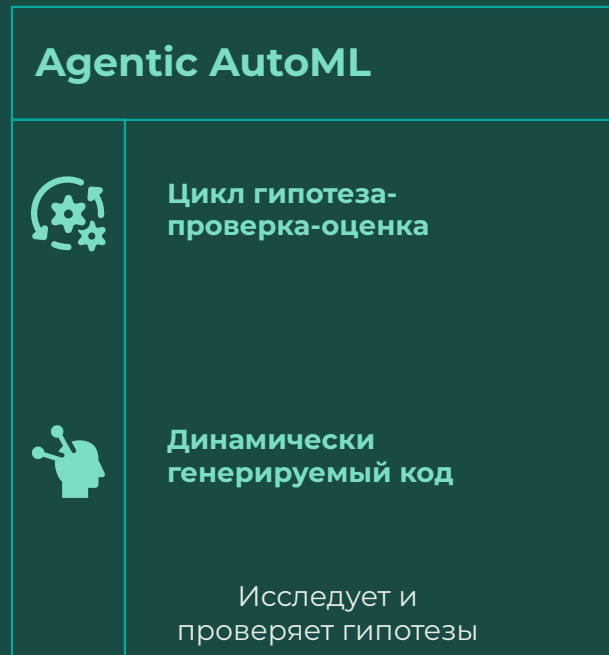


Динамически
генерируемый код

Исследует и
проверяет гипотезы

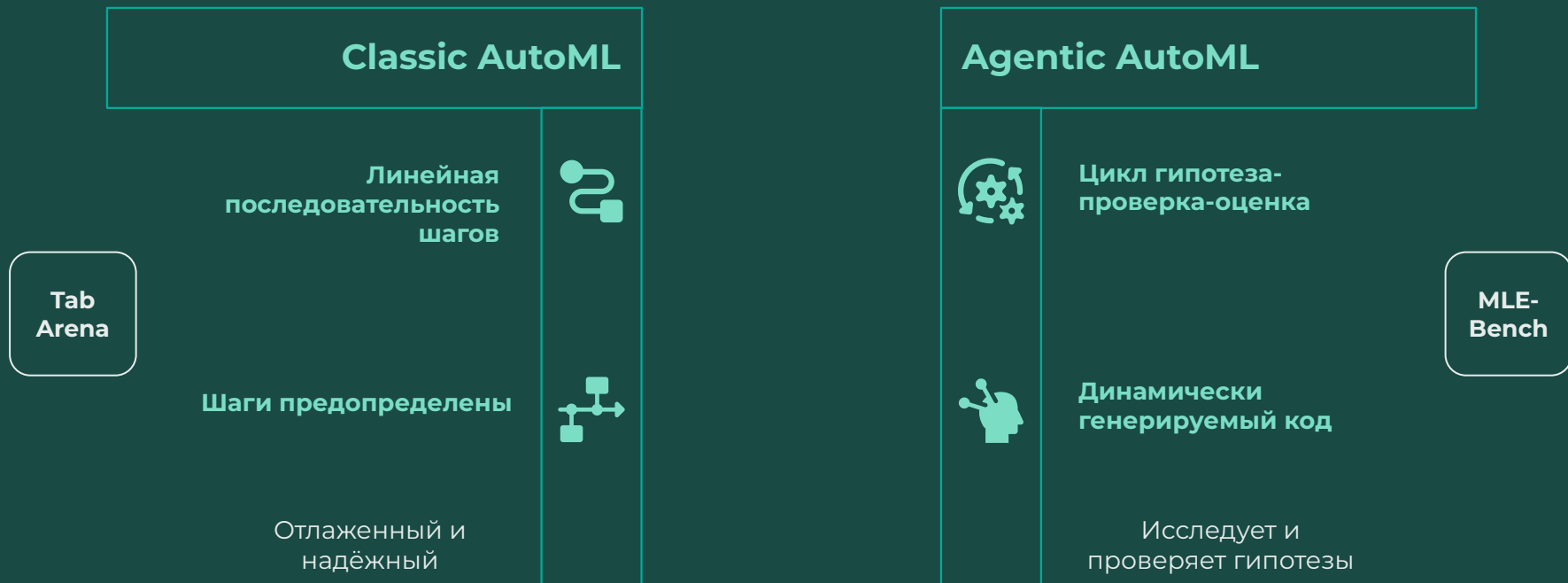
Два мира AutoML

Почти не сравниваются между собой!



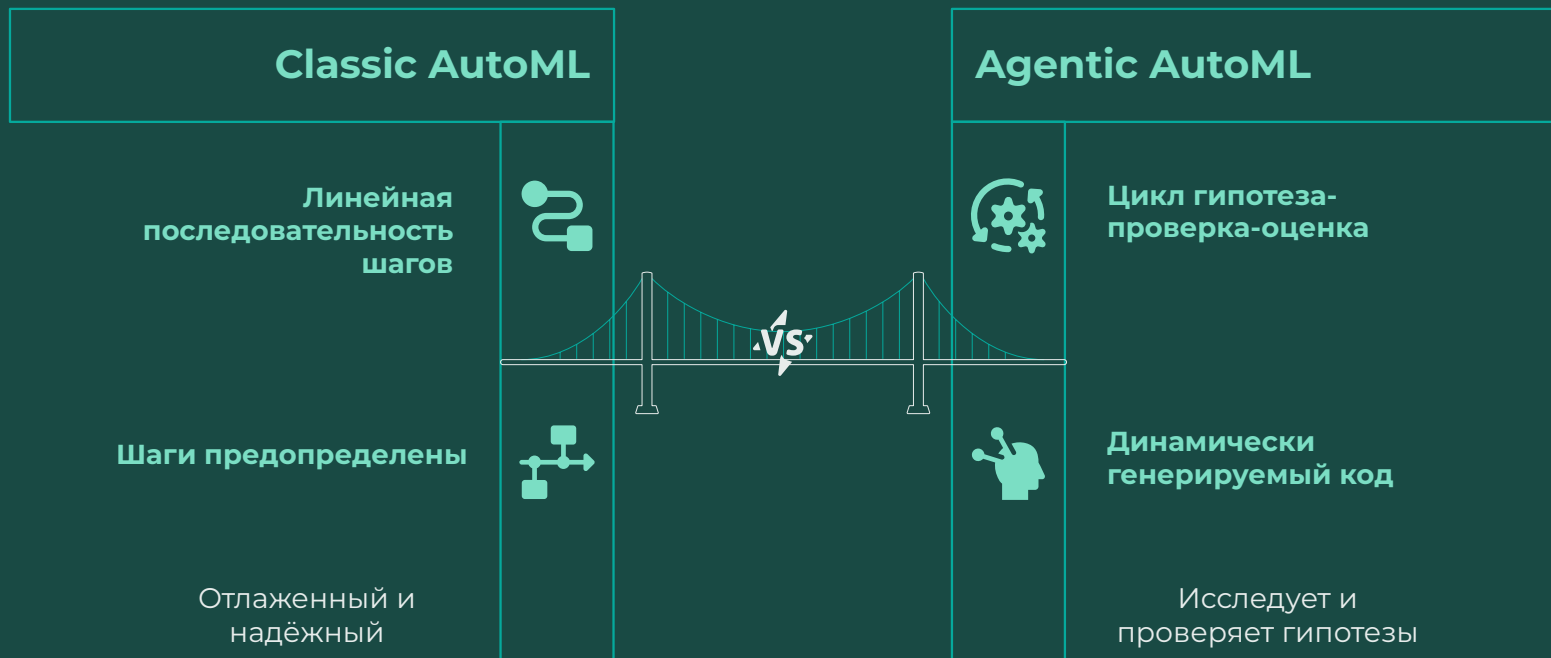
Два мира AutoML

Почти не сравниваются между собой!



Два мира AutoML

Почти не сравниваются между собой!



Эксперимент с двумя целями

01

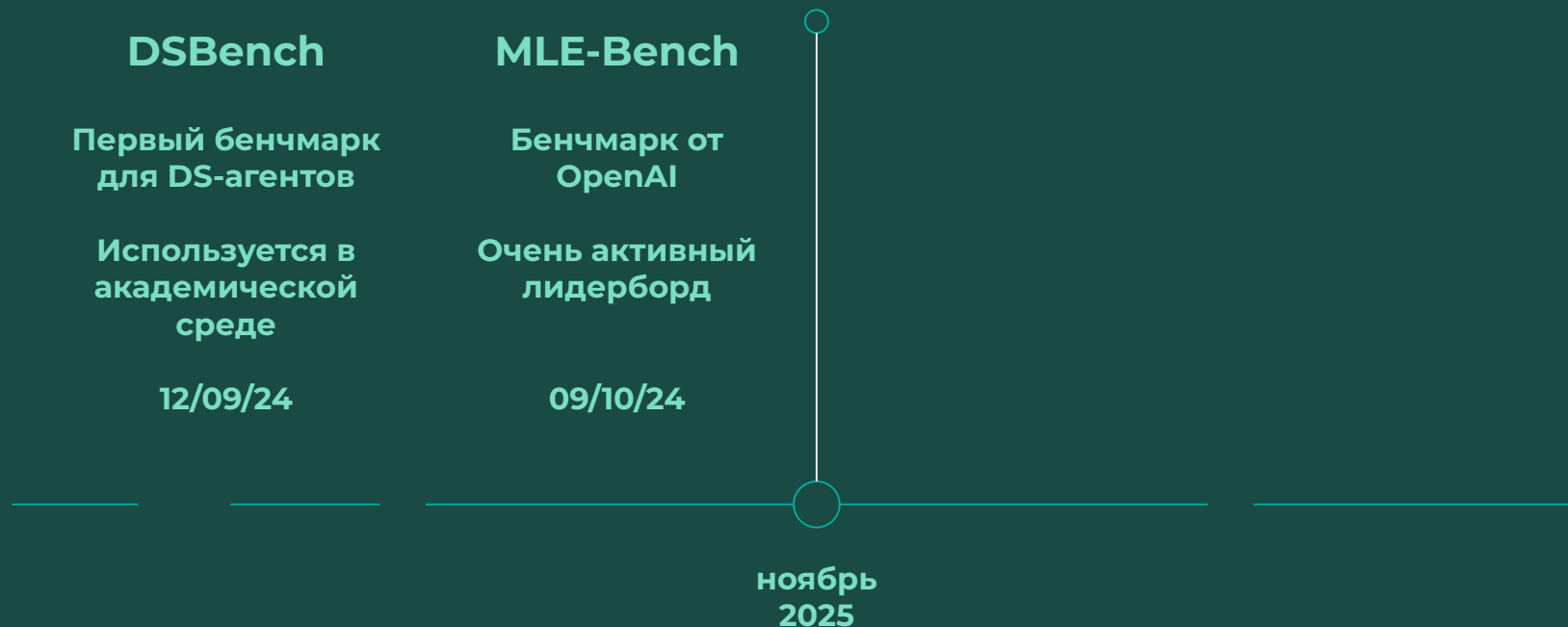
Проверить
Urgini в
Agentic
AutoML

02

Сравнить два
мира на
одинаковых
задачах

Именно отсюда начинается наш
postmortem.

Место преступления: MLE-Bench



Место преступления: MLE-Bench



Место преступления: MLE-Bench

DSBench

Первый бенчмарк
для DS-агентов

Используется в
академической
среде

12/09/24

MLE-Bench

Бенчмарк от
OpenAI

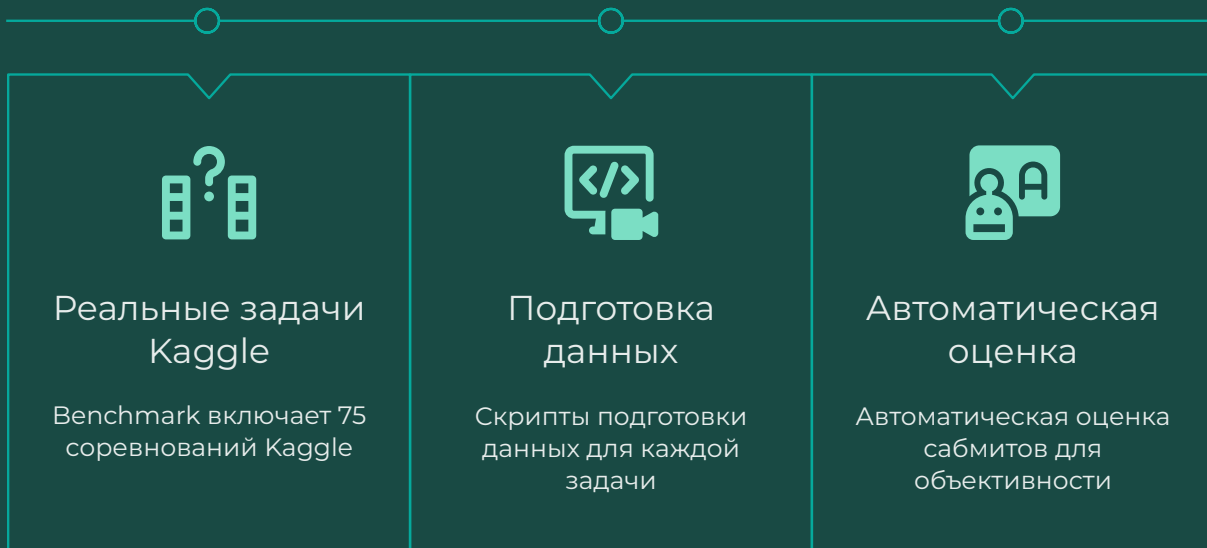
Очень активный
лидерборд

09/10/24

7 Open	75 Closed	Author
1	Add ScienceFlow MLE-Bench Results: 70.22% Overall Medal Rate	
	#151 opened on Jul 25 by pandawei-ele	
1	Support tar archive extraction in utils	
	#146 opened on May 23 by mitre88	
1	Add BaguetteFlow: New Overall SOTA with 65.3% Medal Rate	
	#145 opened on May 18 by Celseca-coder 1 task done	
1	Add Orze MLE-bench Lite results (90.91%, #1 on Lite)	
	#144 opened on May 15 by warlockee 1 task done	
1	Add AiScientist MLE-Bench Lite results	
	#142 opened on Apr 22 by survivi 1 task done	
1	Add DALPHA MLE-Bench Results	
	#129 opened on Mar 16 by BenMyoung	
1	Lazily initialize diskcache to avoid creating cache directory on import	
	#127 opened on Mar 10 by andrewjcg	

ноябрь
2025

Место преступления: MLE-Bench



Место преступления: MLE-Bench

Lite	Medium	High
22 соревнования 158 GB данных	38 соревнований 1.1 TB данных	15 соревнований 2 TB данных

Место преступления: MLE-Bench

	Lite	Medium	High
	22 соревнования 158 GB данных	38 соревнований 1.1 TB данных	15 соревнований 2 TB данных
Famou-Agent-2.0-gemini-3-pro	Gemini-3-Pro-preview, 24 hours, 64 vCPUs, 500GB of RAM, and 1 A800 GPU		

Место преступления: MLE-Bench

		Lite	Medium	High
Upgini ->	Text			
	Image			
	Audio			
	Tabular			
	Seq2Seq			

Место преступления: MLE-Bench

		Lite	Medium	High
Upgini ->	Text			
	Image			
	Audio			
	Tabular	4 соревнования	 проблемы с самими задачами	
	Seq2Seq			

Место преступления: MLE-Bench

	Lite	Medium	High
Text			
Image			
Audio	Агент лидирует = агент хорош во всех задачах?		
Tabular			
Seq2Seq			

Наш агент в расследовании

Как внедриться?



MCP

Надёжный, но не очень гибкий интерфейс



**Python
API**

Гибкость итераций при условии хорошего coding tool



Web

Auto-discovery, но велик риск ошибок

Наш агент в расследовании

Как внедриться?



MCP

Надёжный, но не очень гибкий интерфейс



**Python
API**

Гибкость итераций при условии хорошего coding tool



Web

Auto-discovery, но велик риск ошибок

Я не буду вызывать библиотеку. Я попытаюсь хакнуть ваше API и буду DDOS'ить его с неправильными параметрами сутки напролёт.

User-Agent: facebook.com

(*запрещён в РФ)

Наш агент в расследовании

Urgini + ...

Наш агент в расследовании

Upgini + ...



R&D-Agent

open source, top-1 в лидерборде (nov '25)

Наш агент в исследовании

Upgini + ...

 **R&D-Agent**

open source, top-1 в лидерборде (nov '25)

Ожидание

Попробуем
Upgini

FeaturesEnricher(...)



Research
agent



Development
agent

Реальность

Upgini?...

how code



Research
agent



Development
agent

Наш агент в исследовании

Upgini + ...

 R&D-Agent

open source, top-1 в лидерборде (nov '25)

Ожидание

Какой RMSLE?

0.059



Research
agent



Development
agent

Реальность

Какой RMSLE?

0.17
нет, 0.27
нет, 0.65



Research
agent



Development
agent

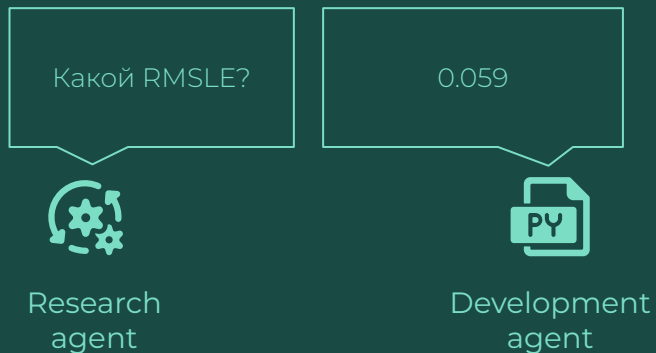
Наш агент в расследовании

Upgini + ...

 **R&D-Agent**

open source, top-1 в лидерборде (nov '25)

Ожидание



Реальность



по крайней мере на 32 CPU 128GB

Наш агент в расследовании

Upgini + ...



AutoGluon

autogluon-assistant
aka **MLZero**

заточен под использование библиотек

Наш агент в расследовании

Urgini + ...



AutoGluon

autogluon-assistant
aka **MLZero**

заточен под использование библиотек

Ожидание

Дайте подумать,
возьмём X, Y, Z



Tool selector

```
use_X()  
use_Y()  
use_Z()
```



Coder

Реальность

Just use
AutoGluon



Tool selector

Согласен, хватит 1
итерации



Coder

Первые результаты Tabular Lite

Агент	Any Medal %
Thesis	50.0 $\pm$ 0.0
R&D-Agent	50.0 $\pm$ 0.0
Upgini+MLZero	50.0 $\pm$ 0.0
...	50.0 $\pm$ 0.0
AIDE	45.8 $\pm$ 2.8

Результат усреднения
3х прогонов

Первые результаты Tabular Lite

Агент	Any Medal %
Thesis	50.0 $\pm$ 0.0
R&D-Agent	50.0 $\pm$ 0.0
Upgini+MLZero	50.0 $\pm$ 0.0
...	50.0 $\pm$ 0.0
AIDE	45.8 $\pm$ 2.8

Причина смерти №1:
неинформативная
метрика

Any Medal % = в сколько
соревнованиях агент получил
любую медаль

границы **фиксированы**

никак не ранжирует внутри
задач!

Меняем метрику

```
mean(  
    if score ≥ BRONZE then 1  
)
```

Any Medal % = в сколько
соревнованиях агент получил
любую медаль

никак не ранжирует внутри
задач



```
mean(  
    (sample_submission_score - score) /  
    (sample_submission_score -  
    gold_medal_score)  
)
```

Normalized score = насколько
агент близок к лучшему
человеческому решению

учитывает лидерборд по
отдельным задачам

Обновлённые результаты

Агент	Any Medal % 	Normalized score
Thesis	50.0 $\pm$ 0.0	0.891 $\pm$ 0.15
R&D-Agent	50.0 $\pm$ 0.0	0.497 $\pm$ 0.574
Upgini+MLZero	50.0 $\pm$ 0.0	0.927 $\pm$ 0.086
...	50.0 $\pm$ 0.0	...
AIDE	45.8 $\pm$ 2.8	0.776 $\pm$ 0.242

Any Medal % = в сколько соревнованиях агент получил любую медаль

Normalized score = насколько агент близок к лучшему человеческому решению

Обновлённые результаты

Агент	Any Medal %	Normalized score 
Upgini+MLZero	50.0 $\pm$ 0.0	0.927 $\pm$ 0.086
Thesis	50.0 $\pm$ 0.0	0.891 $\pm$ 0.15
AIDE	45.8 $\pm$ 2.8	0.776 $\pm$ 0.242
R&D-Agent	50.0 $\pm$ 0.0	0.497 $\pm$ 0.574
...	50.0 $\pm$ 0.0	...

Any Medal % = Автономность

Normalized score = Качество

Обновлённые результаты

Агент	Any Medal %	Normalized score ↓
Upgini+MLZero	50.0 ± 0.0	0.927 ± 0.086
Thesis	50.0 ± 0.0	0.891 ± 0.15
AIDE	45.8 ± 2.8	0.776 ± 0.242
R&D-Agent	50.0 ± 0.0	0.497 ± 0.574
...	50.0 ± 0.0	...

Причина смерти №2:
нестабильный
результат

Нестабильность агента
проявляется не только между
запусками, но и между
задачами

Обновлённые результаты

Причина смерти №3:
ресурсы

Агент	Any Medal %	Normalized score ↓	Ресурсы на 1 задачу	Время на 1 задачу
Upgini+MLZero	50.0 ± 0.0	0.927 ± 0.086	32 vCPU, 128GB RAM	2 часа
Thesis	50.0 ± 0.0	0.891 ± 0.15	24 vCPU, 170GB RAM, 1 H100 GPU	24 часа
AIDE	45.8 ± 2.8	0.776 ± 0.242	6 vCPU, 440 GB RAM, 1 NVIDIA A10 GPU	24 часа
R&D-Agent	50.0 ± 0.0	0.497 ± 0.574	12 vCPU, 220GB RAM, 1 V100 GPU	12 часов
...	50.0 ± 0.0	...	...	...

AutoGluon и ресурсы

Причина смерти №3:
ресурсы

Ресурсы	Время	mean RMSE
32 vCPU, 128GB RAM	1.5 часа	4.09477
1 vCPU, 12GB RAM	1.5 часа	4.27443

Upgini+AutoGluon на
new-york-city-taxi-fare-prediction

Новая линейка – новый лидерборд

Хотите PR с более
точным
лидербордом?

Upgini

Мы не готовы это
поддерживать на
нашей стороне

OpenAI

Хорошо, мы сделаем
свой форк

Upgini



github.com/upgini/mle-bench



huggingface.co/spaces/Upgini/mle-bench-tabular

Причина смерти №0: test set

Февраль 2026

Disarray: Примите наши результаты, они супер-хороши!

OpenAI: Погодите, вы же использовали test для обучения.

OpenAI: Вас надо в отдельную таблицу.

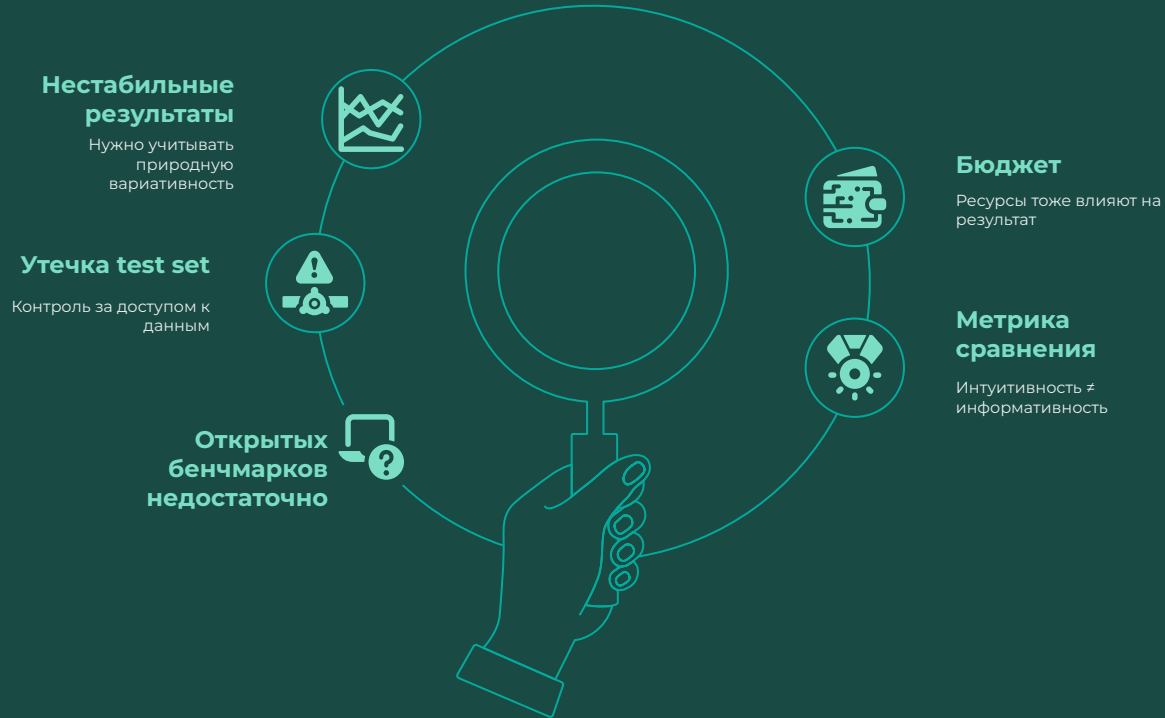
Наблюдатель: У вас пол-лидерборда с закрытым кодом, какая отдельная таблица.

[OpenAI закрыл обновление бенчмарка]

Kaggle: есть private test set для оценки, и он **недоступен**

MLE-Bench: сам делает train и test из public train; **test set может быть использован агентом**

Итоги расследования



Чеклист: что выбрать?

- **Classic AutoML, если:**

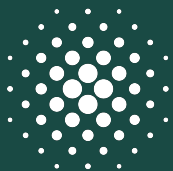
- мало ресурсов
 - нужна устойчивость
 - нужно меньше точек отказа

- **Agentic AutoML, если:**

- достаточно ресурсов
 - готовы к экспериментам
 - нужен нестандартный процессинг данных
 - нужна генерация гипотез
 - есть выверенная метрика сравнения



Не спрашивайте: «Нужен ли нам AI-агент?» Сначала спросите: «Что агент должен уметь делать такого, чего не делает классический AutoML?»



Спасибо за внимание!

Валерия Дымбицкая

tg @Suncelesta

upgini.com