

What Can Transformers Learn In-Context? A Case Study of Simple Function Classes

in context learning

Как работает ICL

- 1) Берется предобученная LLM
- 2) Берется промпт состоящий из примеров решения задачи и тестового примера
- 3) Из LLM получаем решение тестового примера

Как работает ICL

Input: 2014-06-01

Output: !06!01!2014!

Input: 2007-12-13

Output: !12!13!2007!

Input: 2010-09-23

Output: !09!23!2010!

*in-context
examples*

Input: **2005-07-23**

test example

Output: **!07!23!2005!**

!07!23!2005!
| _ _ *model completion*

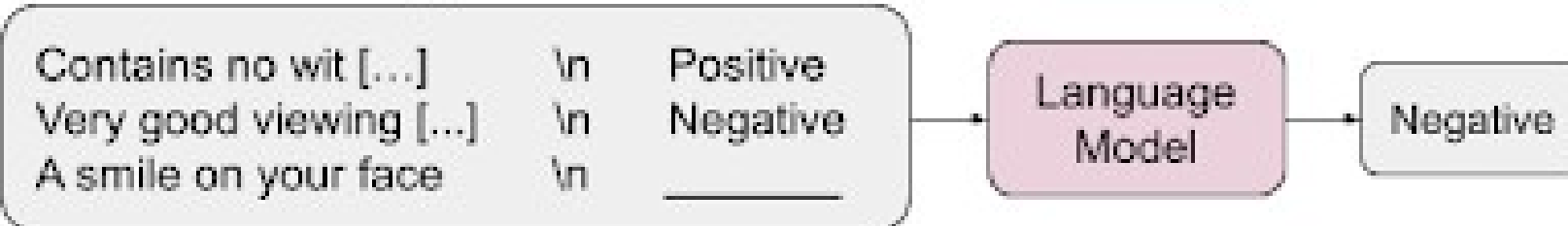
Подходы к ICL

Regular In-Context Learning



Natural language targets: {Positive/Negative} sentiment

Flipped-Label In-Context Learning



Flipped natural language targets: {Negative/Positive} sentiment

Semantically-Unrelated Label In-Context Learning



Semantically-unrelated targets: {Foo/Bar}, {Apple/Orange}, {A/B}

Chain of thought

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain-of-Thought Prompting


Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

ICL подход из статьи

1)Берется трансформер Decoder Only

2)Семплируются x из D_x и f из D_f

3)Обучается с нуля на $(x_1, f(x_1), \dots, x_k, f(x_k), x_{\text{query}}) \rightarrow f(x_{\text{query}})$

Model	Embedding size	#Layers	#Heads	(Total parameters)
Tiny	64	3	2	0.2M
Small	128	6	4	1.2M
Standard	256	12	8	9.5M

ICL на линейных функциях

1) x из нормального распределения

2) $f(x) = wx$

3) w из нормального распределения

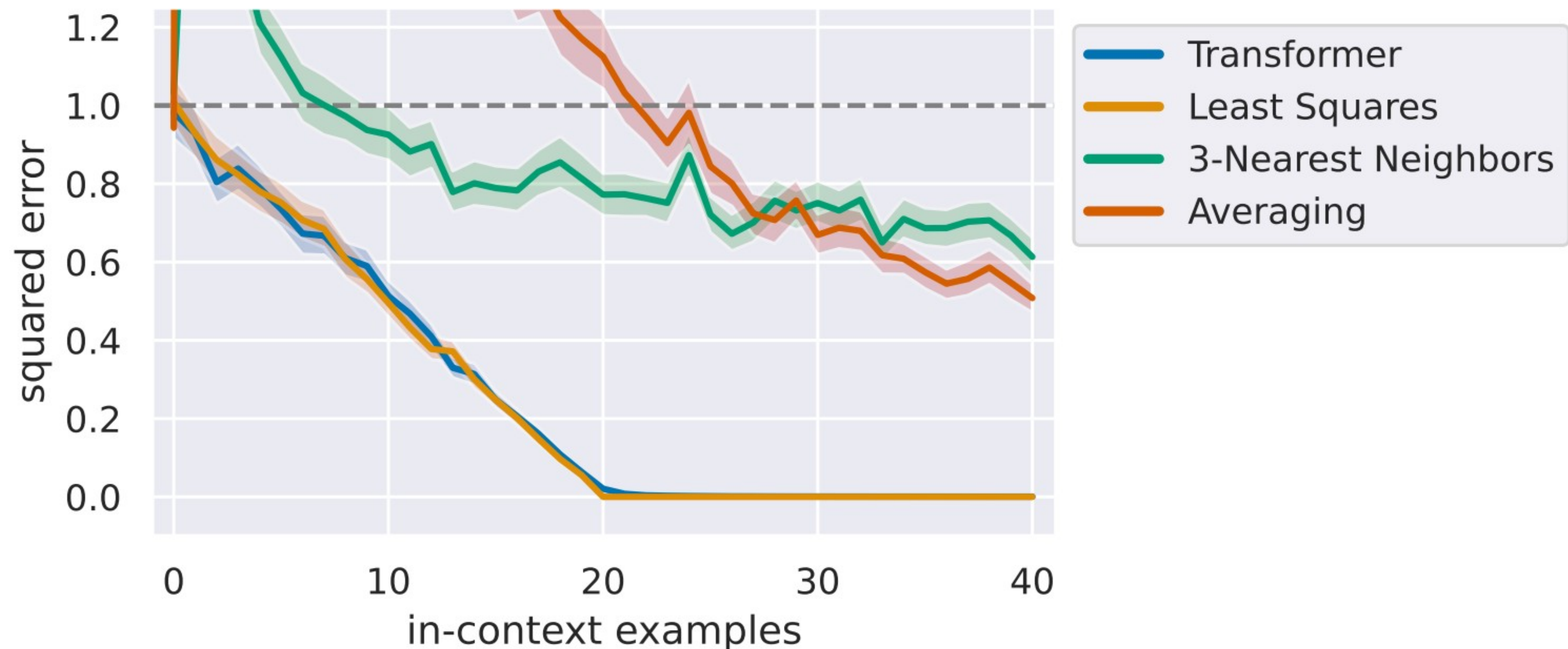
4) Out-of-distribution:

Добавление шума в $f(x)$ при инференсе

Взятие на инференсе x извне распределения

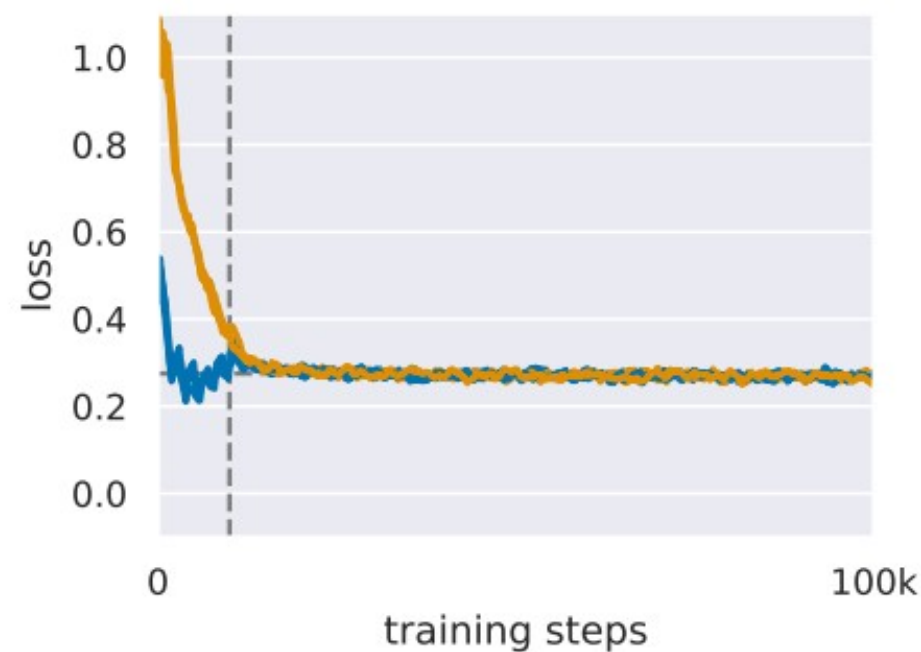
Результаты ICL на линейных функциях

Трансформер достигает минимальной ошибки 0.02 при d примерах и падает до 0.0006 при $2d$ примерах

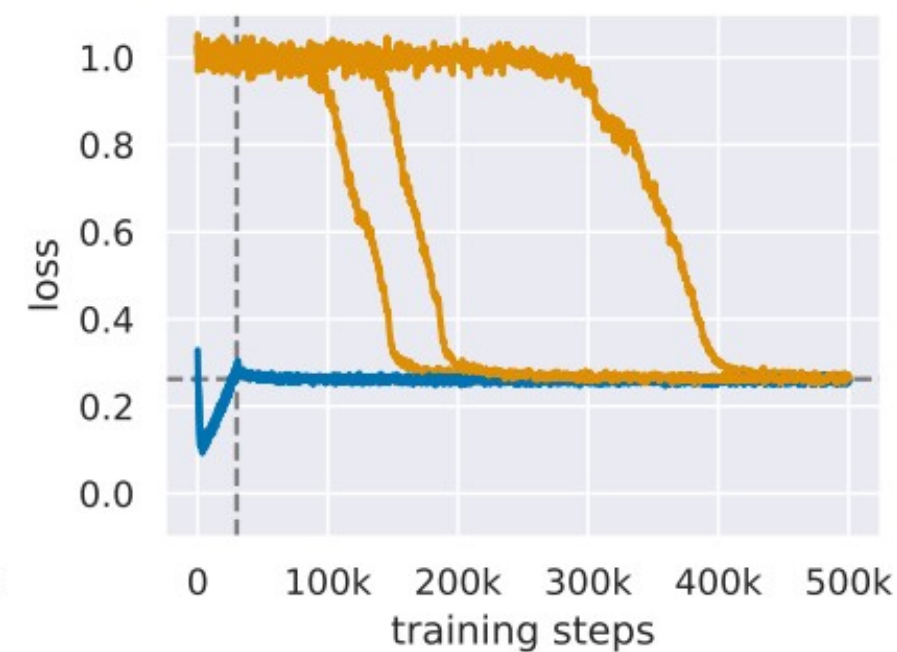


Результаты ICL на линейных функциях

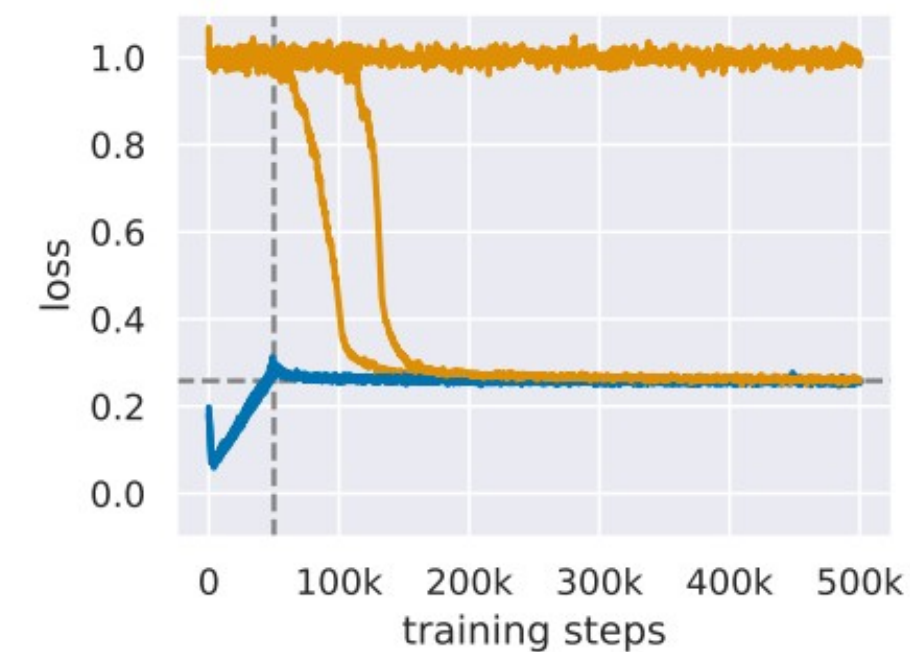
Трансформер достигает лучших результатов при использовании curriculum learning



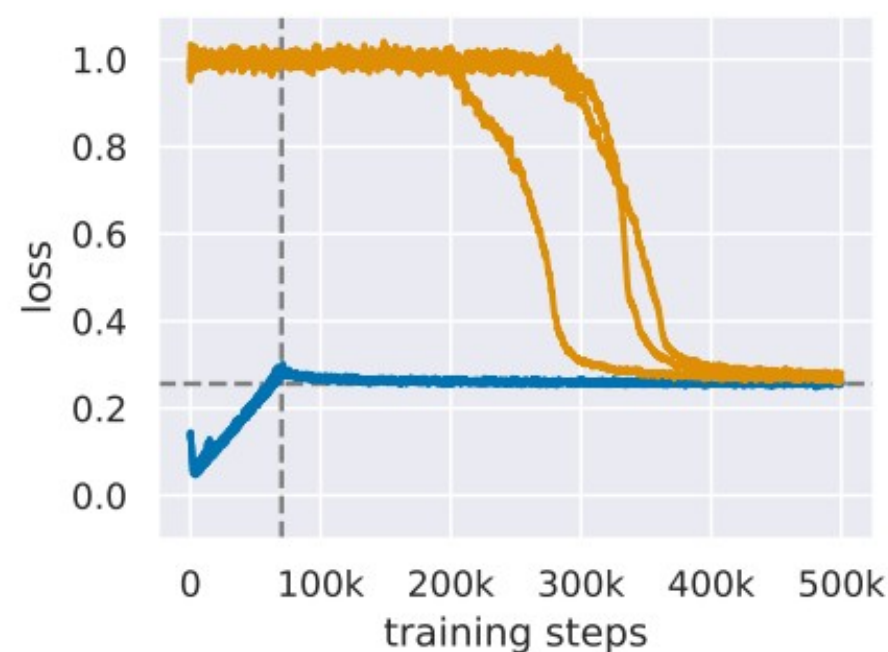
(a) 10 dimensions



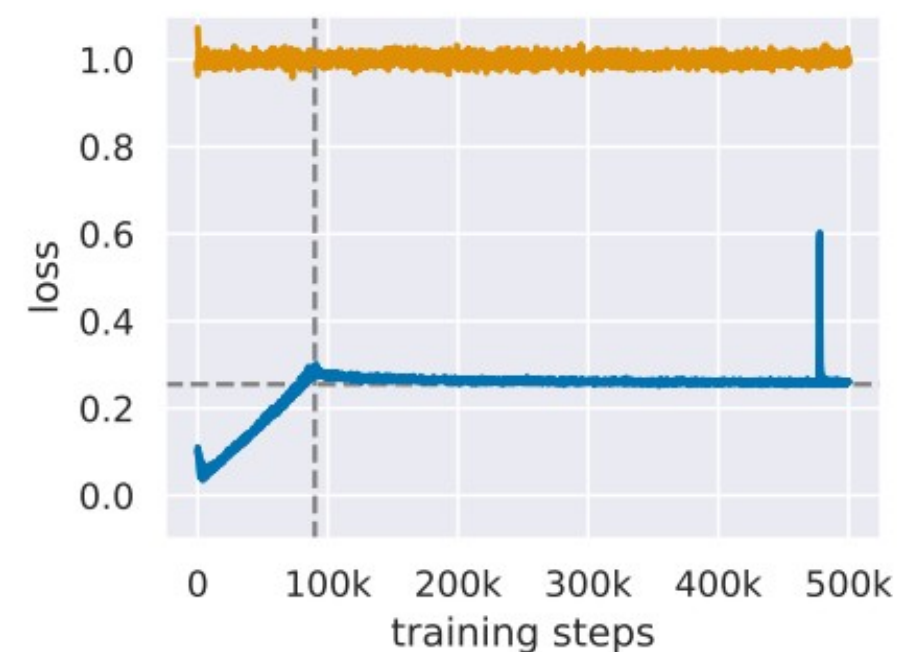
(b) 20 dimensions



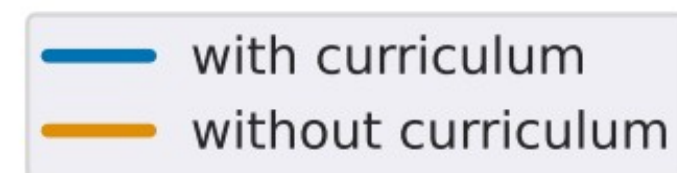
(c) 30 dimensions



(d) 40 dimensions



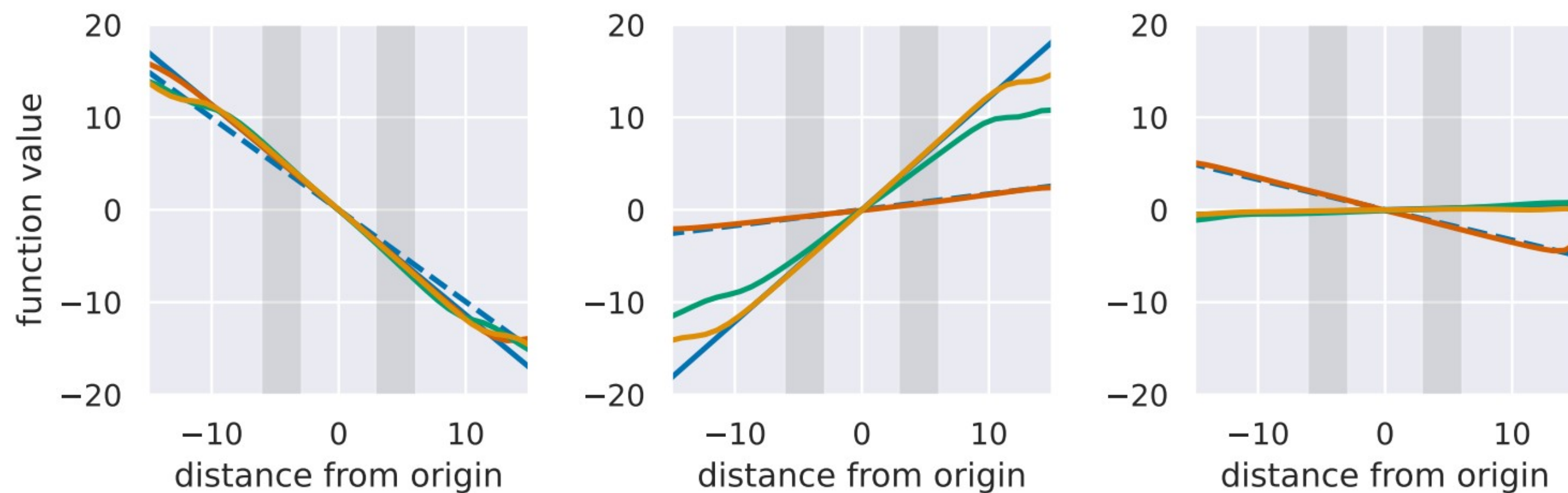
(e) 50 dimensions



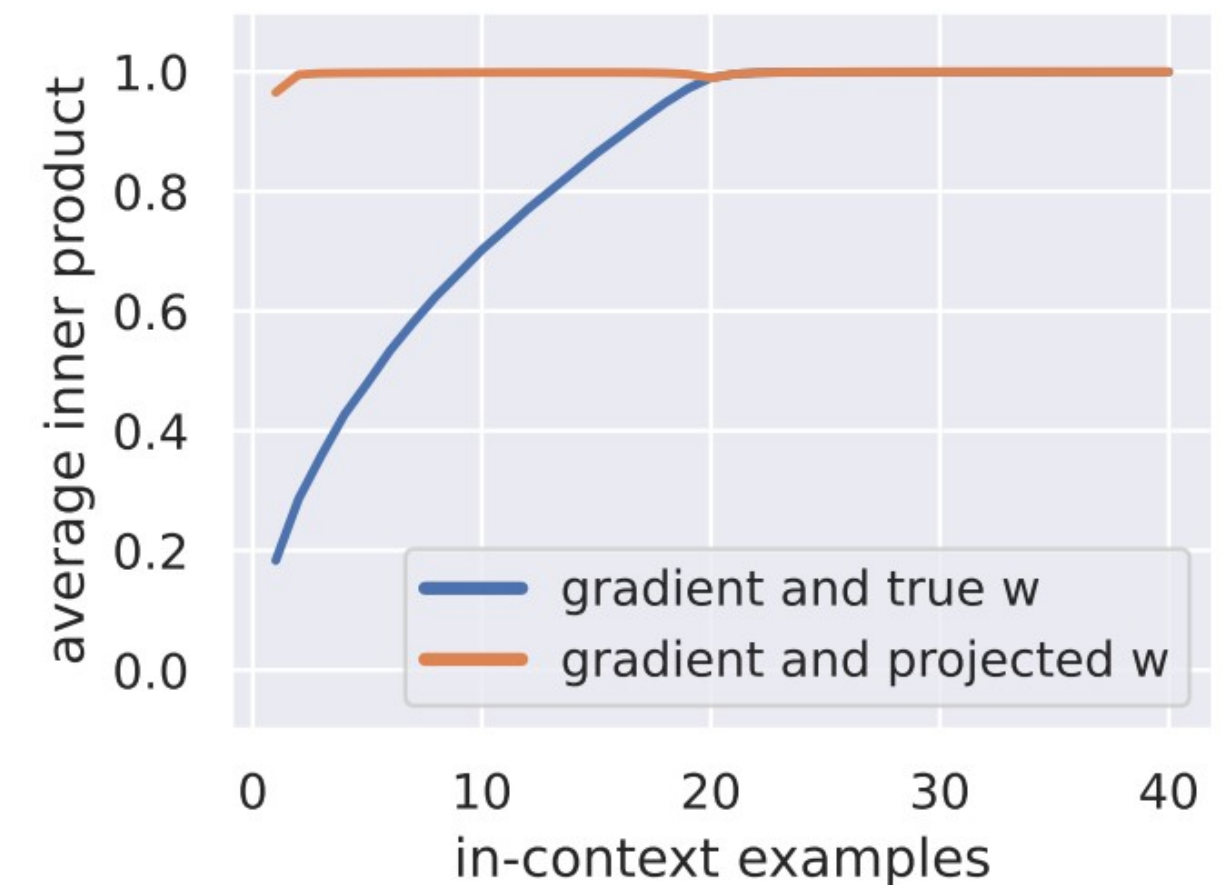
Результаты ISL на линейных функциях

При количестве примеров меньше чем d точного решения не получить, но хочется получить лучшее приближительное

— ground truth — #dims / 2 in-context examples — #dims * 2 in-context examples
- - ground truth projected — #dims in-context examples



(a) function visualizations



(b) gradients

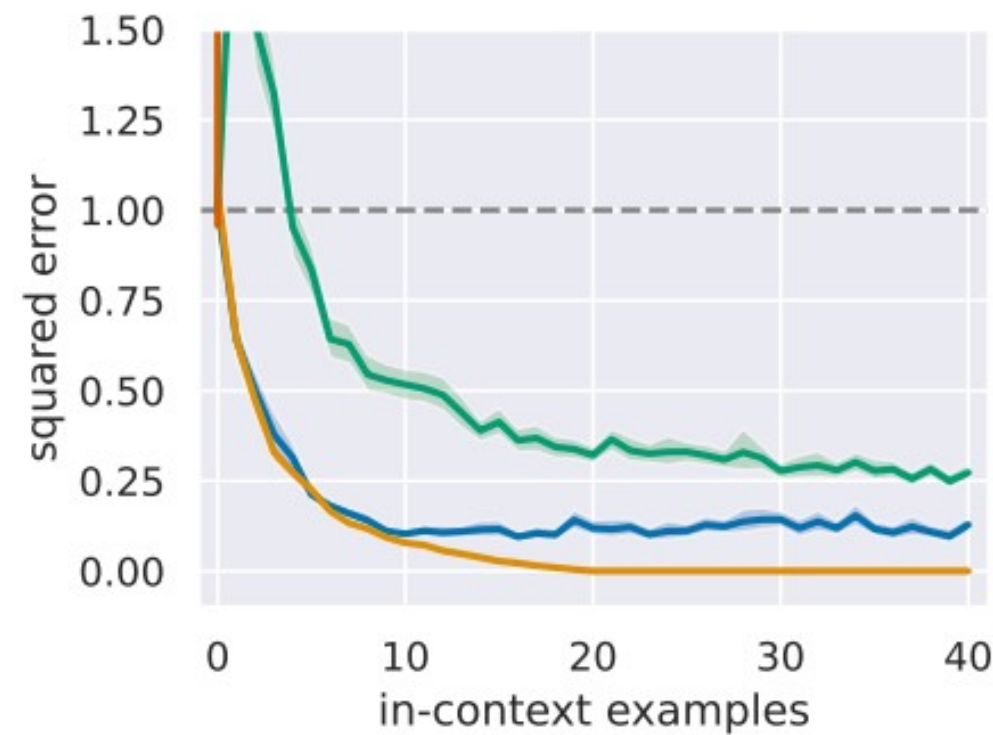
Результаты ICL на линейных функциях

Out of distribution

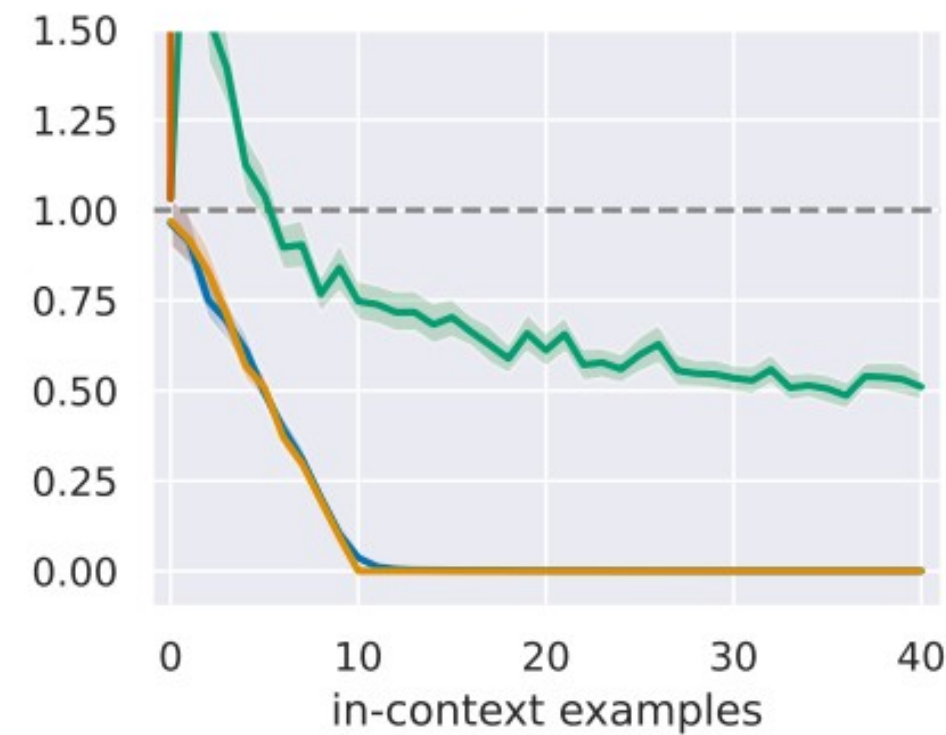
Prompting strategy	$D_{\mathcal{X}}^{\text{train}} \neq D_{\mathcal{X}}^{\text{test}}$	$D_{\mathcal{F}}^{\text{train}} \neq D_{\mathcal{F}}^{\text{test}}$	$D_{\text{query}}^{\text{test}} \neq D_{\mathcal{X}}^{\text{test}}$
Skewed covariance	✓		
$d/2$ -dimensional subspace	✓		
Scale inputs	✓		
Noisy output		✓	
Scale weights		✓	
Different Orthants	✓		✓
Orthogonal query			✓
Query matches example			✓

Результаты ISL на линейных функциях

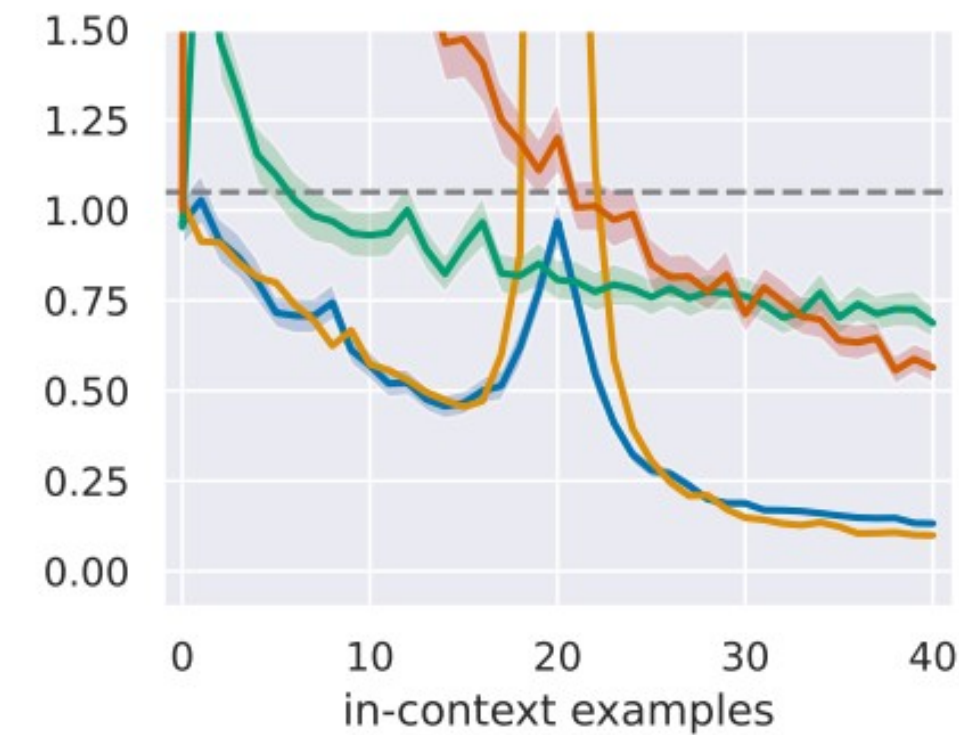
Out of distribution результаты



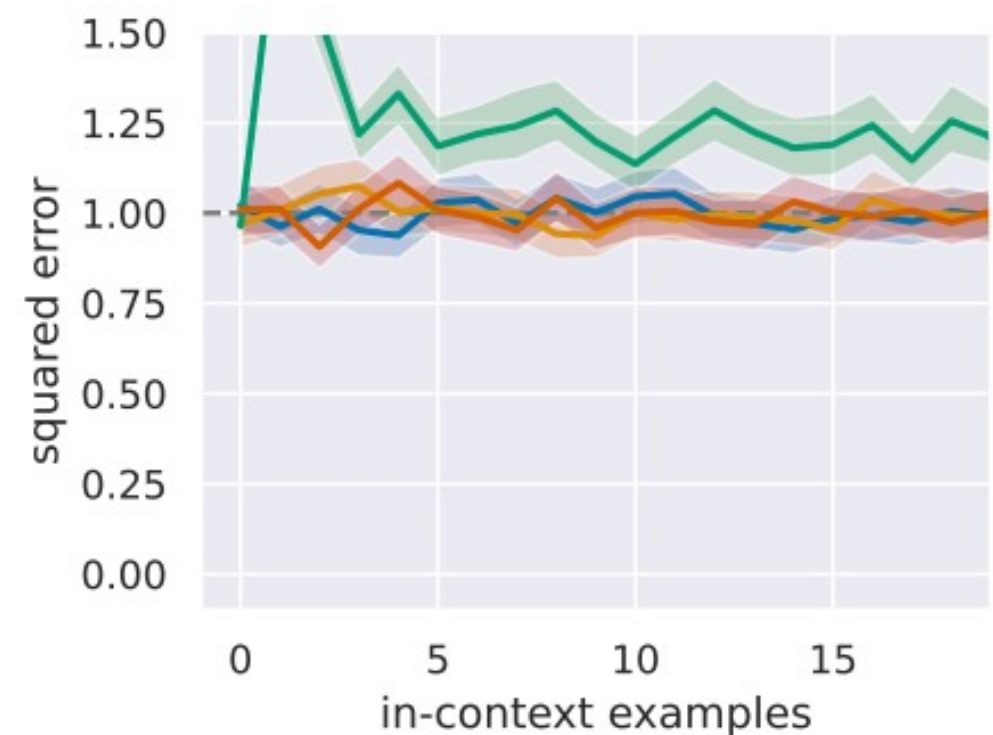
(a) skewed covariance



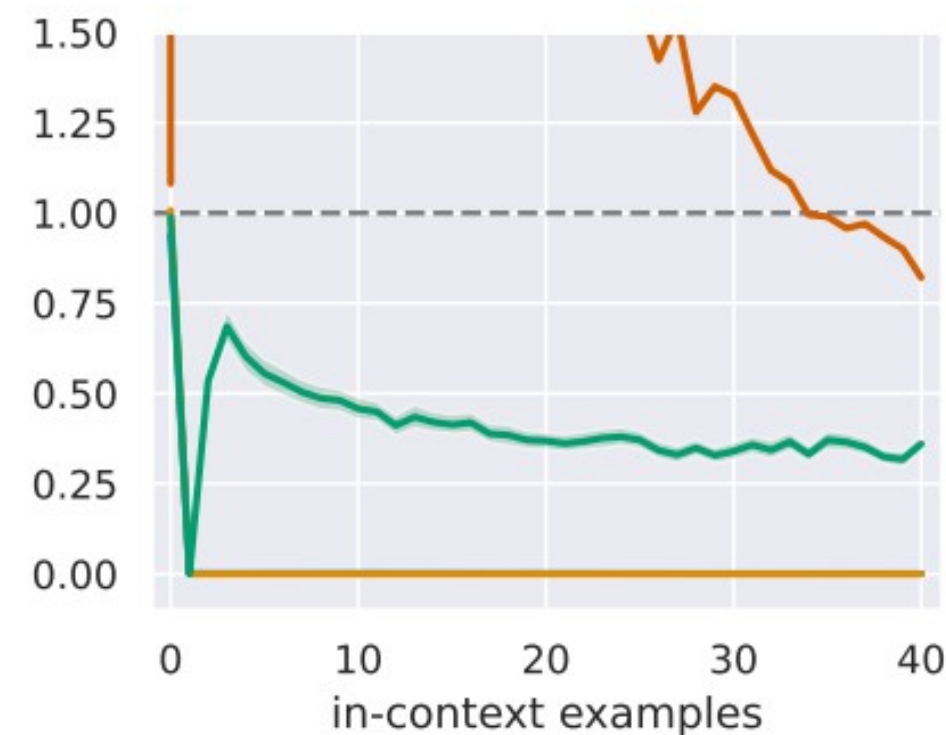
(b) $d/2$ -dimensional subspace



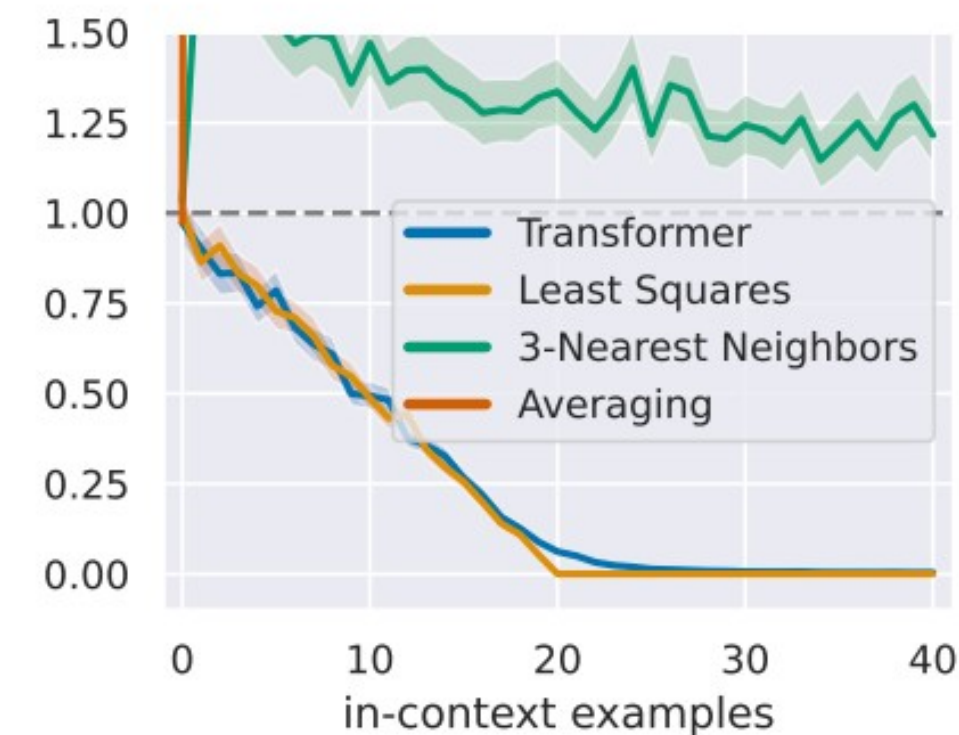
(c) noisy output



(d) orthogonal query



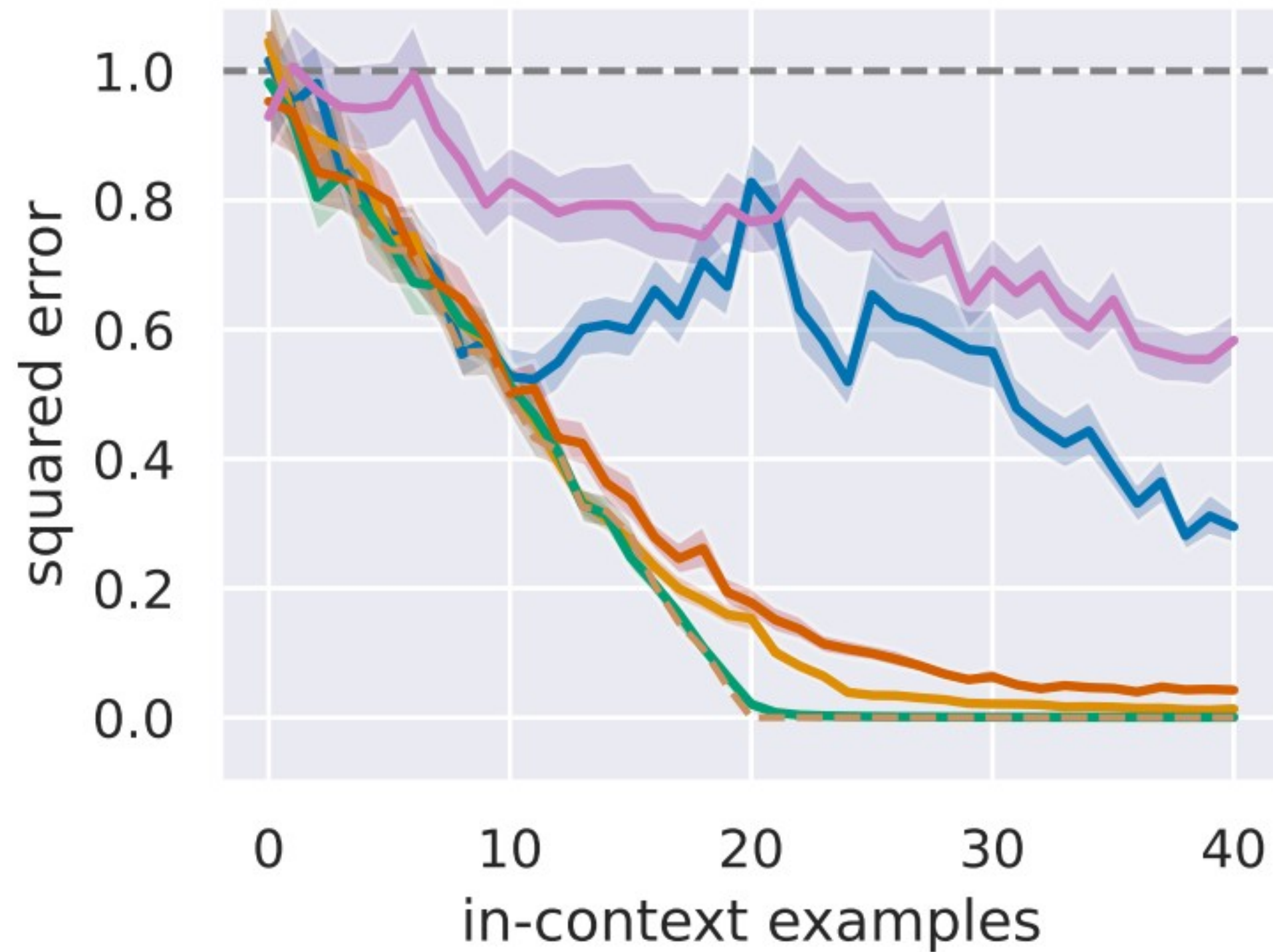
(e) query matches in-context example



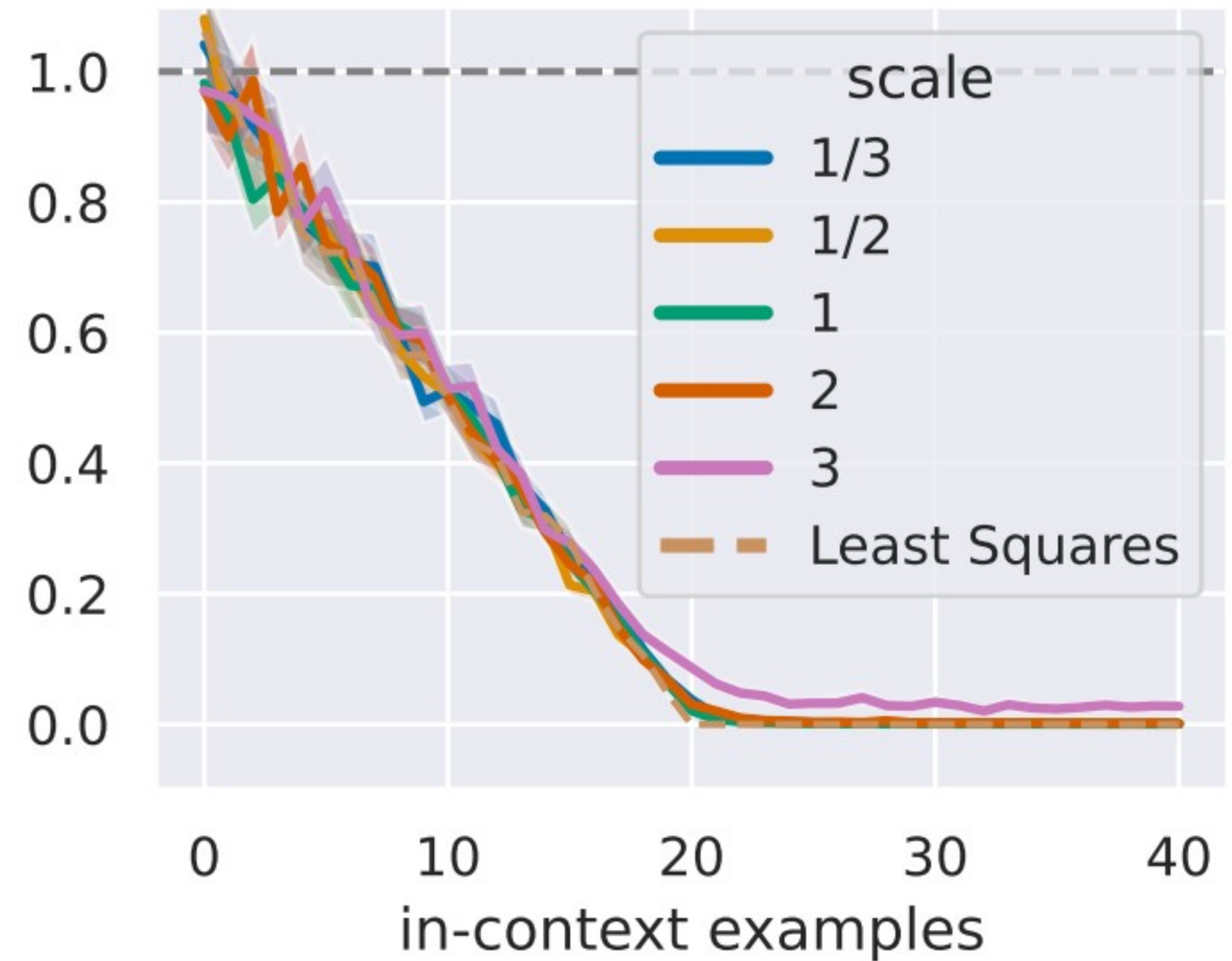
(f) different orthants

Результаты ICL на линейных функциях

Out of distribution при скейле



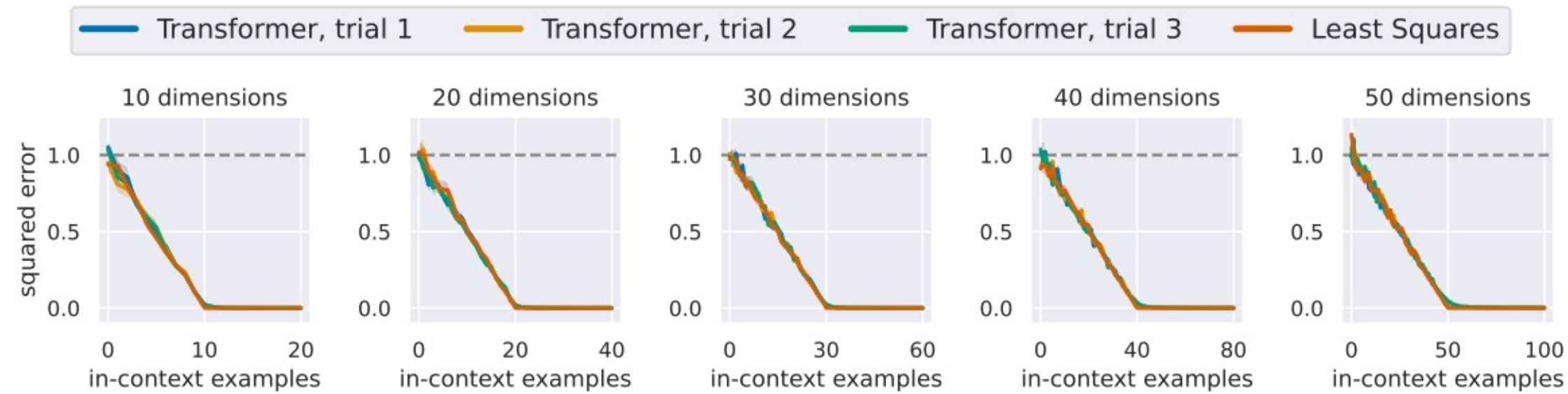
(a) scaled x , Transformer



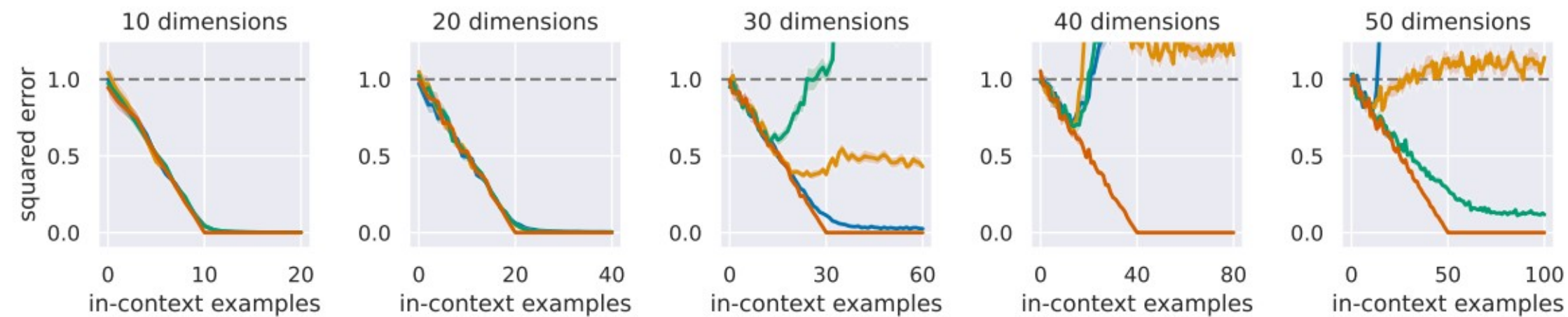
(b) scaled w , Transformer

Результаты ISL на линейных функциях

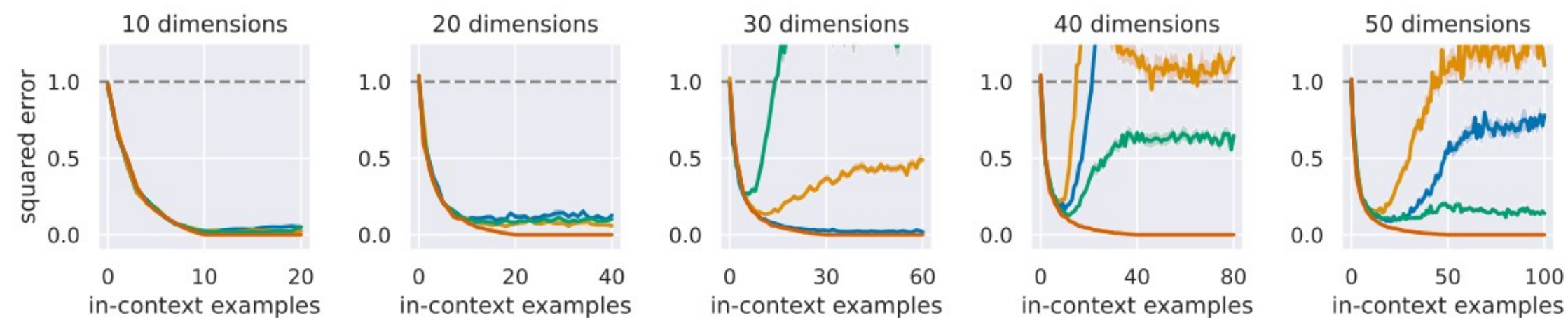
Out of distribution при увеличении размерностей



(a) standard



(b) different orthants



(c) skewed covariance

ICL на более сложных функциях

Разреженные матрицы:

- x из нормального распределения
- $f(x) = wx$
- В w ровно s ненулевых координат из нормального распределения

Деревья решений:

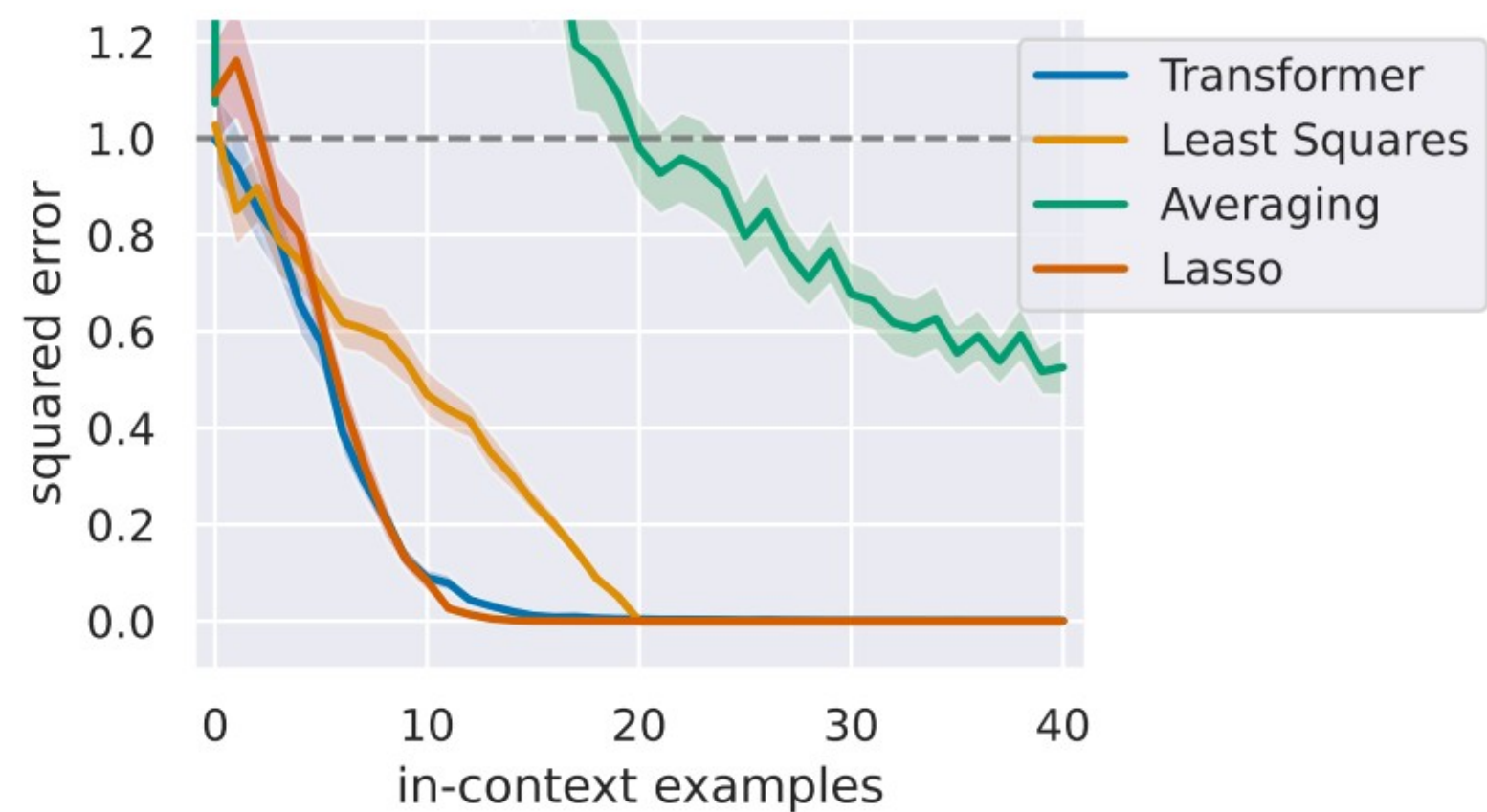
- x из нормального распределения
- f = полное решающее дерево глубины 4
- Предикаты — на знак случайных переменных
- Во всех листьях значения таргета из нормального распределения

NN:

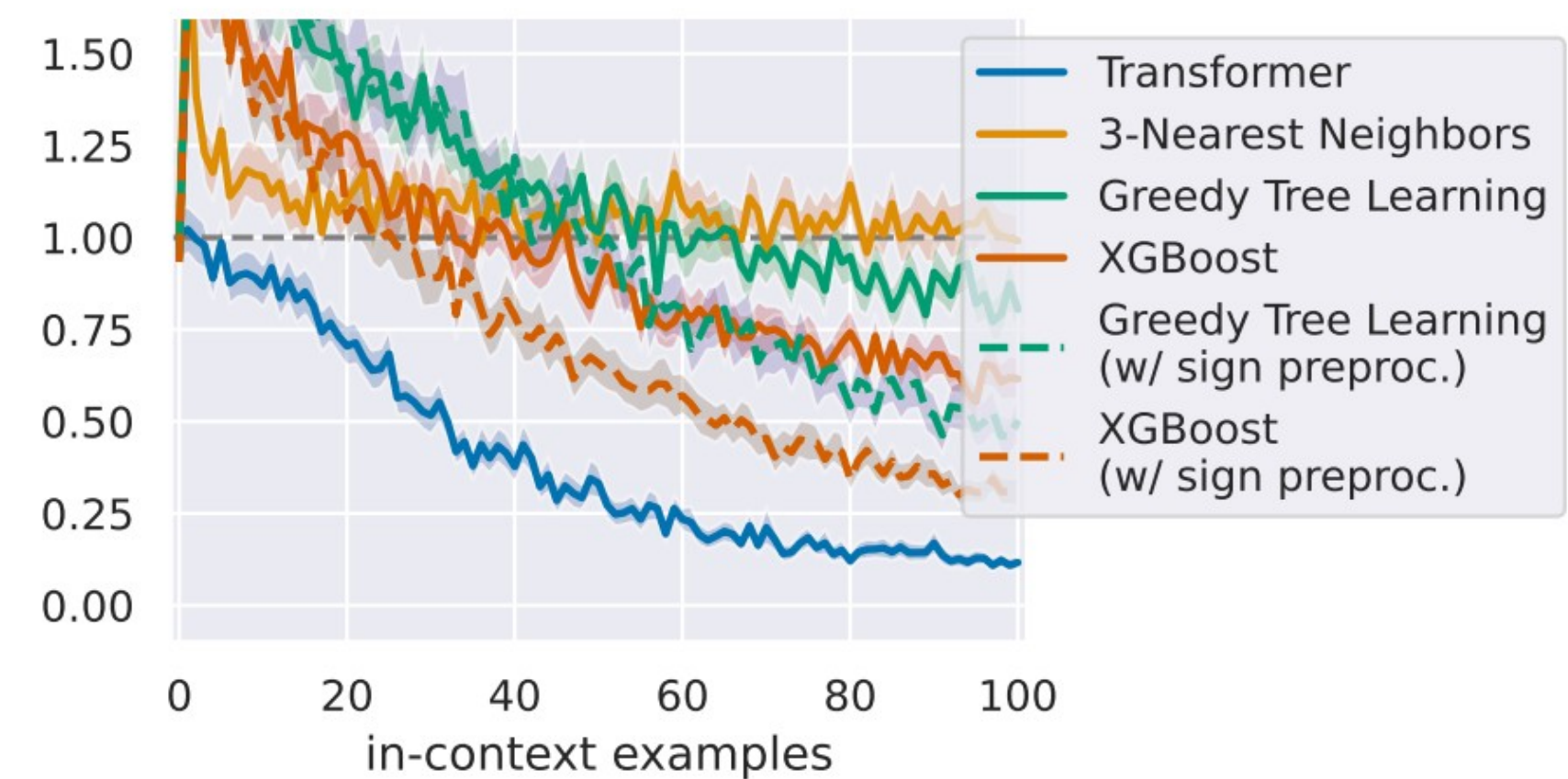
- 2 линейных слоя с relu
- Веса семплируются из нормального распределения

Результаты ICL на более сложных функциях

Результаты:



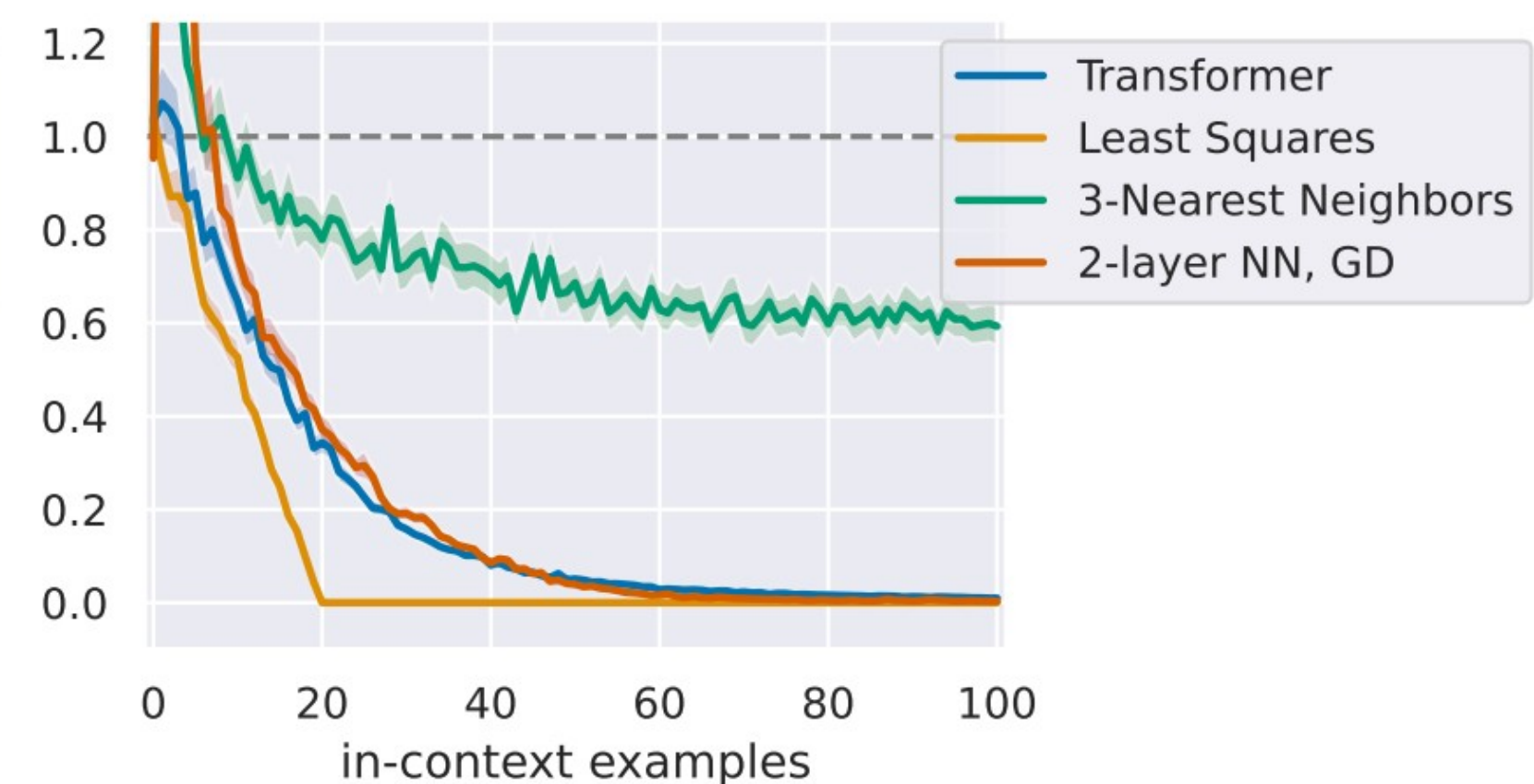
(a) Sparse linear functions



(b) Decision trees



(c) 2-layer NN



(d) 2-layer NN, eval on linear functions