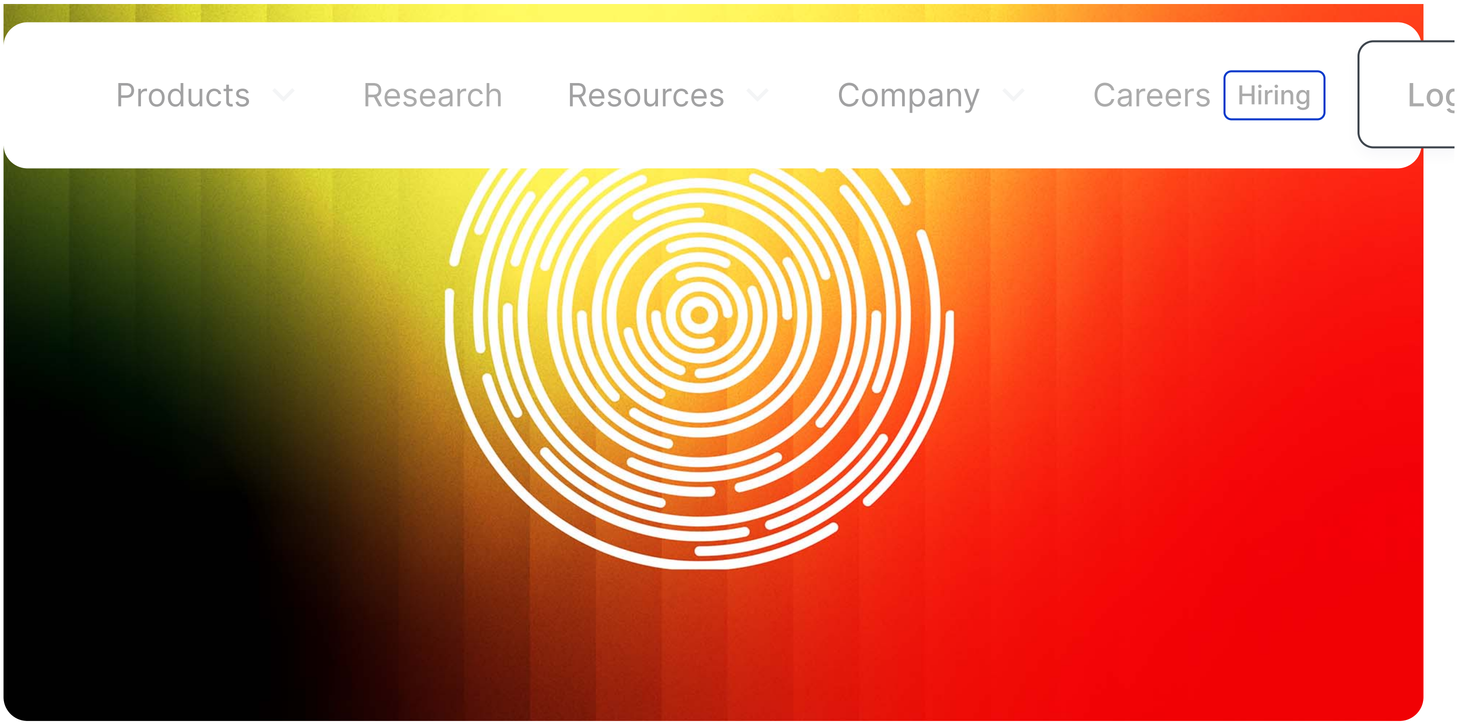
‹ Back to Articles

What is In-context Learning, and how does it work: The Beginner’s Guide

Large Language Models 10min read August 27, 2025

Deval Shah



What is In-context Learning, and how does it work: The Beginner’s Guide

Language models, specifically large language models (LLMs), have significantly advanced the field of natural language processing (NLP). The primary objective of LLMs is to model the generative likelihood of word sequences, enabling the prediction of subsequent tokens. The scalability of LLMs, in terms of training, computing, and model parameters, has been instrumental in enhancing performance across various NLP tasks.

An interesting way of utilizing LLMs post-training is the In-context learning (ICL) approach. Without any gradient update, the model learns to address a new task during inference by receiving a prompt, including task examples. This spans applications from generating contextually relevant text.

In this short article, we'll explore:

1. What is in-context learning (ICL)?
2. How does in-context learning work in LLMs?
3. In-context learning real-life applications
4. Challenges, limitations, and ICL research

Let's dive in.

The Lakera team has accelerated Dropbox's GenAI journey.

> "Dropbox uses Lakera Guard as a security solution to help safeguard our LLM-powered applications, secure and protect user data, and uphold the reliability and trustworthiness of our intelligent features."

If you're interested in digging deeper into how LLMs learn and respond, here are a few more reads worth checking out:

Learn how fine-tuning compares to in-context learning in this guide to LLM fine-tuning best practices.

post breaks down hallucinations and how to

improve prompt effectiveness.

Discover how prompt injections can exploit in-context learning in this essential guide to prompt attacks.

Take a closer look at training data poisoning and how it undermines learning signals in both fine-tuned and in-context settings.

For a systems-level view, explore LLM monitoring and how to track behavior across learning setups.

And if you're evaluating different LLMs for production, this comprehensive LLM comparison is a good place to start.

What is in-context learning (ICL)?

Traditional machine learning models were primarily designed to tackle specific tasks based on their training data. Their capabilities were bound by the input-output pairs they were trained on, and any deviation from this would lead to suboptimal results. However, with the emergence of LLMs, a paradigm shift occurred in how we solved natural language tasks.

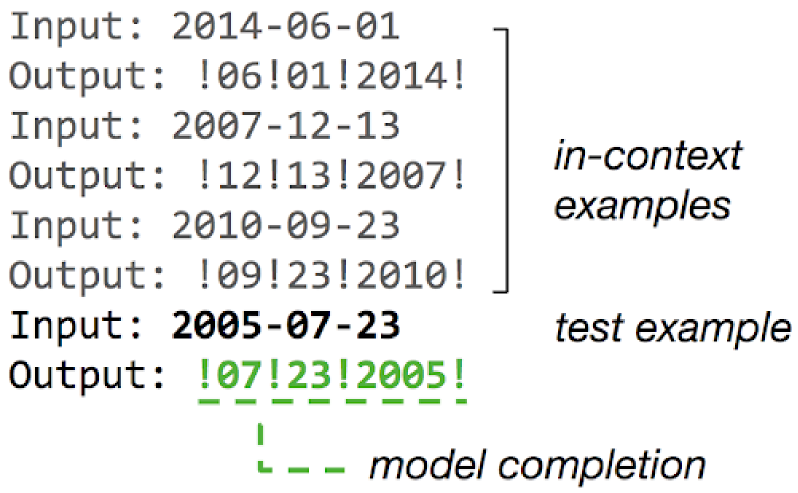
In-context learning (ICL) is a technique where task demonstrations are integrated into the prompt in a natural language format. This approach allows pre-trained LLMs to address new tasks without fine-tuning the model.

parameters. ICL operates without updating these

and generates accurate predictions accordingly.

In-context learning (ICL) is known as *few-shot learning* or *few-shot prompting*. Contrary to conventional models, the knowledge accumulated via this method is transient; post-inference, the LLM does not persistently store this information, ensuring the stability of model parameters.

ICL's efficacy is attributed to its capacity to exploit the extensive pre-training data and the expansive model scale inherent to LLMs. This allows LLMs to comprehend and execute novel tasks without a comprehensive training process of preceding machine learning architectures.



```
Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: 2005-07-23
Output: !07!23!2005!
```

Source

💡 **Pro tip:** Fine-tuning is a crucial step in maximizing the potential of LLMs. Dive into Lakera's comprehensive LLM Fine-Tuning Guide to understand the nuances and best practices for optimal results.

interface for interaction with large language models(LLMs). This framework simplifies the integration of human expertise into LLMs by modifying the sample cases and templates.

ICL's approach mirrors the human cognitive reasoning process, making it a more intuitive model for problem-solving.

Computational overhead for task-specific model adaptation is significantly less and paves the way for deploying language models as a service, facilitating their application in real-world scenarios.

ICL demonstrates competitive performance across various NLP benchmarks, even when compared with models trained on a more extensive labeled data set.

How does in-context learning work in LLMs?

The key idea behind in-context learning is to learn from analogy, a principle that enables the model to generalize from a few input-output examples or even a single example. In this approach, a task description or a set of examples is formulated in natural language and presented as a "prompt" to the model. This prompt is a semantic prior, guiding the model's chain of thought and subsequent output. Unlike traditional machine learning methods like linear regression, which requires labeled data and a separate training process, in-context learning operates on pre-trained models and does not involve any parameter updates.

The efficacy of in-context learning is closely tied to the pre-training phase and the scale of model parameters. Research indicates that the model's ability to perform in-context learning improves as the prior knowledge from the training data, which later

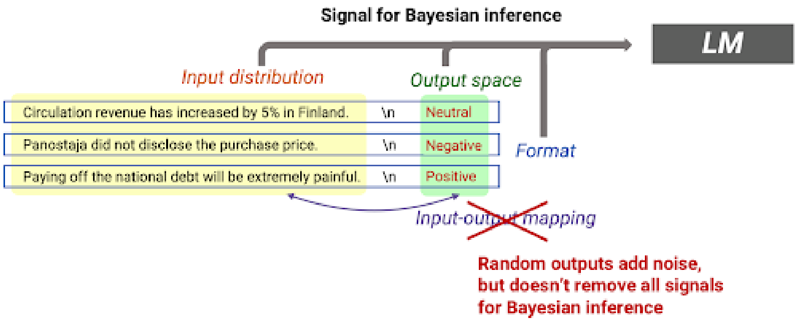
complex tasks with minimal additional input.

In-context learning is often employed in a few-shot learning scenario, where the model is provided with a few examples to understand the task at hand. The art of crafting effective prompts for few-shot learning is known as prompt engineering, and it plays a crucial role in leveraging the model's in-context learning capabilities.

Bayesian Inference Framework

The Stanford AI Lab blog introduces a Bayesian inference framework to understand in-context learning in large language models like GPT-3. The framework suggests that in-context learning is an emergent behavior where the model performs tasks by conditioning on input-output examples without optimizing any parameters. The model uses the prompt to "locate" latent concepts acquired during pre-training. This differs from traditional machine learning algorithms that rely on backpropagation for parameter updates. The Bayesian inference framework provides a mathematical foundation for understanding how the model sharpens the posterior distribution over concepts based on the prompt, effectively "learning" the concept.

It emphasizes the role of latent concept variables containing various document-level statistics. These latent concepts create long-term coherence in the text and are crucial for the emergence of in-context learning. The model learns to infer these latent concepts during pre-training, which later aids in in-context learning. This aligns with the notion that pre-training data is the foundation for in-context learning, Despite this, the model can successfully perform in-context learning if the signal exceeds the noise.


Signal for Bayesian inference
LM
Input distribution
Output space
Circulation revenue has increased by 5% in Finland.
\n
Neutral
Panostaja did not disclose the purchase price.
\n
Negative
Paying off the national debt will be extremely painful.
\n
Positive
Format
Input-output mapping
Random outputs add noise,
but doesn't remove all signals
for Bayesian inference

Source

Interestingly, in-context learning (ICL) is robust to output randomization. Unlike traditional supervised learning, which would fail if the input-output mapping information is removed, in-context learning still performs well. This suggests that other prompt components, such as input and output distribution, provide sufficient evidence for Bayesian inference.

Approaches for In-Context Learning

One of the key aspects of in-context learning is its flexibility in the number of examples required for task adaptation. Specifically, there are three primary approaches:

Few-Shot Learning

In Few-shot learning, the model has *multiple input-output pairs* as examples to understand the task new task. This approach leverages the model's pre-

One-Shot Learning

One-shot learning is a more constrained form of in-context learning where the model is given a *single input-output* example to understand the task. Despite the limited data, the model utilizes its pre-trained parameters and semantic prior knowledge to generate an output that aligns with the task description. This method is often employed when domain-specific data is scarce.

Zero-Shot Learning

The model is not provided with task-specific examples in zero-shot learning. Instead, it relies solely on the task description and pre-existing training data to infer the requirements. This approach tests the model's innate abilities to generalize from its pre-training phase to new, unencountered tasks.

Each approach has advantages and limitations, but they all leverage the model's pre-training and existing model scale to adapt to new tasks. The choice between them often depends on the availability of labeled data, the complexity of the task, and the computational resources at hand.

💡 **Pro tip:** Evaluating the performance and reliability of LLMs is paramount. Explore Lakera's insights on Large Language Model Evaluation to ensure your models deliver accurate and consistent results.

Engineering

Prompt engineering is the art and science of formulating effective prompts that guide the model's chain of thought, enhancing performance on a given task. This involves incorporating multiple demonstration examples across different tasks and ensuring that the input-output correspondence is well-defined.

In large language models (LLMs), prompt engineering has emerged as a crucial strategy to exploit in-context learning (ICL). This technique involves carefully crafting prompts to provide clear instructions and context to the model, enabling it to perform complex tasks more effectively.

Few-shot learning is often combined with prompt engineering to provide a more robust framework. The model can better understand the task description and generate more accurate output by incorporating a few examples within the prompt. This is particularly useful when the available domain-specific data is limited.

While prompt engineering has shown promise, it has challenges. The process can be brittle, with small modifications to the prompt potentially causing large variations in the model's output. Future research is needed to make this process more robust and adaptable to various tasks.

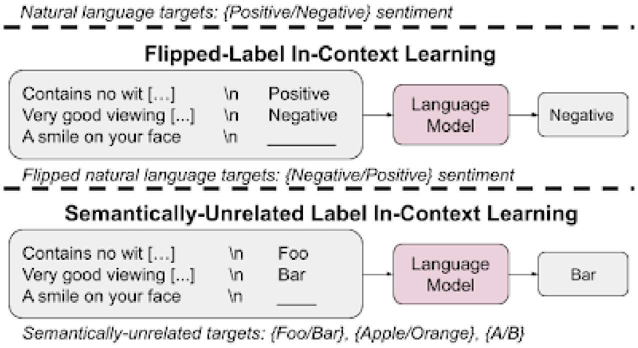
Variants of In-Context Learning in Large Language

model utilizes semantic prior knowledge acquired during the pre-training phase to predict labels based on the format of in-context examples. For instance, if the task involves sentiment analysis, the model will leverage its pre-trained understanding of "positive sentiment" and "negative sentiment" to generate appropriate labels.

Flipped-Label ICL: Flipped-Label ICL introduces complexity by reversing the labels of in-context examples. This forces the model to override its semantic priors, challenging its ability to adhere to the input-label mappings. In this setting, larger models can override their pre-trained semantic priors, a capability not observed in smaller models.

Semantically-Unrelated Label ICL (SUL-ICL): SUL-ICL takes a different approach by replacing the labels of in-context examples with semantically unrelated terms. It directs the model to learn the input-label mappings from scratch, as it can no longer rely on its semantic priors for task completion. Larger models are more adept at this form of learning, indicating their ability to adapt to new task descriptions without relying solely on pre-trained semantic knowledge.

While instruction tuning enhances the model's capacity for learning input-label mappings, it also strengthens its reliance on semantic priors. This dual effect suggests that instruction tuning is an important tool for optimizing ICL performance.


Natural language targets: {Positive/Negative} sentiment
Flipped-Label In-Context Learning
Contains no wit [...] \n Positive
Very good viewing [...] \n Negative
A smile on your face \n
Language Model
Negative
Flipped natural language targets: {Negative/Positive} sentiment
Semantically-Unrelated Label In-Context Learning
Contains no wit [...] \n Foo
Very good viewing [...] \n Bar
A smile on your face \n
Language Model
Bar
Semantically-unrelated targets: {Foo/Bar}, {Apple/Orange}, {A/B}

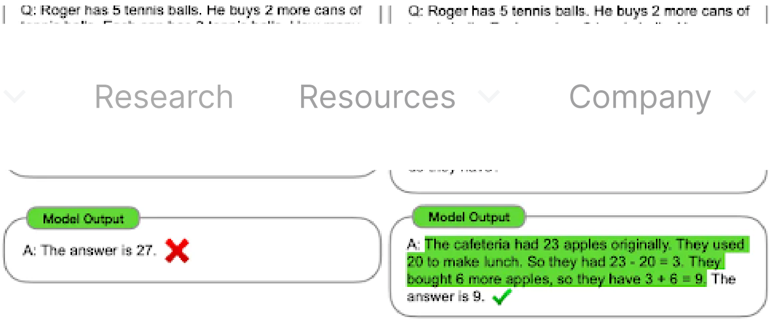
Source

💡 **Pro tip:** Are you curious about the foundational principles behind models like GPT-3? Get a clear understanding with Lakera's Foundation Models Explained article. It's a deep dive into the core mechanics of today's leading LLMs.

Chain-of-Thought Prompting

Chain-of-thought (COT) Prompting is a technique that enhances the reasoning capabilities of large language models (LLMs) by incorporating intermediate reasoning steps into the prompt. This method is particularly effective when combined with few-shot prompting for complex reasoning tasks.

Prompt: *"The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1." Output: "Adding all the odd numbers (9, 15, 1) gives 25. The answer is False."*



Q: Roger has 5 tennis balls. He buys 2 more cans of

Q: Roger has 5 tennis balls. He buys 2 more cans of

A: The answer is 27. ✗

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. ✓

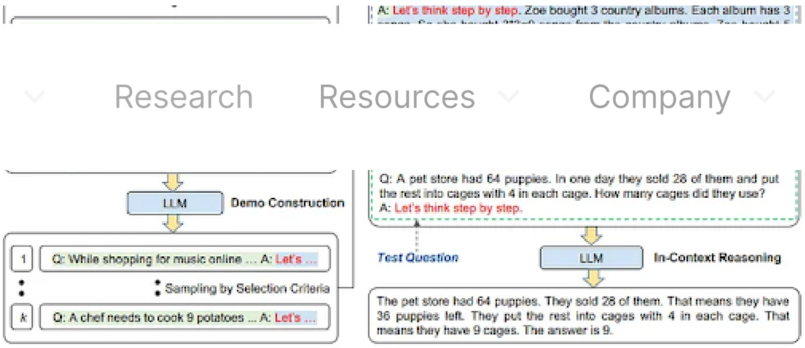
Zero-shot COT Prompting is an extension of COT Prompting that involves adding the phrase "Let's think step by step" to the original prompt. This approach is particularly useful in scenarios with limited examples for the prompt.13

Prompt: *"I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with? Let's think step by step."*

Output: *"First, you started with 10 apples. You gave away 2 apples to the neighbor and 2 to the repairman, so you had 6 apples left. Then you bought 5 more apples, so now you had 11 apples. Finally, you ate 1 apple to remain with 10 apples."*

COT Prompting is closely related to In-Context Learning (ICL) as both techniques aim to leverage LLMs' pre-training data and model parameters for task-specific learning. While ICL focuses on few-shot learning and prompt engineering, COT Prompting emphasizes the chain of thought, prompting complex reasoning.

💡 **Pro tip:** Crafting effective prompts is an art and a science. Enhance your LLM's performance with Lakera's Prompt Engineering Guide. Learn the strategies to guide your model's chain of thought effectively.


A: Let's think step by step. Zoe bought 3 country albums. Each album has 3
Q: A pet store had 64 puppies. In one day they sold 28 of them and put the rest into cages with 4 in each cage. How many cages did they use?
A: Let's think step by step.
LLM
Demo Construction
1
Q: While shopping for music online ... A: Let's ...
Sampling by Selection Criteria
k
Q: A chef needs to cook 9 potatoes ... A: Let's ...
Test Question
LLM
In-Context Reasoning
The pet store had 64 puppies. They sold 28 of them. That means they have 36 puppies left. They put the rest into cages with 4 in each cage. That means they have 9 cages. The answer is 9.

Source

In-context learning real-life applications

In-context learning (ICL) has emerged as a transformative approach in large language models (LLMs), enabling them to adapt to new tasks without explicit retraining. The real-world applications of ICL are vast and span various sectors, showcasing the versatility and potential of this learning paradigm. Here are five key applications where ICL is making or has the potential to make a significant impact:

Sentiment Analysis: Leveraging the power of ICL, LLMs can be fed with a few example sentences and their sentiments (positive or negative). The model can accurately determine its sentiment without explicit training when prompted with a new sentence. This capability can revolutionize customer feedback analysis, market research, and social media monitoring.

Customized Task Learning: Traditional machine learning models require retraining with new data for every new task. However, with ICL, LLMs can learn to perform a task by simply being shown a few examples. This drastically reduces the time and computational resources required, making it feasible model can be prompted to translate new sentences, bridging communication gaps in global businesses.

Code Generation: By feeding the model with a few examples of a coding problem and its solution, the model can generate code for a new, similar problem. This can expedite software development processes and reduce manual coding efforts.

Medical Diagnostics: ICL can be utilized for diagnostic purposes by showing the model a few examples of medical symptoms and their corresponding diagnoses; the model can be prompted to diagnose new cases. This can aid medical professionals in making informed decisions and providing timely care to patients.

Challenges, Limitations, and ICL Research

In-context learning (ICL) allows models to adapt and learn from new input-output pairs without explicit retraining. While ICL has great potential, it has its challenges and limitations, as follows:

Model Parameters and Scale: The efficiency of ICL is closely tied to the scale of the model. Smaller models exhibit a different proficiency in in-context learning than their larger counterparts.

Training Data Dependency: The effectiveness of ICL is contingent on the quality and diversity of the training data. Inadequate or biased training data can lead to suboptimal performance.

dealing with highly specialized domains. Domain

Model Fine-Tuning: Even with ICL, there might be scenarios where model fine-tuning becomes necessary to cater to specific tasks or correct undesirable emergent abilities.

The ICL research landscape is rapidly evolving, and recent advancements have shown how large language models, such as GPT-3, leverage in-context learning. Researchers are probing into the underlying mechanisms, the training data, the prompts, or the architectural nuances that give rise to ICL. The future of ICL holds promise, but there are still many unanswered questions and challenges to overcome.

Ethics and Fairness: In a dynamic learning environment, there's an inherent risk of perpetuating biases and inequalities that the model might have learned from its training data. Ensuring that artificial intelligence operates ethically and fairly, especially when contexts continually evolve, is a formidable challenge.

Privacy and Security: As LLMs integrate more deeply into applications and systems, the potential for security breaches increases. Over time, storage and updating knowledge from different contexts can lead to significant privacy and security concerns. Protecting sensitive information, especially in a domain where the model continually learns, presents a complex challenge.

Large Language Models (LLMs) present a range of security challenges, including vulnerabilities to prompt injection attacks, potential data leakages, and unauthorized access. Lakera is paving the way in building AI solutions for high-stakes environments with decades of experience. While LLM providers may not fully address these inherent risks, Lakera Guard offers robust solutions to protect your LLMs.

(ICL) from 2023:

Learning to Retrieve In-Context Examples for Large Language Models

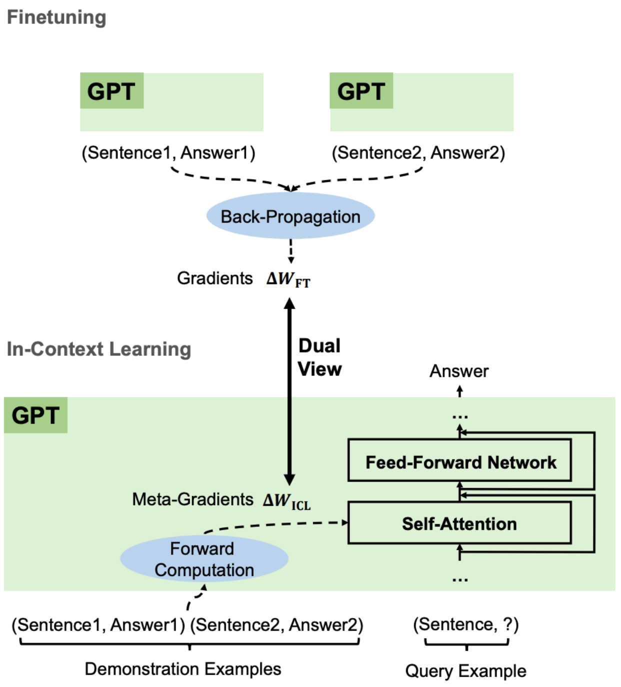
This paper introduces a unique framework to iteratively train dense retrievers that can pinpoint high-quality in-context examples for LLMs. The proposed method first establishes a reward model based on LLM feedback to assess candidate example quality, followed by employing knowledge distillation to cultivate a bi-encoder-based dense retriever. Experimental outcomes across 30 tasks reveal that this framework considerably bolsters in-context learning performance and exhibits adaptability to tasks not seen during training.

Structured Prompting: Scaling In-Context Learning to 1,000 Examples

This paper introduces structured prompting, a method that transcends these length limitations and scales in-context learning to thousands of examples. The approach encodes demonstration examples with tailored position embeddings, which are then collectively attended by the test example using a rescaled attention mechanism. Experimental results across various tasks indicate that this method enhances performance and diminishes evaluation variance compared to conventional in-context learning.

Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers

This paper delves into the underlying mechanism of this phenomenon, proposing that language models act as meta-optimizers and that ICL can be viewed as implicit finetuning. The research identifies a dual meta-gradients based on demonstration examples to



Source

💡 **Pro tip:** Interested in the world of Large Language Models? Discover the latest trends, insights, and best practices at Lakera's official website. Stay updated and informed in the ever-evolving landscape of LLMs.

In context learning: Key takeaways

ICL enables LLMs to learn new tasks using natural language prompts without explicit retraining or fine-tuning.

scale, training data quality, and domain

capabilities.

Research in ICL is advancing with innovations like structured prompting, dense retrievers for LLMs, and understanding the dual relationship between Transformer attention and gradient descent.

Real-world applications of ICL span diverse sectors, including sentiment analysis, code generation, language translation, and medical diagnostics.

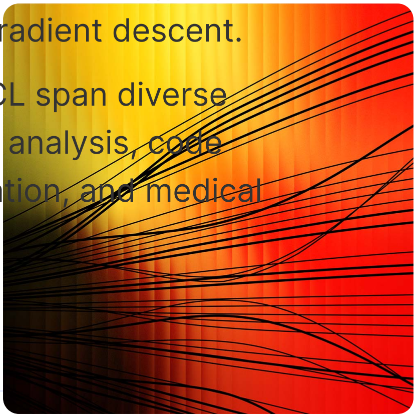
Related Articles

View all related post

Gandalf 5 min read

Inside Agent...

Read more ›



Research 5 min read

Zero-Click Remote...

Read more ›

Follow creator at:

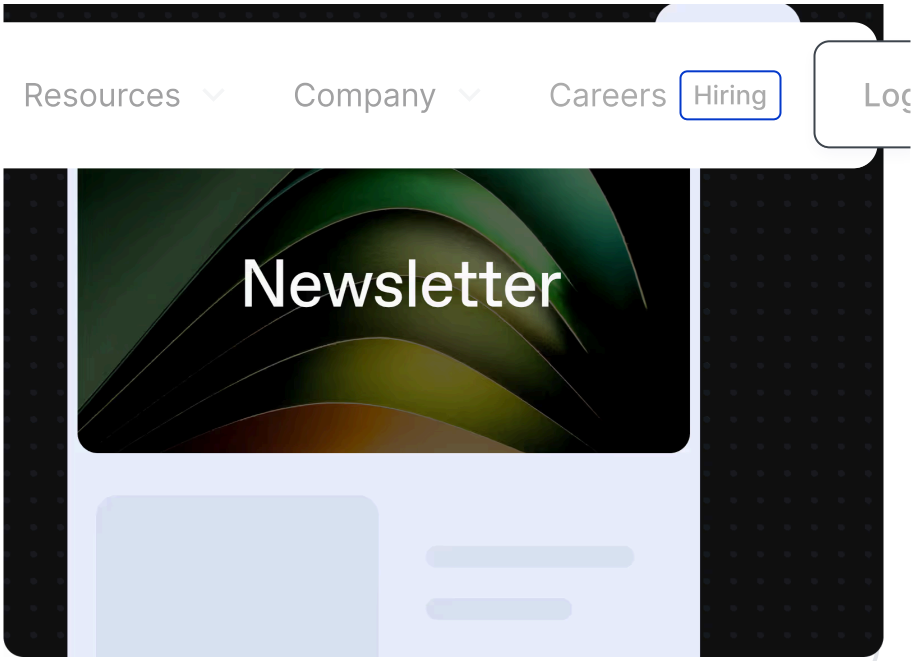
Don't miss the updates!

Subscribe to our newsletter to get the recent updates on Lakera product and other news in the AI LLM world. Be sure you're on track!

Email

unsubscribe anytime. Learn more in our Privacy Policy.

Submit



Book a demo

Start for free

Products

Lakera Guard

Lakera Red

Lakera Gandalf

Resources

Blog

Events

Documentation

Changelog

Security

Privacy Policy

Company

About

Research

News

Careers

Momentum

Contact

Lakera Inc
282 2nd Street, Suite 100