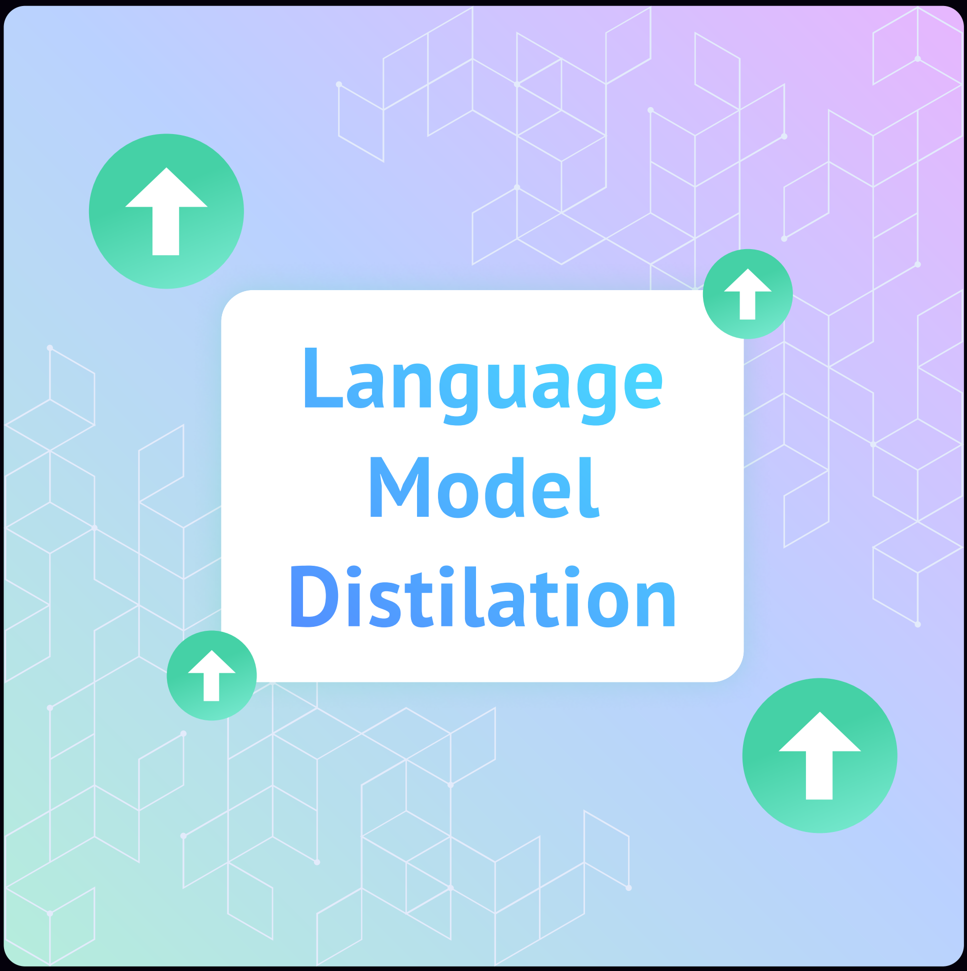


Enhancing Language Model Distillation with Datasaur

Datasaur

Whitepaper
February 2024

Introduction

Language Models (LLMs) are at the forefront of AI advancements, revolutionizing how machines understand and generate human language. However, these models often require significant computational resources, making them **cumbersome**, **expensive**, and challenging to **debug**. Model distillation offers a solution to this challenge by simplifying complex models while retaining their capabilities. Datasaur, a leading platform in data labeling and management, emerges as a critical tool in this process. This paper explores how Datasaur enhances the LLM distillation process and decreases inference time by 80% while maintaining a high accuracy level, potentially improving accuracy by up to 3.26%.

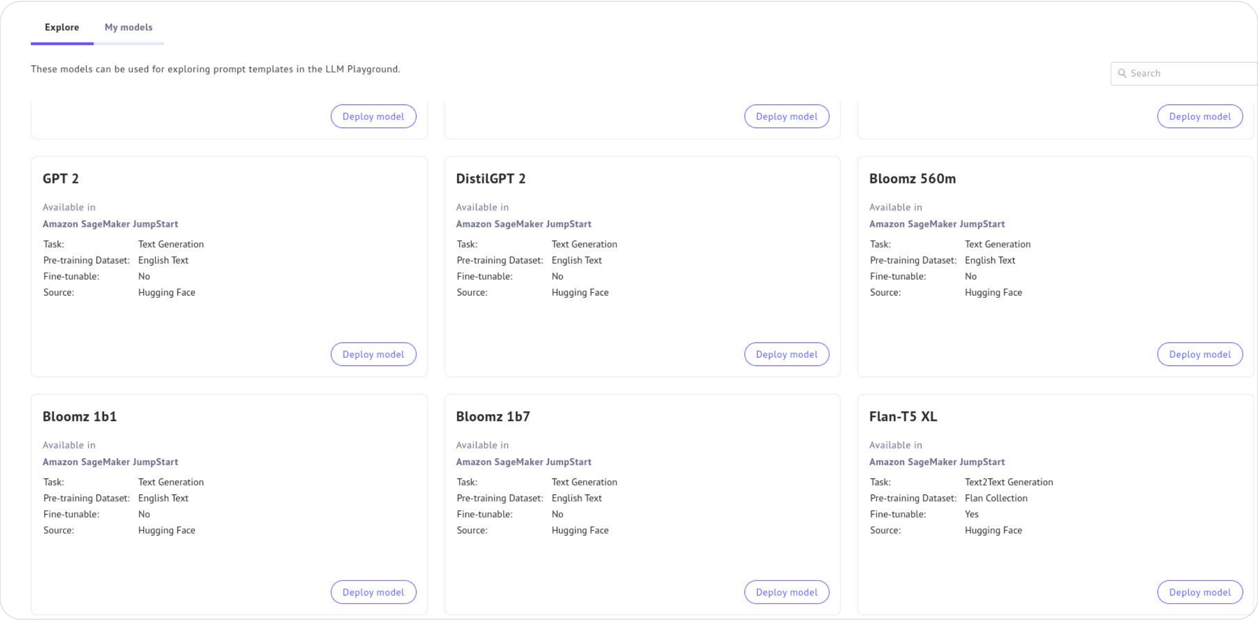
Language Model Distillation: An Overview

The principle of the distillation process involves transferring knowledge from a large, cumbersome model to a smaller model more suitable for deployment. Distilling knowledge means training the smaller model using the outputs of the larger model to generalize data knowledge similarly to the large model. Benefits include quicker deployment and reduced resource requirements, although challenges such as maintaining data quality and managing complexity remain (Hinton, G., Vinyals, O., & Dean, J. (2015)).

The concept of distillation is adapted to address the efficacy problem in LLM applications. LLM distillation compresses large LLMs into smaller, more efficient models, thereby addressing computational demands while maintaining performance. Generally, LLM distillation begins with extracting knowledge from the LLM and then supervising a smaller model using this distilled knowledge. Through this method, the smaller model can incorporate the capabilities of the LLM with significantly improved efficiency. A more comprehensive method is presented by (Cheng-Yu, H, et al. (2023)), demonstrating that LLM distillation is an effective approach for reducing a model's size while enhancing performance.

Advantages of Utilizing LLM Distillation Using Datasaur

LLM distillation is one of the most effective solutions to improve model efficiency and performance. This section explores the advantages of utilizing Datasaur, a leading platform in data labeling and management, for the LLM distillation process.



These models can be used for exploring prompt templates in the LLM Playground.

GPT 2
Available in
Amazon SageMaker JumpStart
Task: Text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face

DistilGPT 2
Available in
Amazon SageMaker JumpStart
Task: Text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face

Bloomz 560m
Available in
Amazon SageMaker JumpStart
Task: Text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face

Bloomz 1b1
Available in
Amazon SageMaker JumpStart
Task: Text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face

Bloomz 1b7
Available in
Amazon SageMaker JumpStart
Task: Text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face

Flan-T5 XL
Available in
Amazon SageMaker JumpStart
Task: Text2Text Generation
Pre-training Dataset: Flan Collection
Fine-tunable: Yes
Source: Hugging Face

Picture 1. LLM Providers integration with Datasaur LLM Labs

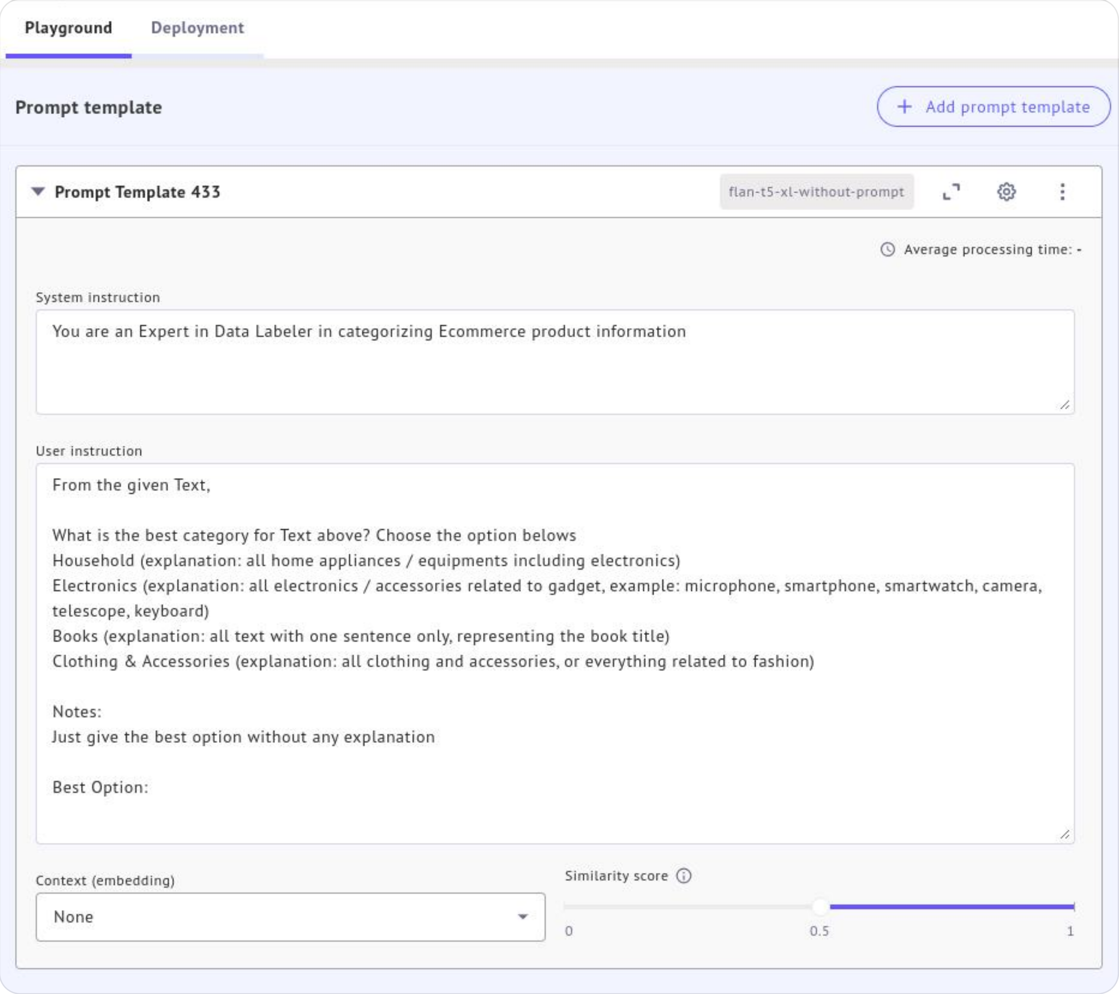
Datasaur helps streamline and simplify each step of the LLM distillation process. Begin by seamlessly managing and deploying the selected LLM provider integrated with Datasaur's platform, such as flan-t5-xl. Next, efficiently label the data using the deployed LLM, encapsulating the LLM's knowledge in the form of generated labels that can be learned by smaller models. Optionally, users can review and evaluate the labeled data through Datasaur's user-friendly interface to enhance data quality. The subsequent crucial step involves training a smaller, more efficient language model using this high-quality data, a process greatly facilitated by Datasaur's feature: Datasaur Dinamic.

LLM distillation culminates in the optimization and deployment of a distilled model that balances efficiency with high accuracy, making it ready for various AI applications. Datasaur enhances the LLM distillation process by offering several advantages, including improvements in data management, quality control, and model training within a unified platform. Datasaur's robust features and ease of use streamline the LLM distillation process, making it more accessible and practical for diverse applications, ranging from small projects to large-scale enterprise solutions.

Case Study

Streamlining User Feedback Classification for an E-commerce Company

E-commerce companies can accumulate a large amount of data; in this example, there are over 50,000 data entries (this case study has been featured in AWS + Datasaur Webinar). In order to pursue a cost-effective approach, 1,000 product description entries were selected as an initial training dataset for the LLM distillation process. These entries were labeled using flan-t5-xl (see Picture 1). To initiate our experimentation, we defined prompt instructions and set the temperature to 0.5 in the LLM settings window (see Picture 2). The selection of temperature to 0.5 was deliberate to ensure consistent output from the LLM. We then connected the model to a text classification project and had the model classify the first 1000 product descriptions (see Picture 3).



System instruction

You are an Expert in Data Labeler in categorizing Ecommerce product information

User instruction

From the given Text,

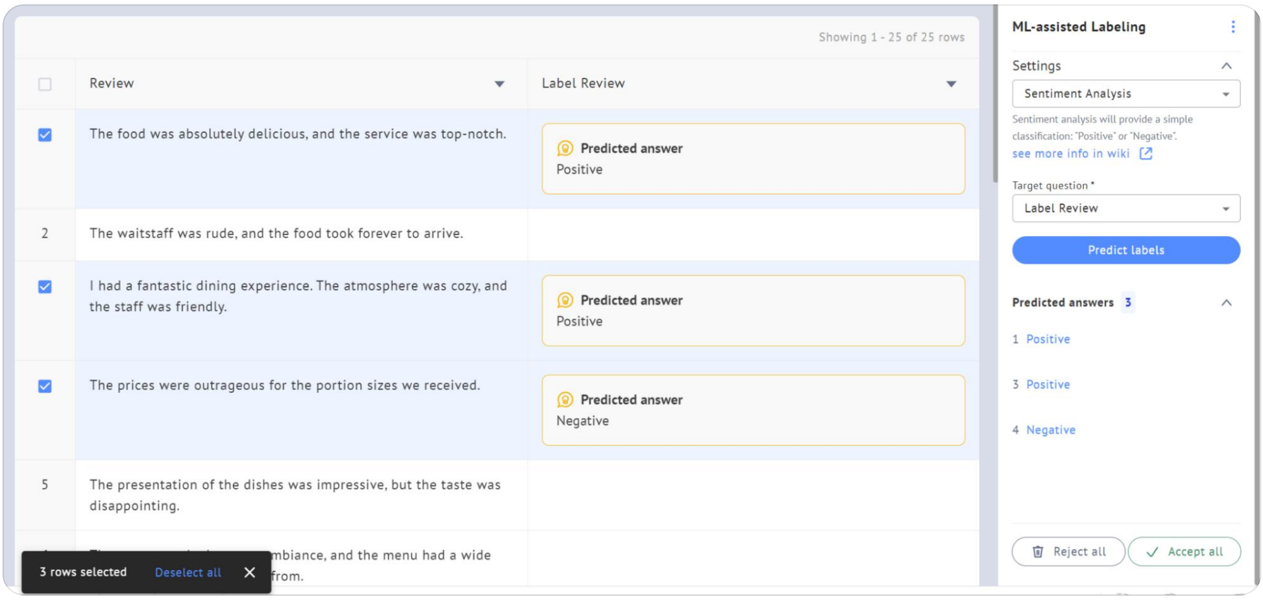
What is the best category for Text above? Choose the option belows
Household (explanation: all home appliances / equipments including electronics)
Electronics (explanation: all electronics / accessories related to gadget, example: microphone, smartphone, smartwatch, camera, telescope, keyboard)
Books (explanation: all text with one sentence only, representing the book title)
Clothing & Accessories (explanation: all clothing and accessories, or everything related to fashion)

Notes:
Just give the best option without any explanation

Best Option:

Context (embedding): None

Picture 2: Define prompt and adjust LLM settings



Showing 1 - 25 of 25 rows

□	Review	Label Review
☑	The food was absolutely delicious, and the service was top-notch.	Predicted answer Positive
2	The waitstaff was rude, and the food took forever to arrive.	
☑	I had a fantastic dining experience. The atmosphere was cozy, and the staff was friendly.	Predicted answer Positive
☑	The prices were outrageous for the portion sizes we received.	Predicted answer Negative
5	The presentation of the dishes was impressive, but the taste was disappointing.	
	mbiance, and the menu had a wide from.	

3 rows selected

ML-assisted Labeling

Settings

Sentiment Analysis

Sentiment analysis will provide a simple classification: "Positive" or "Negative".

see more info in wiki

Target question *

Label Review

Predicted answers 3

1 Positive

3 Positive

4 Negative

Picture 3: Select samples to be labeled by LLM with current settings

This process yielded labels with 90% accuracy, taking 2-3 seconds/entry. In order to enhance the accuracy and relevance of the dataset, we then reviewed and manually validated the labeled data. Utilizing Datasaur Dinamic's AutoML capabilities, we trained a smaller, more efficient NLP model with this refined data. Once deployed, this distilled model successfully classified the remaining 49,000+ product descriptions. After sampling and analyzing 5000 out of the 49,000+ untrained entries, the distilled model achieved an impressive accuracy rate (93.26%), **outperforming the LLM's accuracy (90%)**. Moreover, it accomplished this rapidly at 0.5 seconds per entry, equivalent to processing 2 entries per second.

Metrics	LLM endpoint	Distilled (smaller) model endpoint - reviewed data
Accuracy	90%	93.26%
Inference time	2-3 seconds/entry	0.5 seconds/entry

Through the distillation process, users experienced inference times that were **5x faster** compared to using the LLM endpoint. Despite the increased speed, the accuracy of the distilled model was further enhanced through human review, **surpassing the original LLM's accuracy by up to 3.26%**.

> “ What if there is no time to manually review and fix the generated labels?

To answer this question, we measured the performance of the distilled model when trained with data directly from LLM predictions without undergoing human review. The result indicates that even without human intervention on LLM-generated labels, there’s almost no gap (< 0.5%) between the original LLM label and the distilled model based on the original LLM label.

Metrics	LLM endpoint	Distilled (smaller) model endpoint – non-reviewed data	Distilled (smaller) model endpoint – reviewed data
Accuracy	**90.24%**	**89.9%**	**93.26%**
Inference time	2-3 sec/entry	0.5 seconds/entry	0.5 seconds/entry

In this scenario, given the high performance of LLM, the distillation process has proven successful in delivering a high-quality, smaller model. This distilled model not only replicates the LLM’s data knowledge but can also enhance the accuracy of LLM-generated labels when subjected to human review and refinement. Below are the details and a comparison of time consumption in our case, categorizing over 49,000 product descriptions for E-commerce.

Notes:

- Number of training data: 1000
- Number of inference data (all data except the trained data): 49000+
- Time consumption: Time needed per data x number of data
- Formula for the Human Review process: reviewing time x size of the training data (1000)

Metrics	**LLM endpoint**	**Distilled (smaller) model — non-reviewed data**	**Distilled (smaller) model — reviewed data**
Labeling time	0 hours	0.7 hours	0.7 hours
Human review time	0 hours	0 hours	0.28 hours
Training time	0 hours	0.1 hours	0.1 hours
Inference time	34 hours	6.5 hours	6.5 hours
Total	**34 hours**	**7.3 hours**	**7.58 hours**

This case exemplifies how Datasaur facilitates efficient and accurate large-scale data processing, demonstrating its value in reducing costs and maintaining high data quality standards in LLM distillation, particularly in practical business applications. Through this case, users also have the flexibility to either prioritize the automated process or the accuracy of their dataset. Both approaches are viable considering the accuracy and processing time, with a noteworthy gain of 3.26% accuracy achievable within 0.28 hours using the human reviewing process.

Conclusion

LLM distillation is an increasingly popular technique, with emerging trends focusing on creating more efficient and accurate models. Datasaur's continuous development positions it as a significant contributor to future advancements in this field.

LLM distillation is essential for the practical deployment of advanced AI models. Datasaur enhances this process by ensuring efficient data management and quality control, establishing itself as an indispensable tool in AI development.

References

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531. [https://arxiv.org/abs/1503.02531]

Cheng-Yu, H. (2023). Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. ACL Anthology. [https://aclanthology.org/2023.findings-acl.507.pdf]

Datasaur