

Apris 15, 2025



Accelerating scientific  
research with the truly open,  
fully local AI

**scholasticAI helps to navigate and to efficiently extract knowledge from massive complex multilingual scientific corpora, providing both complete privacy and access to the best-in-class resources.**

# Siloed Science, Gen AI might just save the day

Scientific knowledge grows exponentially more complex and siloed, with **millions of papers published yearly making traditional literature discovery nearly impossible** -

just as LLMs emerge with their unprecedented ability to understand, synthesize, and discover hidden connections across domains, creating a **unique opportunity to revolutionize how researchers explore and build upon existing knowledge.**

# There has been no viable open alternative

## Privacy & Control Gap

Researchers are forced to rely on proprietary black-box platforms, giving up control of their sensitive data and research processes.

**We need private, sovereign alternatives.**

## Need For Infrastructure

Even advanced gen AI models aren't enough - scientific research requires dedicated infrastructure for search, retrieval, processing, and reasoning over complex documents.

**No open solution combines these critical capabilities.**

## Small Language Models Performance Issues

Current open-source SLMs are still too big and lack multilingual capabilities to be viable and efficient locally, while proprietary platforms lead to economic dependency and usage of massive compute resources.

**We need frugal yet performant solutions that can run locally on modest hardware.**

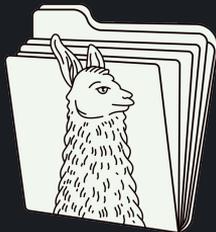


# LLM Stack Tailored For Research

## Tiny Models Local Orchestration



The assistant is powered by an orchestration of 3 models with LlamaFiles inference local server.



## Complex Document Processing, Fast



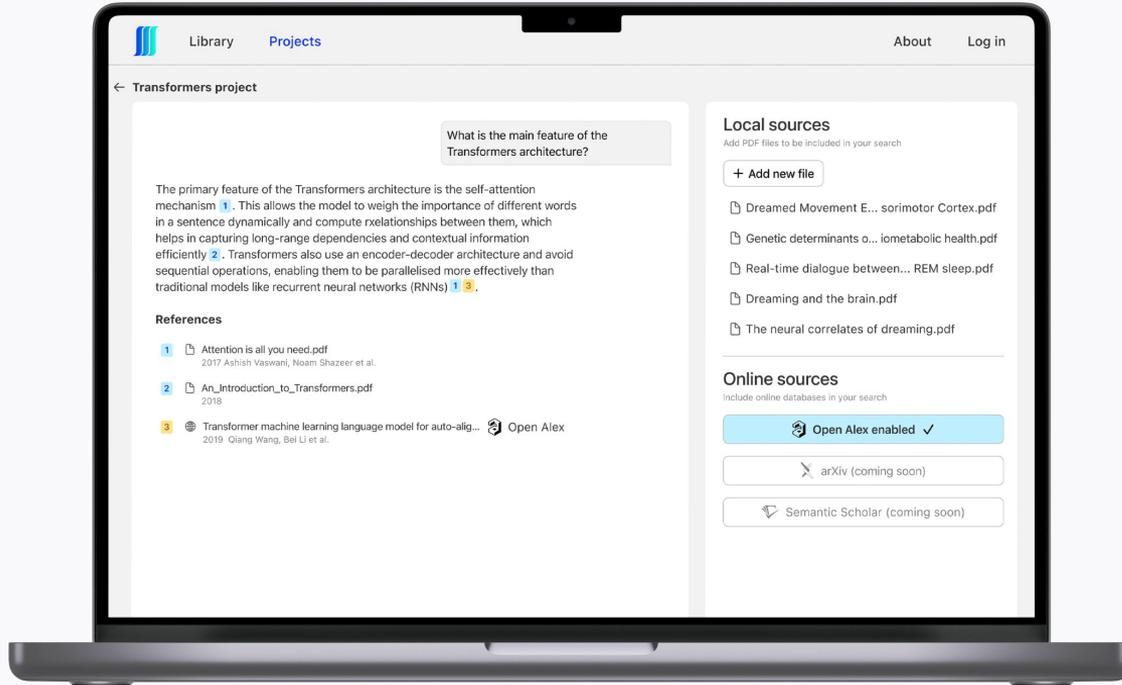
The retrieval is based on a document processing pipeline, integrating a fine-tuned yolov model, allowing to efficiently transform scientific articles with complex layouts into usable and contextualized embedded chunks.

## LLM-Enhanced Search

Users' questions are transformed into augmented structured json queries to OpenAlex API, allowing for far more pertinent, yet computationally efficient search results.



# Local Assistant Running On Modest Hardware



Efficiently running with the following minimum specifications:

- OS: Windows/macOS
- Memory: 8 giga ram
- Processors:
  - Apple Silicon: M1 or newer
  - Intel: Core i5 (8th gen or newer)
  - AMD: Ryzen 3 (3000 series or newer), Ryzen 5 (3000 series or newer)

# LLM-enhanced Search In OpenAlex

Library Projects About Log in

← New research project

Type your question or choose one of the queries

Create a summary of all uploaded files

Why do some people never remember their dreams?

Can we communicate with dreamers while they sleep?

What's the role of the prefrontal cortex in dreaming?

Type a message

↑ Save as a note

### Local sources

Add PDF files to be included in your search

+ Add new file

- Dreamed Movement E... sorimotor Cortex.pdf
- Genetic determinants o... iometabolic health.pdf
- Real-time dialogue between... REM sleep.pdf
- Dreaming and the brain.pdf
- The neural correlates of dreaming.pdf

### Online sources

Include online databases in your search

Enable Open Alex

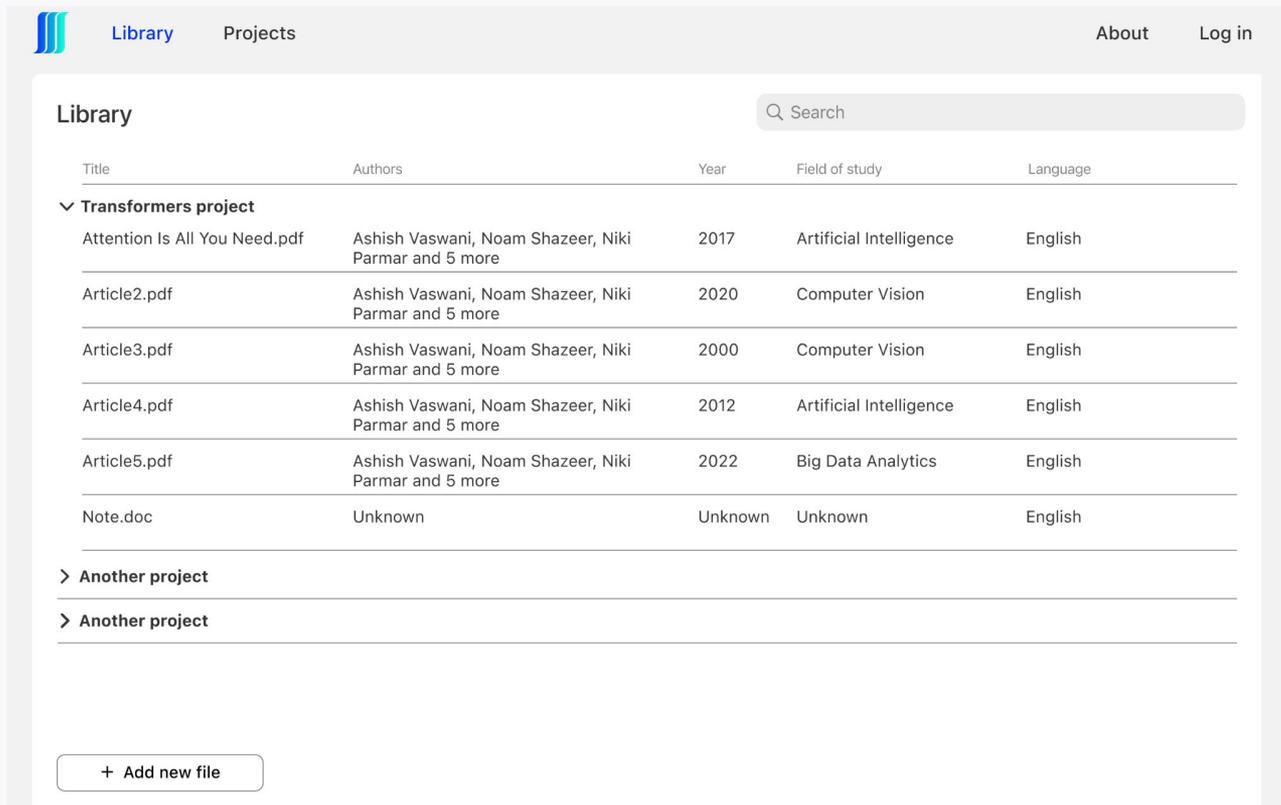
arXiv (coming soon)

Semantic Scholar (coming soon)

With the online mode, users gain access to the largest constantly updated **scientific publications database - OpenAlex**.

This ensures users can tap into the latest scientific works and resources, enhancing their research capabilities with the serendipitous discovery of resources through the **250 millions works**, including books and journal articles, in **30+ languages**.

# Personal Library



The screenshot shows a web interface for a 'Personal Library'. At the top, there is a navigation bar with 'Library' and 'Projects' on the left, and 'About' and 'Log in' on the right. Below the navigation bar, the main content area is titled 'Library' and features a search bar with a magnifying glass icon and the text 'Search'. Below the search bar is a table with the following columns: 'Title', 'Authors', 'Year', 'Field of study', and 'Language'. The table contains several rows of data, including a section for 'Transformers project' which is expanded to show a list of documents. At the bottom of the table, there are two rows for 'Another project'. Below the table, there is a button labeled '+ Add new file'.

Title	Authors	Year	Field of study	Language
▼ Transformers project				
Attention Is All You Need.pdf	Ashish Vaswani, Noam Shazeer, Niki Parmar and 5 more	2017	Artificial Intelligence	English
Article2.pdf	Ashish Vaswani, Noam Shazeer, Niki Parmar and 5 more	2020	Computer Vision	English
Article3.pdf	Ashish Vaswani, Noam Shazeer, Niki Parmar and 5 more	2000	Computer Vision	English
Article4.pdf	Ashish Vaswani, Noam Shazeer, Niki Parmar and 5 more	2012	Artificial Intelligence	English
Article5.pdf	Ashish Vaswani, Noam Shazeer, Niki Parmar and 5 more	2022	Big Data Analytics	English
Note.doc	Unknown	Unknown	Unknown	English
▶ Another project				
▶ Another project				

+ Add new file

The assistant allows users to operate locally with their own locally hosted files, similar to Zotero, enabling **seamless management and categorization** of personal (or previously downloaded via OpenAlex) scientific materials.



# Generation with Sources Citation

The screenshot shows the ScholasticAI interface. At the top, there is a navigation bar with 'Library' and 'Projects' on the left, and 'About' and 'Log in' on the right. Below the navigation bar, the page title is '← Transformers project'. The main content area is divided into two columns. The left column contains a question box with the text 'What is the main feature of the Transformers architecture?'. Below the question box is a generated answer: 'The primary feature of the Transformers architecture is the self-attention mechanism 1. This allows the model to weigh the importance of different words in a sentence dynamically and compute relationships between them, which helps in capturing long-range dependencies and contextual information efficiently 2. Transformers also use an encoder-decoder architecture and avoid sequential operations, enabling them to be parallelised more effectively than traditional models like recurrent neural networks (RNNs) 1 3.' Below the answer is a 'References' section with three items: 1. 'Attention is all you need.pdf' (2017 Ashish Vaswani, Noam Shazeer et al.), 2. 'An\_Introduction\_to\_Transformers.pdf' (2018), and 3. 'Transformer machine learning language model for auto-align...' (2019 Qiang Wang, Bei Li et al.) with an 'Open Alex' logo. The right column is titled 'Local sources' and contains a list of PDF files: 'Dreamed Movement E... sorimotor Cortex.pdf', 'Genetic determinants o... iometabolic health.pdf', 'Real-time dialogue between... REM sleep.pdf', 'Dreaming and the brain.pdf', and 'The neural correlates of dreaming.pdf'. Below this is the 'Online sources' section, which includes 'Open Alex enabled ✓', 'arXiv (coming soon)', and 'Semantic Scholar (coming soon)'. At the bottom of the interface, there is a text input field with the placeholder 'Type a message', an upward arrow button, and a 'Save as a note' button.

ScholasticAI enables transparent research by generating responses with direct citations to both local documents and academic databases like Open Alex. Each claim is linked to its source through numbered references, letting researchers **track and verify every piece of information.**



# Advanced References Auditability

Library Projects About Log in

← Transformers project

The primary feature of the Transformers mechanism 1. This allows the model to v in a sentence dynamically and compute r helps in capturing long-range dependenc efficiently 2. Transformers also use an er sequential operations, enabling them to b traditional models like recurrent neural ne

References

- 1 Attention is all you need.pdf 2017 Ashish Vaswani, Noam Shazeer et a
- 2 An\_Introduction\_to\_Transformers.pdf 2018
- 3 Transformer machine learning langua 2019 Qiang Wang, Bei Li et al.

EXPLORING THE LIMITS OF TRANSFER LEARNING

Scaling strategy	GLUE	CNN3M	SQuAD	SGLUE	EnDe	EnFr	EnRo
* Baseline	83.28	19.24	80.88	71.36	26.08	39.82	27.65
1x size, 4x training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1x size, 4x batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2x size, 2x training steps	<b>86.18</b>	19.66	<b>84.18</b>	77.18	27.52	<b>41.03</b>	28.19
4x size, 1x training steps	<b>85.91</b>	19.73	<b>83.86</b>	<b>78.04</b>	27.47	40.71	28.10
4x ensemble	84.77	<b>20.10</b>	83.09	73.74	<b>28.05</b>	40.53	<b>28.97</b>
4x ensemble, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

Table 13: Comparison of different methods of scaling up our baseline model. All methods except ensembling fine-tuned models use 4x the computation as the baseline. "Size" refers to the number of parameters in the model and "training time" refers to the number of steps used for both pre-training and fine-tuning.

for 4x as many steps (Shallue et al., 2018). We include an additional experiment where we train our baseline model with a 4x larger batch size to compare these two cases.

It is common practice on many of the benchmarks we consider to ensemble additional performance by training and evaluating using an ensemble of models. This provides an orthogonal way of using additional computation. To compare other scaling methods to ensembling, we also measure the performance of an ensemble of 4 separately pre-trained and fine-tuned models. We average the logits across the ensemble before feeding them into the output softmax nonlinearity to obtain an aggregate prediction. Instead of pre-training 4 separate models, a cheaper alternative is to take a single pre-trained model and produce 4 separate fine-tuned versions. While this does not use our entire 4x computational budget, we also include this method to see if it produces competitive performance to the other scaling methods.

The performance achieved after applying these various scaling methods is shown in Table 13. Unsurprisingly, increasing the training time and/or model size consistently improves the baseline. There was no clear winner between training for 4x as many steps or using a 4x larger batch size, though both were beneficial. In general, increasing the model size resulted in an additional bump in performance compared to solely increasing the training time or batch size. We did not observe a large difference between training a 2x bigger model for 2x as long and training a 4x bigger model on any of the tasks we studied. This suggests that increasing the training time and increasing the model size can be complementary means of improving performance. Our results also suggest that ensembling provides an orthogonal and effective means of improving performance through scale. In some tasks (CNN3M, WMT English to German, and WMT English to Romanian), ensembling 4 completely separately trained models significantly outperformed every other scaling approach. Ensembling models that were pre-trained together but fine-tuned separately also gave a substantial performance increase over the baseline, which suggests a cheaper means of improving performance. The only exception was SuperGLUE, where neither ensembling approach significantly improved over the baseline.

We note that different scaling methods have different trade-offs that are separate from their performance. For example, using a larger model can make downstream fine-tuning and

35

+ Add new

When generating a response, users can instantly view the exact page or section being referenced by clicking on citations, allowing them to verify context and accuracy in real-time.



# Why this moment is important

We are experiencing **growing interest from both clients and partners**, supported by an increased emphasis to privacy-first and sustainable solutions, as well as maturing AI regulations in EU and internationally.

With the **market conditions being so favourable**, it is vital for us to be able to double down on our go-to-market efforts, which **requires funding**.

