  
  
(crew\_env) C:\Users\USER\Downloads\Test\_Agent\Test\_5>python run\_redel.py

✅ AutoAgentSystem 啟動！直接輸入你的總任務，Ctrl+C / exit 可退出。

USER: 請使用 Semantic Scholar 搜尋 "large language models" 的研究文獻，摘要三篇重點，並指出引用數最高的是哪一篇。把這個任務委派給一位研究助理。

AI:

[🛠️ Tool Call]: [ToolCall(id='call\_xp0nemhYbrXoKNLdC4a223eO', type='function', function=FunctionCall(name='delegate', arguments='{"instructions":"請搜尋 \\"large language models\\" 的相關研究文獻，並摘要三篇研究文獻的重點內容。請特別指出其中引用數最高的論文。","who":"Research Assistant"}'))]

[✅ 任務指派] Agent: alpha

📄 被指派的任務：

請搜尋 "large language models" 的相關研究文獻，並摘要三篇研究文獻的重點內容。請特別指出其中引用數最高的論文。

----------------------------------------

AI:'alpha' is helping you with this request.

AI:

[🛠️ Tool Call]: [ToolCall(id='call\_RoQNdjmcH10ODIEOtlanb38a', type='function', function=FunctionCall(name='wait', arguments='{"until":"next"}'))]

AI:alpha:1. Harefuah. 2025 May;164(5):276-280.

[ARTIFICIAL INTELLIGENCE TOOLS AND THEIR USE IN MEDICINE CHATGPT - NOT THE ONLY

PLAYER IN THE ARENA].

[Article in Hebrew]

Weizman Z(1), Degany O(2), Shoenfeld Y(3).

Author information:

(1)Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva,

Israel.

(2)Tel-Aviv University Sackler School of Medicine and Medint Medical

Intelligence, Tel-Aviv, Israel.

(3)Reichman University ,Herzelia, Israel, and Zabludowicz Center for Autoimmune

Diseases, Sheba Medical Center, Tel-Hashomer, Ramat-Gan, Israel.

In recent years, there has been a remarkable growth in the development and use

of artificial intelligence tools in medicine based on large language models.

This review will describe the main existing tools and their various applications

for medical staff and patients. Despite its popularity, we will show that

ChatGPT is not the only tool and that other tools are sometimes preferable. We

will review research comparisons between different tools' effectiveness in

various tasks. It will be shown that these tools lack specific performances,

such as accuracy and reliability in providing information, understanding

clinical context, and making diagnoses. The number of studies on these topics is

small, and sometimes their presented results contradict each other. Additional

quality research is needed to characterize and improve these tools and designate

specific tools for different medical uses. Despite the many advantages and

enormous potential inherent in these models, they should be used cautiously, as

they only aid the treating physician and do not replace his knowledge,

professional experience, and human judgment.

PMID: 40413583 [Indexed for MEDLINE]

1. Gastroenterology. 2025 May 22:S0016-5085(25)00774-7. doi:

10.1053/j.gastro.2025.05.012. Online ahead of print.

The Role of Artificial Intelligence in Chronic Liver Diseases and Liver

Transplantation.

Spann A(1), Strauss AT(2), Davis SE(3), Bhat M(4).

Author information:

(1)Division of Gastroenterology, Hepatology, and Nutrition, Department of

Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

Electronic address: ashley.spann@vumc.org.

(2)Department of Medicine, Johns Hopkins University, School of Medicine,

Baltimore, Maryland, USA.

(3)Department of Biomedical Informatics, Vanderbilt University Medical Center,

Nashville, Tennessee, USA.

(4)Division of Gastroenterology and Hepatology, Department of Medicine,

University of Toronto, Toronto, Ontario, Canada.

In hepatology, pattern recognition in laboratory data and clinical

characteristics is the hallmark of clinical care. Artificial intelligence (AI)

tools like machine or deep learning and large language models provide

interesting mechanisms for facilitating care advancement. The complexity and

diversity of data, alongside genetic, environmental and lifestyle factors, all

contribute to individualized recommendations intuitively made by clinicians for

patients with liver disease. AI tools provide the opportunity to train on high

volume data and simulate the clinician's subconscious thought processes in

decision-making. With tremendous growth in hepatology-focused AI, critical

efforts are needed for considering multicenter efforts and enabling collection

of clean data that is as free as possible of bias. Prospective evaluation of AI

tools seamlessly integrated into workflows, especially through clinical trials,

as well as patient partner and clinical stakeholder engagement will be key to

building trust in the individualized predictions provided. In this review, we

delve into the AI literature in hepatology for diagnostic, prognostic and

therapeutic applications.

Copyright © 2025 AGA Institute. Published by Elsevier Inc. All rights reserved.

DOI: 10.1053/j.gastro.2025.05.012

PMID: 40412666

1. Adv Sci (Weinh). 2025 May 24:e03135. doi: 10.1002/advs.202503135. Online ahead

of print.

deepTFBS: Improving within- and Cross-Species Prediction of Transcription Factor

Binding Using Deep Multi-Task and Transfer Learning.

Zhai J(1)(2), Zhang Y(1), Zhang C(1)(2), Yin X(3), Song M(1)(2), Tang C(3), Ding

P(1)(2), Li Z(1)(2), Ma C(1)(2).

Author information:

(1)State Key Laboratory for Crop Stress Resistance and High-Efficiency

Production, Center of Bioinformatics, College of Life Sciences, Northwest A&F

University, Yangling, Shaanxi, 712100, China.

(2)Key Laboratory of Biology and Genetics Improvement of Maize in Arid Area of

Northwest Region, Ministry of Agriculture, Northwest A&F University, Yangling,

Shaanxi, 712100, China.

(3)College of Life Sciences, Northwest A&F University, Yangling, Shaanxi,

712100, China.

The precise prediction of transcription factor binding sites (TFBSs) is crucial

in understanding gene regulation. In this study, deepTFBS, a comprehensive deep

learning (DL) framework that builds a robust DNA language model of TF binding

grammar for accurately predicting TFBSs within and across plant species is

presented. Taking advantages of multi-task DL and transfer learning, deepTFBS is

capable of leveraging the knowledge learned from large-scale TF binding profiles

to enhance the prediction of TFBSs under small-sample training and cross-species

prediction tasks. When tested using available information on 359 Arabidopsis

TFs, deepTFBS outperformed previously described prediction strategies, including

position weight matrix, deepSEA and DanQ, with a 244.49%, 49.15%, and 23.32%

improvement of the area under the precision-recall curve (PRAUC), respectively.

Further cross-species prediction of TFBS in wheat showed that deepTFBS yielded a

significant PRAUC improvement of 30.6% over these three baseline models.

deepTFBS can also utilize information from gene conservation and binding motifs,

enabling efficient TFBS prediction in species where experimental data

availability is limited. A case study, focusing on the WUSCHEL (WUS)

transcription factor, illustrated the potential use of deepTFBS in cross-species

applications, in our example between Arabidopsis and wheat. deepTFBS is

publically available at https://github.com/cma2015/deepTFBS.

© 2025 The Author(s). Advanced Science published by Wiley‐VCH GmbH.

DOI: 10.1002/advs.202503135

PMID: 40411397📄 Chain of Thought Prompting Elicits Reasoning in Large Language Models

🔗 https://www.semanticscholar.org/paper/1b6e810ce0afd0dd093f789d2b2742d047e316d5

📚 Cited: 8727

📝 We explore how generating a chain of thought -- a series of intermediate reasoning steps -- significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called chain of thought prompting, where a few chain of thought demonstrations are provided as exemplars in prompting. Experiments on three large language models show that chain of thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a 540B-parameter language model with just eight chain of thought exemplars achieves state of the art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

📄 BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

🔗 https://www.semanticscholar.org/paper/3f5b31c4f7350dc88002c121aecbdc82f86eb5bb

📚 Cited: 4310

📝 The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models. This paper proposes BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. BLIP-2 achieves state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods. For example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model's emerging capabilities of zero-shot image-to-text generation that can follow natural language instructions.

📄 LoRA: Low-Rank Adaptation of Large Language Models

🔗 https://www.semanticscholar.org/paper/a8ca46b171467ceb2d7652fbfb67fe701ad86092

📚 Cited: 9710

📝 An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned models, each with 175B parameters, is prohibitively expensive. We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. We also provide an empirical investigation into rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the integration of LoRA with PyTorch models and provide our implementations and model checkpoints for RoBERTa, DeBERTa, and GPT-2 at https://github.com/microsoft/LoRA.📄 \*\*Lost in Translation: Large Language Models in Non-English Content Analysis\*\*

In recent years, large language models (e.g., Open AI's GPT-4, Meta's LLaMa,

Google's PaLM) have become the dominant approach for building AI systems to

analyze and generate language online. However, the automated systems that

increasingly mediate our interactions online -- such as chatbots, content

moderation systems, and search engines -- are primarily designed for and work

far more effectively in English than in the world's other 7,000 languages.

Recently, researchers and technology companies have attempted to extend the

capabilities of large language models into languages other than English by

building what are called multilingual language models.

In this paper, we explain how these multilingual language models work and

explore their capabilities and limits. Part I provides a simple technical

explanation of how large language models work, why there is a gap in available

data between English and other languages, and how multilingual language models

attempt to bridge that gap. Part II accounts for the challenges of doing

content analysis with large language models in general and multilingual

language models in particular. Part III offers recommendations for companies,

researchers, and policymakers to keep in mind when considering researching,

developing and deploying large and multilingual language models.

🔗 http://arxiv.org/abs/2306.07377v1

📄 \*\*Cedille: A large autoregressive French language model\*\*

Scaling up the size and training of autoregressive language models has

enabled novel ways of solving Natural Language Processing tasks using zero-shot

and few-shot learning. While extreme-scale language models such as GPT-3 offer

multilingual capabilities, zero-shot learning for languages other than English

remain largely unexplored. Here, we introduce Cedille, a large open source

auto-regressive language model, specifically trained for the French language.

Our results show that Cedille outperforms existing French language models and

is competitive with GPT-3 on a range of French zero-shot benchmarks.

Furthermore, we provide an in-depth comparison of the toxicity exhibited by

these models, showing that Cedille marks an improvement in language model

safety thanks to dataset filtering.

🔗 http://arxiv.org/abs/2202.03371v1

📄 \*\*How Good are Commercial Large Language Models on African Languages?\*\*

Recent advancements in Natural Language Processing (NLP) has led to the

proliferation of large pretrained language models. These models have been shown

to yield good performance, using in-context learning, even on unseen tasks and

languages. They have also been exposed as commercial APIs as a form of

language-model-as-a-service, with great adoption. However, their performance on

African languages is largely unknown. We present a preliminary analysis of

commercial large language models on two tasks (machine translation and text

classification) across eight African languages, spanning different language

families and geographical areas. Our results suggest that commercial language

models produce below-par performance on African languages. We also find that

they perform better on text classification than machine translation. In

general, our findings present a call-to-action to ensure African languages are

well represented in commercial large language models, given their growing

popularity.

🔗 http://arxiv.org/abs/2305.06530v1我已經搜尋了關於 "large language models" 的相關研究文獻，並選擇了三篇研究文獻進行摘要。以下是摘要內容：

1. \*\*Chain of Thought Prompting Elicits Reasoning in Large Language Models\*\*

- \*\*引用數：8727\*\*

- 這篇論文探討了如何通過鏈式思考（Chain of Thought）提示來顯著提升大型語言模型在複雜推理任務中的表現。作者展示了在足夠大的語言模型中，鏈式思考能力自然地出現，這是通過提供一些鏈式思考範例作為提示來實現的。實驗結果顯示，這種方法在多種算術、常識和符號推理任務中提高了性能。特別是在GSM8K數學文字問題基準上，使用540B參數的語言模型僅需八個鏈式思考範例即可達到最新的準確性，甚至超過了經過微調的GPT-3。

2. \*\*BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models\*\*

- \*\*引用數：4310\*\*

- 本文提出了一種通用且高效的預訓練策略BLIP-2，通過使用現成的預訓練圖像編碼器和大型語言模型來引導視覺語言預訓練。BLIP-2使用輕量級的Querying Transformer來橋接模態差距，並在兩個階段進行預訓練。即使訓練參數顯著少於現有方法，BLIP-2在多個視覺語言任務中仍能達到最先進的性能。

3. \*\*LoRA: Low-Rank Adaptation of Large Language Models\*\*

- \*\*引用數：9710\*\*

- 本文提出了一種低秩適應（LoRA）方法，該方法通過凍結預訓練模型權重並在每個Transformer層中注入可訓練的秩分解矩陣，大大減少了下游任務的可訓練參數。LoRA在RoBERTa、DeBERTa、GPT-2和GPT-3上表現與微調相當或更好，儘管其可訓練參數更少，訓練吞吐量更高，且不像適配器那樣增加額外的推理延遲。

在這三篇文獻中，引用數最高的是 "Chain of Thought Prompting Elicits Reasoning in Large Language Models"，引用數為8727。

AI:我已經委派了一位研究助理來搜尋關於 "large language models" 的研究文獻，並為您總結了三篇重點內容。以下是總結：

1. \*\*Chain of Thought Prompting Elicits Reasoning in Large Language Models\*\*

- \*\*引用數：8727\*\*

- 這篇論文探討如何利用鏈式思考（Chain of Thought）提示來顯著提升大型語言模型在複雜推理任務中的表現。研究顯示，這種方法在多種推理任務中提高了性能，特別是在數學問題基準測試中顯示出顯著提升。

2. \*\*BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models\*\*

- \*\*引用數：4310\*\*

- 本文提出了一種高效的預訓練策略BLIP-2，通過使用預訓練圖像編碼器和語言模型來進行視覺語言預訓練，並在多個視覺語言任務中達到最先進的性能。

3. \*\*LoRA: Low-Rank Adaptation of Large Language Models\*\*

- \*\*引用數：9710\*\*

- 本文提出了一種低秩適應（LoRA）方法，通過凍結預訓練模型並注入可訓練的秩分解矩陣，減少了下游任務的可訓練參數，同時保持或提升模型性能。

在這三篇文獻中，引用數最高的是 \*\*"Chain of Thought Prompting Elicits Reasoning in Large Language Models"\*\*，引用數為8727。

USER: exit

👋 使用者中斷。再見！