Academic Paper Writing in Artificial Intelligence

Related Work

Instructor: Ying-Jia Lin
Octobor 22th, 2026

Course Website
Download slide here!

Outline

- Week 6 Assignment Summary (20 min)
- What a Related Work section is (10 min)
- Related Work in **3 steps** and examples (40 min)
- **Your Turn for revision and new assignment (80 min)**
  - Feel free to discuss with me !!

Week 6 Assignment Summary

- The details are not sufficient.
  - ❌ We used cross-entropy loss to train the language model.
  - ✅ We used **autoregressive** cross-entropy loss to train the language model. **The loss was averaged over valid target tokens, excluding prompt and padding tokens.**
  - ✅ We trained the model using AdamW **with a learning rate of 2×10^{-5}, a batch size of 4, and a maximum of three epochs. The checkpoint with the lowest validation loss was selected for final evaluation**.

Where we are

Main Body (Your Main Paper)
Abstract
Introduction
Related Work (Literature Review)
Methods
Results
Discussion
Conclusion

Depends on where you submit

Depends on where you submit

Related Work builds the gap

- A Related Work section is not a summary of everything you have read.
  - *Its job is to make your paper necessary.*
- Every group of papers you describe should leave the reader thinking "so somebody still needs to solve that." ➡️ *Your paper is that somebody.*
- If a reader finishes your Related Work and thinks the problem sounds solved already, the section has worked against you.

Two Examples Last Week

- 💡 Click and browse the **Related Work** section

  - Li, Mingjie, et al. "Advancing In-Context Learning for Efficient and Stable Medical Report Generation." IEEE Transactions on Pattern Analysis and Machine Intelligence (2026).

  - Yan, Sixing, et al. "Learning self-critiquing mechanisms for region-guided chest x-ray report generation." International Conference on Learning Representations. Vol. 2026. 2026.

No “Related Work” in many journals

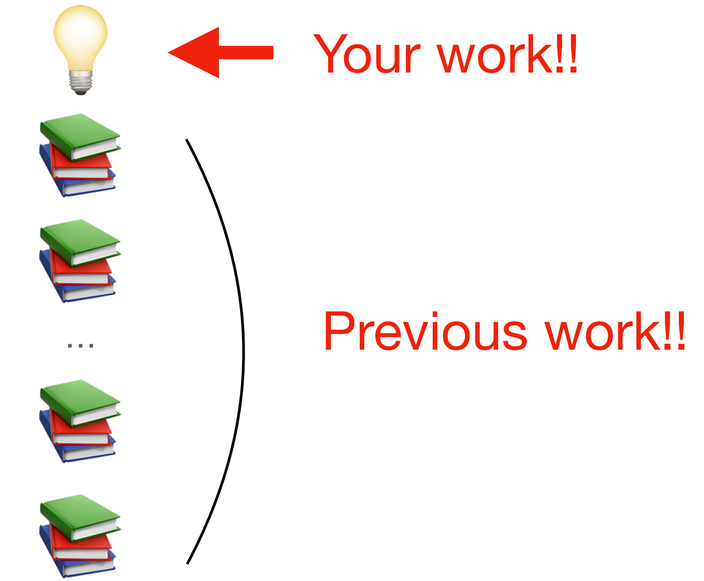
- 💡 Click and browse **Related Work** that is hidden **inside Introduction**

  - Tseng, LW., Lu, YC., Tseng, LC. et al. Model-dependent prompt effects challenge reliable deployment of medical LLMs in traditional Chinese medicine. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-03397-5

  - Levin, C., et al. "Nursing judgment in the age of generative artificial intelligence: A cross-national study on clinical decision-making performance among emergency nurses." International journal of nursing studies (2025): 105216.

Standing on the shoulders of giants





Three steps to write the Related Work section

1. Group: sort your references into two to four families, by the approach

2. Describe: one clause per citation

3. Compare: close each family with what it cannot do

Three steps to write the Related Work section

1. Group: sort your references into two to four families, by the approach

2. Describe: one clause per citation

3. Compare: close each family with what it cannot do

💡 Target length: 400–700 words, two to four paragraphs, one per family.
💡 Step 3 is the one that gets skipped, and it is the only reason the section exists.

Step 1: Group

Step 1: Group

❌ Grouped by **time**:

"In 2019, … In 2020, … More recently, in 2023, …”

❌ Grouped by **paper**:

"Smith et al. proposed X. Jones et al. proposed Y. Lee et al. proposed Z."

✅ Grouped by **approach**:

"Existing work falls into two groups: methods that map structured diagnosis codes, and methods that learn directly from clinical text.”

Example of Step 1 (You can see the sorting)

Tseng, LW., Lu, YC., Tseng, LC. et al. Model-dependent prompt effects challenge reliable deployment of medical LLMs in traditional Chinese medicine. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-03397-5

On the model side, several studies have attempted to enhance LLM capabilities in TCM through domain-specific training. Yang et al. developed TCM-GPT-7B using LoRA-based fine-tuning[9], and Dai et al. developed TCMChat-7B with supervised fine-tuning across six TCM scenarios[10]. However, these models were not evaluated against a common standardized public benchmark, such as national licensing examinations, making direct performance comparisons across studies difficult. More recently, frontier general-purpose models, such as DeepSeek and Qwen, have also been evaluated on national TCM licensing examinations[11,12], yet none of these studies systematically examined the role of prompt engineering. Existing resources such as TCMBench and TCM-Eval provide comprehensive or extensible benchmarks for evaluating general TCM competence[13,14]. In contrast, the present study uses a standardized national examination as a controlled test set to investigate how prompt effects vary across models and question characteristics.

Regarding prompt and context engineering, prior research has established these as effective means of enhancing LLM performance[15], with common strategies including role specification, few-shot demonstrations[16], and CoT prompting[17]. Under certain LLM models, prompt engineering seems much more important than model fine-tuning[18]. However, these strategies do not yield uniform benefits. Different knowledge domains and modifications on prompt engineering may have diverse effects on different models[19,20]. CoT prompting produced no significant accuracy gain for GPT-3.5 on USMLE-style questions (62.8% vs. 61.7%; $p = 0.734$)[21], whereas GPT-4 maintained diagnostic accuracy while additionally generating interpretable clinical reasoning through specialized prompts[22]. The same prompting method can also yield divergent results across tasks: traditional CoT achieved 88.4% accuracy on clinical note comprehension but only 61.7% on medical entrance questions within the same model[23]. Reasoning-optimized models have also shown markedly stronger performance than non-reasoning architectures on the same clinical reasoning tasks[24], further suggesting that the effectiveness of prompt strategies is highly dependent on model characteristics[25]. Inappropriate prompt or context may even lead to wrong clinical judgment in some LLM models, whether open or closed[26–28]. Yet this interaction has not been examined in TCM.

interactions between model selection and prompt engineering, and their subsequent influence on performance in this domain, remain underexplored.

This study aimed to identify the key factors that can be leveraged to improve LLM performance in acquiring and applying TCM knowledge, using the 2025 Taiwan national TCM licensing examination as a standardized benchmark. We systematically evaluated 16 LLMs under seven prompt conditions, extending the single-model experiment from our prior work. The study had three main objectives: (1) to establish each model's baseline performance and quantify advances in LLM capability in TCM across successive generations; (2) to analyze the interactions between model selection, prompt engineering strategies, and question characteristics through systematic prompt ablation experiments, and to investigate whether there is a question-characteristics-dependent prompt effect; and (3) to apply the three-category error classification framework from our previous study (misinterpretation, misunderstanding, misapplication) to identify the sources of performance gaps.

Results

Baseline performance of LLMs in the TCM national examination

Following the evaluation pipeline summarized in Fig. 1, baseline performance under the pure zero-shot condition, without any prompt or context engineering, varied substantially across models (Fig. 2). Among all evaluated models, pure zero-shot accuracy ranged from 35.37% (Gemma-3n-e4b-it) to 85.89% (Claude-opus-4.5), reflecting considerable variation in baseline model capability.

Using the official passing criterion of the Taiwanese National Examination for Chinese Medicine (≥285 correct responses), approximately half of the evaluated models achieved or exceeded the passing threshold, even under this pure zero-shot setting. High-performing models such as Claude-opus-4.5 (85.89%), GPT-5.2 (84.00%), Gemini-3-flash-preview (83.58%), and Grok-4 (83.37%) clearly surpassed the passing requirement, whereas several mid-tier models demonstrated only marginal performance near the cutoff. In contrast, some open-weight and freely accessible models, including DeepSeek-V3.2 (49.68%) and Gemma-3n-e4b-it (35.37%), failed

How to find your group?

Lay your papers out and ask of each one: what does this paper assume it has?

Papers that assume the same thing belong together. For example:

- papers that need **paired data** vs papers that work **without pairs**
- papers that need **labelled examples** vs papers that **do not**
- papers that use **structured fields** vs papers that use **free text**

Example2 of Step 1

Lee, Changsun, et al. "Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation." Medical Image Analysis (2026): 104077.

et al., 2024; Chaves et al., 2024), these AI models still face significant limitations in broader applications such as radiology report generation or visual question answering (VQA).

The limited success of 3D imaging AI models can be attributed to several key factors. First, there is a fundamental shortage of paired 3D image-report datasets compared to the abundance of datasets available for 2D imaging modalities such as X-rays (Johnson et al., 2019; Demner-Fushman et al., 2012; Irvin et al., 2019), resulting in significant data insufficiency. This challenge is further exacerbated by the immense data volume and computational demands of 3D imaging, which are several hundred times greater than those of 2D imaging (Litjens et al., 2017). Second, the inherent complexity of 3D medical imaging poses significant challenges for effective encoding, as current 3D vision encoders often lack the capacity to capture the full scope of information required for accurate 3D image interpretation (Litjens et al., 2017; Dou et al., 2017; Isensee et al., 2021). Existing vision-language models (VLMs) also struggle to efficiently process the complex and voluminous data inherent in 3D imaging, leading to notable inefficiencies. Specifically, most state-of-the-art models (Wu et al., 2023; Blankemeier et al., 2024; F. Bai et al., 2024) rely on 3D vision encoders that partition image volumes into fixed-size 3D patches. This approach introduces processing inefficiencies by creating unnecessary correlations across the z-axis, further limiting their effectiveness in handling 3D medical imaging data.

To address these challenges, we propose a novel VLM inspired by the way radiologists interpret 3D medical images. Unlike traditional methods that treat the entire 3D volume as a uniform entity, our model, termed MultiSlice Vision Language Model (MS-VLM), processes 3D volumes as a collection of 2D planes. This approach enables the model to encode information from each 2D plane independently before combining them to produce a comprehensive interpretation. By leveraging this method, MS-VLM achieves significantly higher data and computational efficiency while delivering superior performance compared to conventional approaches. Additionally, the structural design of MS-VLM allows it to handle volumetric data with variable slice lengths, enhancing its versatility and adaptability across diverse clinical applications. Notably, MS-VLM excels in scenarios where specific planes and phases are critical for accurate diagnosis, such as rectal MRI, where interpretations across different phases (e.g., pre-contrast and post-contrast) and planes (e.g., sagittal and axial) are essential. By incorporating reconstructions of images from multiple planes and phases, MS-VLM closely mirrors the interpretive processes employed by radiologists, leading to improved diagnostic performance.

Our contributions can be summarized as follows:

- Inspired by radiologists, we propose a novel 3D imaging interpretation method that treats 3D volumes as a collection of 2D planes, significantly improving data and computational efficiency.
- By introducing a novel Z-former, the proposed MS-VLM model handles volumetric data with variable slice lengths, enhancing its practical utility in diverse clinical scenarios.
- Specifically, by incorporating reconstructions from various planes and phases through the Z-former, MS-VLM aligns with radiologists' interpretative workflows, resulting in improved performance.

2. Related work

In this section, we provide a concise overview of recent research on the development of multi-modal foundation models in the medical domain and the approaches of the VLM for 3D medical imaging.

2.1. Multi-modal foundation model in medical domain

Healthcare is inherently multi-modal, requiring clinicians to synthesize information from various sources, such as medical images, clinical notes, lab results, electronic health records (EHR), genomic data, etc. to provide effective patient care. In recent years, significant progress has been made in developing AI systems that achieve expert level performance within individual modalities, such as interpreting radiographic images, analyzing pathology slides, or detecting genetic mutations (Pham et al., 2022; Taori et al., 2023; Qu et al., 2021). However, the complexity and multi-modal nature of healthcare require AI models capable of integrating these diverse data types into a cohesive framework. The early attempts to achieve this integration leverage contrastive learning models, such as CLIP (Radford et al., 2021). CLIP maps images and corresponding contextual information into a shared embedding space through large-scale contrastive learning, demonstrating promising results in integrating simple medical imaging data, like chest radiographs, with radiology reports. Medical adaptions of CLIP such as MedCLIP, CheXzero and CXR-CLIP (Wang et al., 2022; Tiu et al., 2022; You et al., 2023) perform image–text retrieval on chest X-ray (CXR) images and the corresponding reports demonstrating bidirectional understanding of the two modalities.

Building on this foundation, more sophisticated approaches emerged to directly integrate data from different modalities. Unlike earlier CLIP-based methods, these newer approaches enable a more comprehensive understanding of multi-modal data, supporting advanced applications such as detailed medical report generation and VQA. Vision encoders for image data and text encoders for textual information are used together, with cross-attention and self-attention mechanisms facilitating seamless integration of complex information. This deeper level of integration not only correlates visual and textual data but also supports holistic tasks that require a nuanced understanding of medical data. Notable examples include PPKED, MedViLL, and Medical X-VL (Liu et al., 2021; Moon et al., 2022; Park et al., 2024).

More recently, the impressive capabilities of large language models (LLMs) have driven efforts to develop multi-modal models with a generalist architecture. These models aim to handle multiple modalities using a single framework, enhancing flexibility and simplifying the integration process compared to specialized encoder modules. Med-PaLM M (Tu et al., 2024) is a pioneering effort in this area, demonstrating the ability to process text, imaging, and genomic data within a unified framework. This design maximizes flexibility and information transfer across modalities, providing a strong foundation for future multi-modal AI systems. Med-PaLM M has also shown potential to outperform specialized models in specific tasks, suggesting that generalist models may offer superior performance even in domain-specific applications. Subsequent models, such as BioMedGPT and PathChat (K. Zhang et al., 2024; Lu et al., 2024), have further built on these capabilities, underscoring the potential of foundational models to provide generalized solutions in the medical domain.

2.2. VLMs for 3D medical imaging interpretation

Most initial efforts to develop effective multi-modal models in the medical domain have largely focused on simpler 2D imaging modalities, such as plain radiographs and pathology slides (Tu et al., 2024; Li et al., 2024). However, developing such models for 3D medical imaging modalities, like CT and MRI, have presented unique challenges: limited training data and high data dimensionality. Furthermore, current vision encoders often lack the capacity to adequately process the intricate information found in 3D medical images, as they require handling large-scale data while understanding complex spatial relationships.

Unlike natural 3D data such as video, which contains significant redundancy across numerous frames and can comprise thousands of individual frames over a short duration, medical 3D imaging typically consists of fewer slices along the z-axis, usually ranging from 20 to 300.

Step 2: Describe

2.1. Multi-modal foundation model in medical domain

Smith et al. (2020) first pre-trained a medical foundation model using CT image-report pairs.

Jin et al. (2021)

Swift et al. (2023)

2.2. VLMs for 3D medical imaging interpretation

Zhang et al. (2022)

Lee et al. (2023)

Two ways to cite the same paper

Two ways to cite the same paper

1️⃣ **Author as the subject** — the sentence is people >= idea.

“Smith et al. (2020) proposed a graph-based reranker.”

2️⃣ **In brackets** — the sentence is about the idea.

“Graph-based rerankers improve recall on long documents (Smith et al., 2020).”

💡 Related Work is about ideas, so **most** of your citations should sit in **brackets**.

⚠️ Use the author form: Authors’ action (-> the paper itself) is very important;

model doesn’t have a name.

Example of Step 2

Liu, Zhuang, et al. "A convnet for the 2020s." 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, 2022.

Recent convolution-based approaches. Han et al. [25] show that local Transformer attention is equivalent to inhomogeneous dynamic depthwise conv. The MSA block in Swin is then replaced with a dynamic or regular depthwise convolution, achieving comparable performance to Swin. A concurrent work ConvMixer [4] demonstrates that, in small-scale settings, depthwise convolution can be used as a promising mixing strategy. ConvMixer uses a smaller patch size to achieve the best results, making the throughput much lower than other baselines. GFNet [56] adopts Fast Fourier Transform (FFT) for token mixing. FFT is also a form of convolution, but with a global kernel size and circular padding. Unlike many recent Transformer or ConvNet designs, one primary goal of our study is to provide an in-depth look at the process of modernizing a standard ResNet and achieving state-of-the-art performance.

⚠️ Make sure the cited paper matches your words

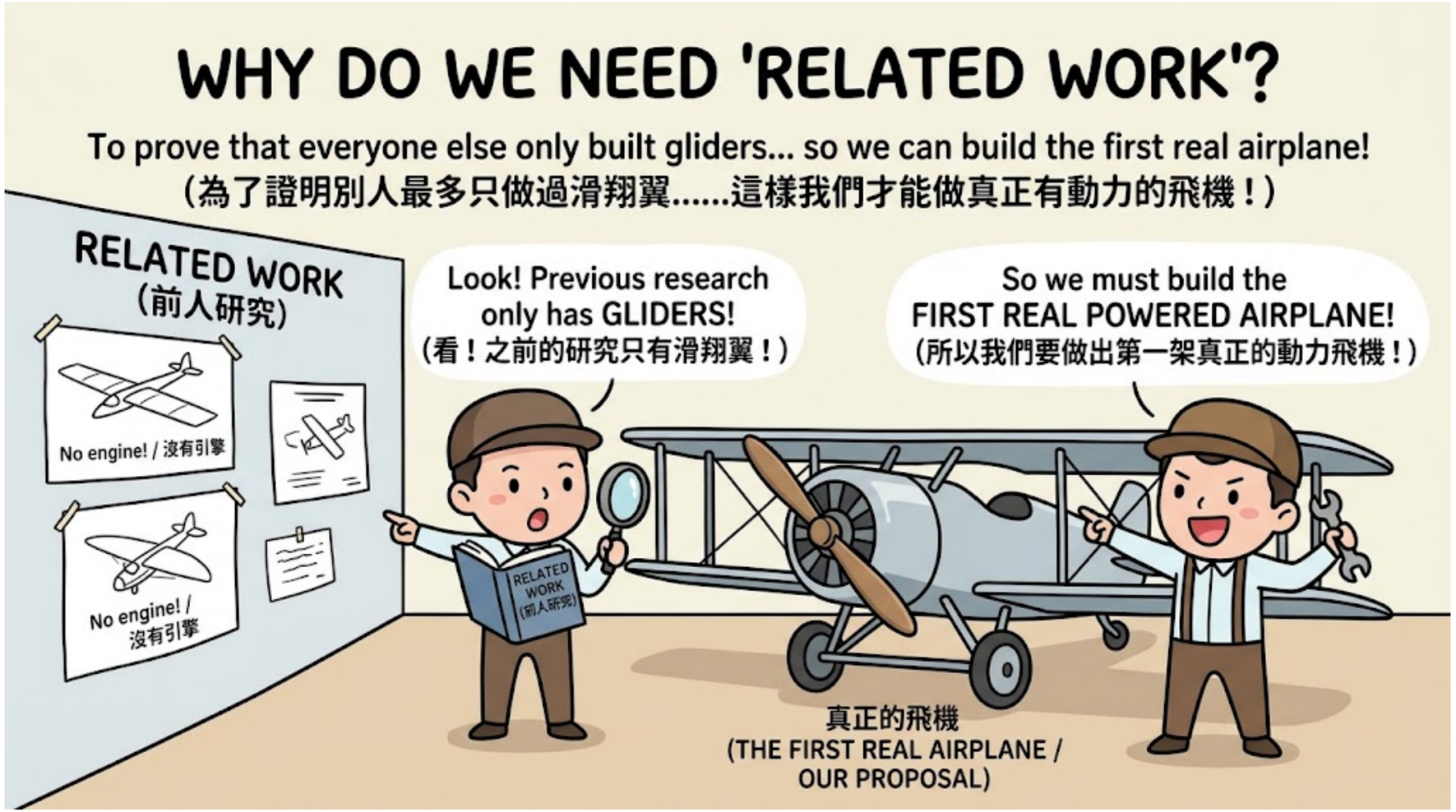
- You should care fully check the references and write the relevant descriptions.
  - Smith et al. (2020) proposed a pre-trained large language model.
    - If reviewers are familiar the paper of Smith et al. (2020), they will check the relevance.

Example of Step 3

Liu, Zhuang, et al. "A convnet for the 2020s." 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, 2022.

Recent convolution-based approaches. Han et al. [25] show that local Transformer attention is equivalent to inhomogeneous dynamic depthwise conv. The MSA block in Swin is then replaced with a dynamic or regular depthwise convolution, achieving comparable performance to Swin. A concurrent work ConvMixer [4] demonstrates that, in small-scale settings, depthwise convolution can be used as a promising mixing strategy. ConvMixer uses a smaller patch size to achieve the best results, making the throughput much lower than other baselines. GFNet [56] adopts Fast Fourier Transform (FFT) for token mixing. FFT is also a form of convolution, but with a global kernel size and circular padding. Unlike many recent Transformer or ConvNet designs, one primary goal of our study is to provide an in-depth look at the process of modernizing a standard ResNet and achieving state-of-the-art performance.

Related Work is complete “Only if you have Step 3”



Step 3: Compare and point out what others “cannot”

Lin, Ying-Jia, et al. "Cfever: A chinese fact extraction and verification dataset." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

“However, these approaches all rely on structured diagnosis codes, which are not available until after discharge coding is complete.”

“In contrast, we follow FEVER (Thorne et al. 2018a) to provide a fixed fact database and ask models to first extract relevant documents from Wikipedia before verifying the claims with the evidence sentences. We consider our approach, along with FEVER (Thorne et al. 2018a), to be more realistic for real- world fact verification scenarios.”

Another writing strategy: find the current best

Despite these challenges, progress has been made. RadFM (Wu et al., 2023) supports both 2D and 3D images, focusing on text generation tasks like report generation, rationale generation and VQA. Another significant contribution is CT-CLIP (Hamamci et al., 2024a), introducing the open-source dataset CT-RATE, which consists of over 25,000 chest CT volume-report pairs. CT-CLIP demonstrates that contrastive learning can align CT images with reports using a 3D vision encoder and text encoder, similar to CLIP. Both CT2Rep (Hamamci et al., 2024b) and 3D-CT-GPT (Chen et al., 2024) employ the 3D vision encoder architecture from CT-CLIP as part of its chest CT report generation model. Unlike CT2Rep that leverages a simple transformer decoder for report generation only, 3D-CT-GPT leverages LLM and perform both report generation and VQA. The M3D-LaMed model (Bai et al., 2024) further advances the field by integrating visual information into LLM through a perceiver module, enabling tasks like captioning, VQA, and segmentation. However, M3D-LaMed still relies on a 3D Vision Transformer (ViT) (Dosovitskiy, 2020) with fixed patch embedders, limiting its ability to capture detailed visual information, and retaining challenges in representing complex 3D medical structures effectively.

RadFM < CT-CLIP < CT2Rep < 3D-CT-GPT < M3D-LaMed

How long and how many papers?

- Two to four paragraphs, one per family, 400–700 words in total
- Roughly 20–40 citations for a conference paper; fewer if the area is small
- Most citations appear in brackets, grouped two or three to a sentence
- One paper gets extended treatment: the closest one

If you are over 700 words, you are probably writing a survey. If you are under 200, you probably have not found your groups yet.

Sentence Bank of Related Work

- "Existing work falls into two groups: … and …"
- "A first line of work … (A, 2020; B, 2021)."
- "These approaches share one assumption: …"
- "However, all of them rely on …, which is not available when …"
- "The work closest to ours is …"
- "There are two main differences between … and our work. First, … Second, …"
- "To our knowledge, no previous work has …"

Your Turn

1. Group: sort your references into two to four families, by the approach

2. Describe: one clause per citation

3. Compare: close each family with what it cannot do

NEXT WEEK ATTENTION!!

- Next week: **Results** section: Tables
  - Paragraph for a table
  - Caption for a table
- Please bring your preliminary results.

Results are not ready?

- If your experiments are still running, bring the table with empty cells and the row and column headings filled in. That is enough to work on.
- If you don’t have your results ready, please run small experiments this week.

Thank you for your participation 📚