

Large Language Models for Alzheimer's Disease

Cognitive AI for Science (CogAI4Sci) Laboratory, NUS

Boneshwar V K

Indian Institute of Technology, Madras

19 October 2024

GitHub: bs20b012 **Email:** bs20b012@smail.iitm.ac.in

- ① Introduction
- ② Literature Survey Overview
- ③ DALK
- ④ BioMedGPT
- ⑤ AD-AutoGPT
- ⑥ Med-PaLM
- ⑦ Conclusion

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT
- 5 AD-AutoGPT
- 6 Med-PaLM
- 7 Conclusion

Introduction

- Alzheimer's disease is a major global health concern due to its prevalence as the most common cause of dementia, affecting millions of people worldwide, particularly those over 65. There is currently no cure, but ongoing research aims to develop effective treatments that can slow disease progression and improve quality of life for patients and their caregivers.
- Large language models (LLMs) can assist in understanding and managing Alzheimer's by analyzing vast amounts of medical data to support diagnosis, treatment planning, and caregiver resources.
- Large Language Models have shown their promising performances in generic tasks, leveraging the large knowledge of LLMs, several works have been enforced in recent times for domain-specific tasks.

Introduction

- The medical field needs advanced knowledge tools like large language models (LLMs) to advance. These models must be tailored to the specific area of medicine to prevent mistakes, as errors in this field can have serious consequences.
- Here we review a few of the reputed papers relevant to our research topic
 - ① DALK: *Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature*[1].
 - ② BiomedGPT: *A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks*[2].
 - ③ AD-AutoGPT: *An Autonomous GPT for Alzheimer's Disease Infodemiology*[3].
 - ④ Med-PaLM: *Language Models Encode Clinical Knowledge*[4].

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT
- 5 AD-AutoGPT
- 6 Med-PaLM
- 7 Conclusion

Literature Survey Overview

DALK

- Leverages the concept of incorporating LLMs for Knowledge Graph generation and Knowledge Graphs into LLMs for better reasoning.
- The KGs are being constructed and utilized for reasoning dynamically.

BioMedGPT

- The crux of this paper is to provide a generalist model designed for handling a wide range of biomedical tasks, from medical imaging to text summarisation.
- Leverages a BERT-style encoder over corrupted text and a GPT-style left-to-right auto-regressive decoder.

Literature Survey Overview (contd.)

AD-AutoGPT

- Leverages AutoGPT, an AI agent that automates the process of complex tasks by breaking them down into smaller steps, solving each subtask, and executing decisions without requiring human intervention.
- The tool streamlines and highlights its ability to automate complex data processing tasks, including the extraction of key themes from news articles and analyses of Alzheimer's Disease narratives which can be further adapted for broader applications.

Literature Survey Overview (contd.)

Large Language Models Encode Clinical Knowledge

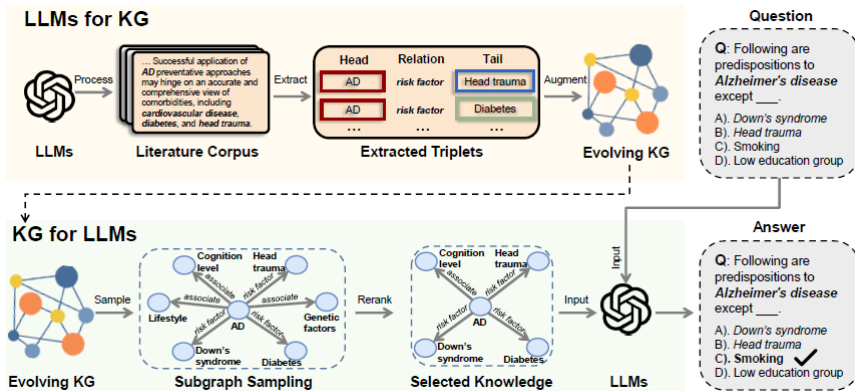
- The paper introduces a medical domain specific model titled Med-PaLM[4], a model released by Google Research & Deep Mind. inspired by PaLM[5] and Flan-PaLM[6].
- This paper also enlightens us about the dataset titled MultiMedQA, a new benchmark designed for this evaluation, and demonstrate that Flan-PaLM achieves state-of-the-art performance on multiple medical datasets, significantly surpassing previous models.

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK**
- 4 BioMedGPT
- 5 AD-AutoGPT
- 6 Med-PaLM
- 7 Conclusion

DALK: Understanding DALK

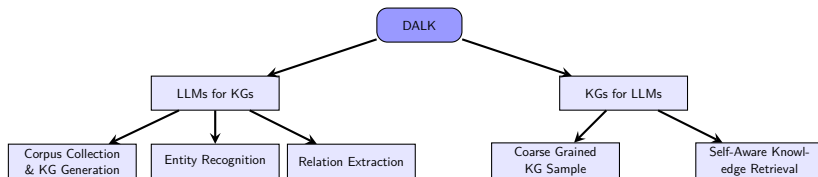
- **Summary:** The paper contributes by constructing Knowledge Graphs highly specific to Alzheimer's disease, developing a self-aware knowledge retrieval system that removes the noise and irrelevant data by re-ranking and filtering out irrelevant triples dynamically and thereby showing that LLM's performance increases with more focus on AD where the KGs play a vital role.
- **Research Question:** How can the integration of LLMs and dynamically evolving Knowledge Graphs with a self-aware knowledge retrieval system improve the performance of LLMs in answering AD domain-specific questions where the common strategies like LLMs integrated with RAGs fail due to the limitations in long-tail memory, domain-specific knowledge, data quality, and efficiency & scale issues where re-training the LLMs from scratch is not feasible.

DALK: Understanding DALK (contd.)



DALK: Understanding DALK-Proposed Solution

- **Proposed Solution:** The proposed solution can be classified broadly into two categories as stated in the original paper, LLMs for KG and KG for LLMs.



DALK: LLMs for KG Construction

Corpus Collection

- The AD-specific corpus is created from 9,764 articles, focusing on post-2011 publications to ensure recent and relevant knowledge.
- Knowledge graph (KG) generation is based on this domain-specific corpus, following methodologies like those proposed by Pu et al. (2023).

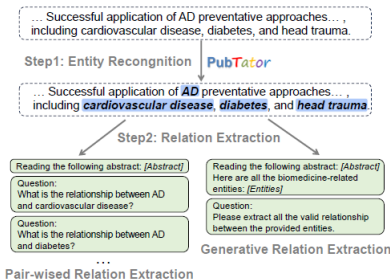
Entity Recognition

- Entities such as genes, diseases, chemicals, and mutations are extracted using PubTator Central (PTC), a tool from NCBI.
- Extracted entities are used as nodes in the Alzheimer's Disease (AD) knowledge graph.

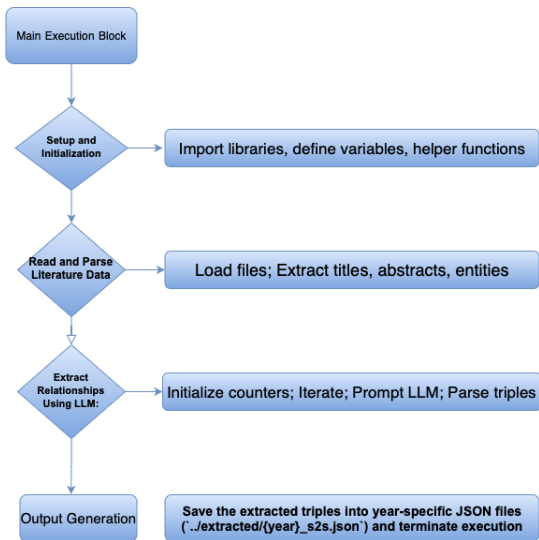
DALK: LLMs for KG Construction (contd.)

Relation Extraction

- Pair-wised Relation Extraction prompts LLMs to describe relationships between two entities in a text segment.
- Generative Relation Extraction allows LLMs to output entity pairs and their relationships directly, creating two versions of knowledge graphs, KG_{pair} and KG_{gen} .



DALK: LLMs for KG Construction-Flowchart of LLM4KG



DALK: KGs for LLMs reasoning

Coarse-grained KG Sample

- In the coarse-grained knowledge sampling process, given a query Q , the large language models (LLMs) first extract domain-specific entities $E = \{e_1, e_2, \dots\}$. These entities are linked to an evolving knowledge graph G using semantic similarity models to create embeddings H_G and H_E , and links are established based on cosine similarity.
- This process produces an initial entity set E_G , which is then used to build evidence subgraphs for boosting the LLM's reasoning ability.

DALK: KGs for LLMs reasoning (contd.)

Self-aware Knowledge retrieval

- The self-aware knowledge retrieval method addresses the noise in sampled knowledge by prompting the LLM to re-rank and filter out irrelevant information. The template

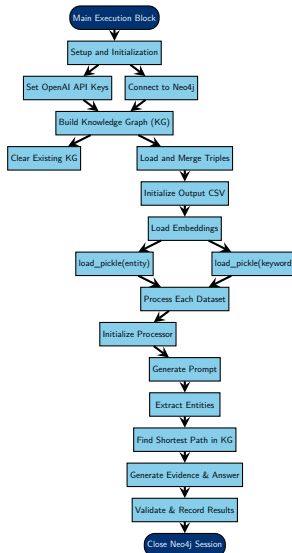
$$p_{self} = Template_{self}(Q, G_Q, k) \quad (1)$$

is used to prompt the LLM to retrieve the top k triples relevant to a query. The final inference is done using

$$p_{inference} = Template_{inference}(Q, G_Q^{self}), \quad (2)$$

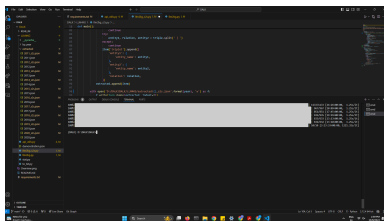
allowing the LLM to generate answers based on the most relevant knowledge, ensuring better accuracy in the reasoning process.

DALK: KGs for LLMs reasoning: Flowchart of KG4LLM

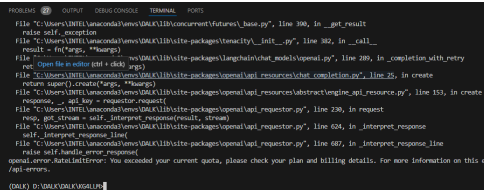


DALK: Implementation Results

DALK Source code was utilised to recreate the results. The repo basically has two python files for the two key concepts that were discussed earlier-LLM4KG & KG4LLM. Knowledge Graph Generation was successful but the LLM reasoning using KGs failed due to lack of Neo4j subscription. For details refer to my GitHub and the snippets shared below.



(a) KG Generation using ILM2kg



(b) LLM Reasoning

图 1: Implementation Results

DALK: Future Directions and Potential Improvements Mentioned in Paper

- **Entity Recognition:** Replacing PubTator by LLMs. Also another way to extract entities would be from images using either LLMs or OCR techniques.
- **ADQA Benchmark Limitations:** Alzheimer's Disease Question Answering (ADQA) benchmark primarily includes questions derived from medical school exam datasets, which may not fully represent the depth of scientific literature. The authors suggest future work should expand this dataset to include more AD-specific questions more like from PubMedQA.
- **KG Construction:** There's a limitation in the KG construction method where the relation extraction between entities is not as accurate in the **pair-wise** method compared to the **generative** method.

DALK: Future Directions and Potential Improvements-Others

- **Handling Noisy Data:** Handling of noise using self aware knowledge retrieval is built upon existing KGs. Refining the KGs by *context aware retrieval* would reduce the noise.
- **Scaling and Efficiency:** Scaling LLMs using techniques like LoRA, Q-LoRA, replacing by lighter LLMs would be efficient.
- **Interactive Feedback Mechanism & Evaluation Metrics :** The paper focuses on the static retrieval of the most relevant triples and answers. Incorporating an interactive feedback loop where the system can receive corrections or updates from domain experts in real-time and using different evaluation metrics should add more interpretability.

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT**
- 5 AD-AutoGPT
- 6 Med-PaLM
- 7 Conclusion

BioMedGPT: Understanding BioMedGPT

- **Summary:** This paper introduces BioMedGPT which is claimed to be the first open-sourced Vision-Language model for generic healthcare applications. Leveraging BERT-style encoder and GPT-style decoder??, this model outperforms SOTA models in different tasks such as radiology visual question answering, report generation, and summarisation.
- **Research Question:** This paper tries to bridge the gap between Generalist AI solutions and biomedical applications by developing the model BioMedGPT to understand and interpret diverse data types and provide customized outputs. By doing so, BioMedGPT addresses the challenges posed by continuously evolving data and the integration of multimodal data such as scans and medical reports.

BioMedGPT: Understanding BioMedGPT

- **Proposed Solution:** The aim is to develop an open-source generalist AI model, BioMedGPT, for biomedical sciences, capable of capturing the complexities of multimodal datasets. BioMedGPT combines a BERT encoder with a GPT-style left-to-right autoregressive decoder, outperforming models that rely solely on encoders or decoders in learning joint representations and aligning input-output across modalities.
- To handle diverse modalities without relying on task-specific output structures, unifying the embeddings obtained from VQGAN, BPE Tokenizer, Pix2Seq Tokenizer for Masked Language Modelling, text tokenizing and Bounding Box tokenizing respectively. Here they incorporate the position embeddings using a decoupling method to separate position correlation to avoid unnecessary randomness in the attention.

BioMedGPT: Understanding BioMedGPT

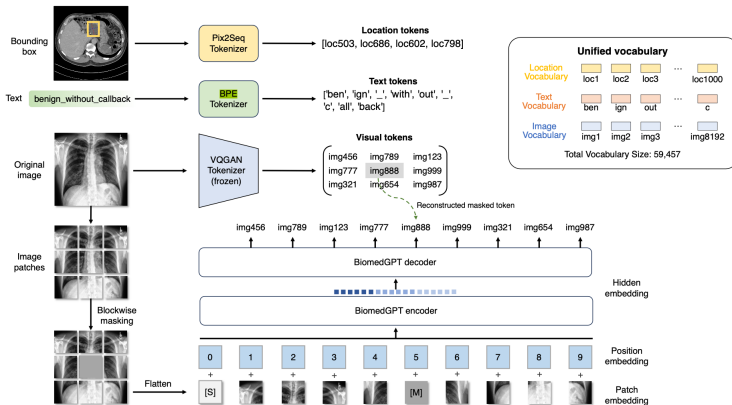


图 2: BioMedGPT Tokenization to form Unifying Vocabulary

BioMedGPT: Understanding BioMedGPT

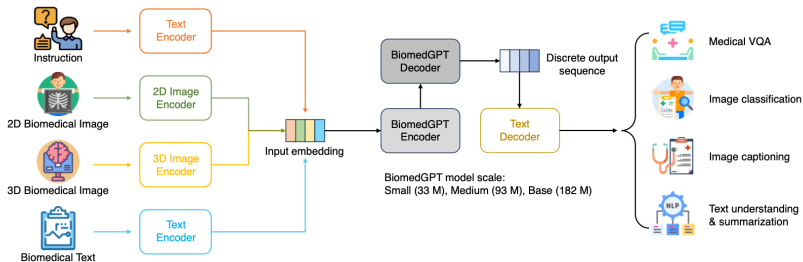


图 3: BioMedGPT Architecture

BioMedGPT: Tasks, Datasets, and Evaluation Metrics

The following table summarises all datasets and their respective evaluation metrics used in this paper.

Task	Datasets Used	Evaluation Metrics
Visual Question Answering (VQA)	VQA-RAD, SLAKE, PathVQA	Accuracy, Weighted F1 Score
Image Captioning	IU X-Ray, PEIR GROSS, MIMIC-CXR	ROUGE-L, METEOR, CIDEr
Image Classification	MedMNIST-Raw, CBIS-DDSM, MC-CXR	Accuracy, F1 Macro Score
Report Summarization	MIMIC-CXR, HealthCareMagic	ROUGE-L, F1 Score
Text Summarization	MIMIC-CXR, HealthCareMagic	ROUGE-L, F1 Score
Natural Language Inference (NLI)	MedNLI	Accuracy, F1 Macro Score
Clinical Trial Matching	TREC 2022	Accuracy, Mean Average Precision (MAP)

 4: BioMedGPT Tasks, Datasets and their Evaluation Metrics

BioMedGPT: Model Fine-Tuning, Pre-training and Zero Shot Inference

- **Model Pre-Training:** For a model being parameterized θ it is trained by minimising the following equation 3. In BioMedGPT, input \mathbf{x} , \mathbf{x} includes both linguistic and visual tokens, such as subwords, image codes, and location tokens, used during pre-training. Subwords, generated by a BPE tokenizer, have 15% of their tokens masked for masked language modeling, given the high overlap among medical terms. For object detection, location tokens are created using Pix2Seq, based on pixel inputs.
- Hyperparameters used for Pre-training the BioMedGPT are- AdamW optimizer, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e-}8$, dropout = 0.1 and learning rate = $1\text{e-}4$.

$$L_{\theta}(\mathbf{x}_{1,1}, \dots, \mathbf{x}_{i,b}) = -\sum_{b=1}^B \log \prod_{i=1}^I p_{\theta}(\mathbf{x}_{i,b} | \mathbf{x}_{1,b}, \dots, \mathbf{x}_{i-1,b}) = -\sum_{b=1}^B \sum_{i=1}^I \log p_{\theta}(\mathbf{x}_{i,b} | \mathbf{x}_{<1,b}). \quad (3)$$

BioMedGPT: Model Fine-Tuning, Pre-training and Zero Shot Inference (contd.)

- Biomedical images require preprocessing to remove trivial elements like black backgrounds and adjust input size. The images are cropped to the object's bounding box, resized to 256×256 , and the central 128×128 region is fed into a pre-trained VQ-GAN to generate sparse image codes for masked image modeling. The same tokenization approach applies to vision-language tasks. Fine-tuning involves seq2seq learning, adapted for different datasets and tasks.

BioMedGPT: Model Fine-Tuning, Pre-training and Zero Shot Inference (contd.)

- **Fine-Tuning:** BioMedGPT fine-tunes its pre-trained structure without adding extra components, aligning with the pre-training workflow. Hyperparameters like beam search size and output length are optimized based on task needs and training data. Best-performing checkpoints are used for inference, with K-fold cross-validation applied when official splits are unavailable.
- Trie-based beam search improves generation quality by restricting token candidates, preventing invalid outputs, and enhancing efficiency. This method also accelerates validation during fine-tuning, boosting speed by up to 16x.

BioMedGPT: Model Fine-Tuning, Pre-training and Zero Shot Inference (contd.)

- **Zero Shot Inference and Model Instruction-tuning:**
Generally the model instruction involves a set of pre-defined answers just like the one found in Llava-Med. This paper utilizes an open-vocabulary setting, allowing the model to operate without a predefined set of answer.
- Several medical datasets used for zero-shot experiments: RSNA Pneumonia Detection Challenge, MedMNIST v2, BreastMNIST, DermaMNIST, OCTMNIST, PathMNIST, and MIMIC-CXR [7]. BiomedGPT's performance was tested across tasks like pneumonia detection, tumor identification, and report generation, with comparisons to Med-PaLM M. Preliminary studies with instruction-tuned models highlighted gaps in in-context learning, emphasizing the need for further research in medical AI.

BioMedGPT: Evaluation Metrics

表 1: Evaluation Metrics for BiomedGPT

Metric	Task	Formula
Accuracy	Image Classification, VQA, NLI	$\frac{TP+TN}{TP+TN+FP+FN}$
F1 Score	Classification with Class Imbalance	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
Weighted F1	VQA with Weighted Classes	Weighted F1 = $\sum_i \frac{n_i}{N} \times F1_i$
F1-Macro	CBIS-DDSM Classification	F1-Macro = $\frac{1}{N} \sum_i F1_i$
ROUGE-L	Text Summarization, Captioning	ROUGE-L = $\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}}$
METEOR	Text Generation Quality	METEOR = $(1-p) \frac{PR}{\alpha P + (1-\alpha)R}$
CIDEr	Image Captioning	CIDEr(c, S) = $\frac{1}{M} \sum_i \frac{g_i(c) \cdot g_i(S_i)}{\ g_i(c)\ \ g_i(S_i)\ }$

BioMedGPT: Implementation Results

The author provided a Colab notebook for testing Fairseq-free inference with BioMedGPT-base, as shown in the fig.34. However, tasks like pre-training and fine-tuning were not executed due to limited computational resources and the absence of a pre-configured Linux environment. The model understands modality data but does not efficiently answer domain specific questions like the one shown below.

```
img = Image.open('/content/chestray.png')
txt = "what modality is used here?"
inputs = tokenizer(txt), return_tensors="pt").input_ids
patch_img = patch_resize_transform(img).unsqueeze(0)

gen = model.generate(inputs, patch_images=patch_img, num_beams=5, no_repeat_ngram_size=3, max_length=16)
results = tokenizer.batch_decode(gen, skip_special_tokens=True)

result = results[0]
result = re.sub(r'["\s]', '', result).strip()

result
```

'XRay'



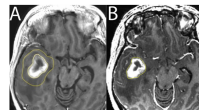
```
img = Image.open('/content/2_img')
txt = "Does this MRI image show any signs of having brain tumor?"
inputs = tokenizer(txt), return_tensors="pt").input_ids
patch_img = patch_resize_transform(img).unsqueeze(0)

gen = model.generate(inputs, patch_images=patch_img, num_beams=5, no_repeat_ngram_size=3, max_length=16)
results = tokenizer.batch_decode(gen, skip_special_tokens=True)

result = results[0]
result = re.sub(r'["\s]', '', result).strip()

result
```

'anterior cingulate cortex'



BioMedGPT: Future Directions and Potential Improvements mentioned in the paper

- In-context learning abilities and text comprehension.
- Expanding the modality to 3D scans and other medical imaging classification problem statements.
- Incorporation of new metrics for evaluation such as F1-RadGraph.
- Expanding solutions to time-series and sequential data.
- Data imbalances, the dataset contents are more orientated towards Radiology hence creating an imbalance in the model's learning.

BioMedGPT: Future Directions and Potential Improvements-Others

- Incorporating ***speech transcripts*** to diagnose dementia diseases such as Alzheimer's disease.
- Incorporating multi-modality data such as 3D MRI scans with genetic data such as gene expression profiles-scRNA sequenced data, bulk RNA sequenced data. In fact using a separate pipeline to remove noisy data by ***ML based imputation*** and then inferring the gene expression patterns using graph based methods could potentially give us some insights on the disease.
- New evaluation metrics based on the specific use case could add some interpretability of the model's performance.

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT
- 5 AD-AutoGPT**
- 6 Med-PaLM
- 7 Conclusion

AD-AutoGPT: Prelims

- **AutoGPT:** AutoGPT is not just a standalone model; it's an exciting experiment that harnesses the incredible power of GPT-4/GPT-3. By leveraging the capabilities of this Large Language Model (LLM), AutoGPT aims to automate various tasks. It compiles a list of instructions from the LLM and executes them, often revolving around programming and implementing logical steps.
- Automation tasks performed by AutoGPT are Code execution, Google search, Web browsing, web scrapping, File operation and even Twitter integration.

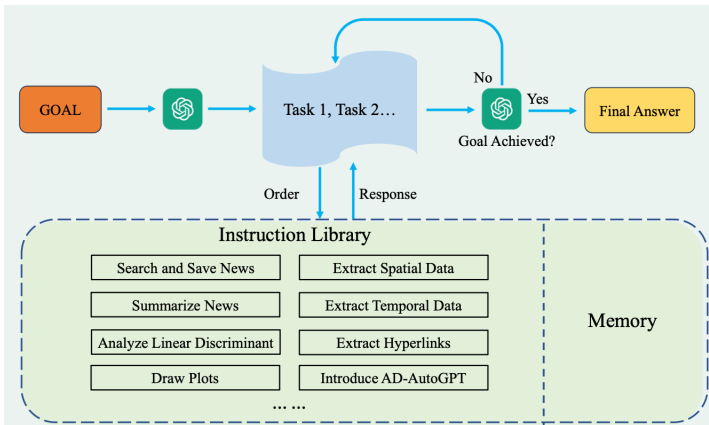
AD-AutoGPT: Understanding AD-AutoGPT

- **Summary:** This paper leverages AuotGPT, built on Langchain framework and integrated with GPT-3.5/4, following ***chain of thoughts*** to break the whole process into sub-parts and accomplish the these sub-tasks iteratively.
- **Research Question:** Can an autonomous LLM-based tool efficiently collect, process, and analyze health-related information, overcoming the limitations of manual data handling and improving understanding of the complex narratives surrounding Alzheimers's Disease?

AD-AutoGPT: Understanding AD-AutoGPT

- **Proposed Solution:** The proposed solution, AD-AutoGPT, effectively addresses the research question by leveraging LLMs for automated health data collection, spatiotemporal analysis, topic modeling, and dynamic visualization. It offers a scalable, efficient, and autonomous approach to enhance public health research on Alzheimer's Disease, paving the way for future AI-assisted infodemiology efforts.
 - **Automation of Infodemiology:** AD-AutoGPT transforms traditional manual health data collection and analysis by automating the entire pipeline, making it faster, more accessible, and less labor-intensive.
 - **Overcoming LLM Limitations:** Token limitations of LLMs are overcome through techniques like map-reduce summarization and spatiotemporal extraction
 - **Facilitating Public Health Research:** By visualizing trends and automatically identifying hot topics.

AD-AutoGPT: Pipeline



 5: AD-AutoGPT Pipeline

AD-AutoGPT: Flowchart

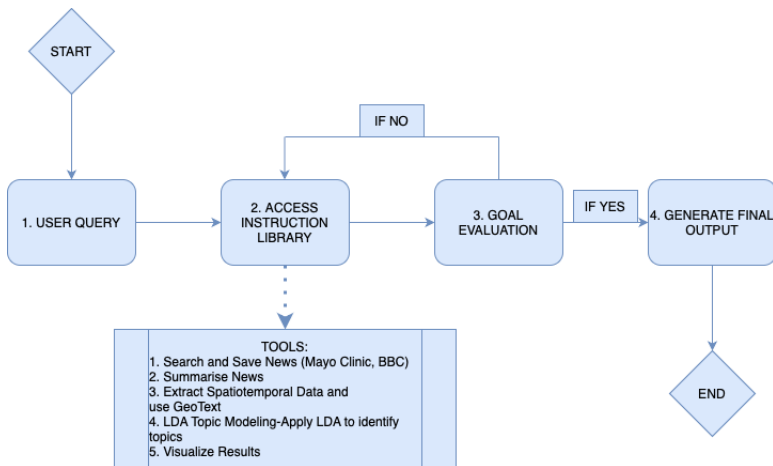


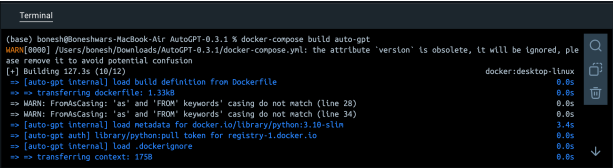
图 6: Flowchart of AD-AutoGPT Functionalities

AD-AutoGPT: Future Directions and Potential Improvements

- The authors have stated that future improvements could involve adding data from scientific journals, electronic health records (EHRs), and clinical trial databases.
- Expanding the system to support multiple languages would enable the analysis of global health narratives, especially in regions where AD research and awareness are evolving.
- Real-time Monitoring and prediction along with integration of more advanced topic modeling techniques, such as BERT-based topic models to capture evolving trends more accurately over time.
- A bias detection module could be incorporated to ensure the tool provides balanced insights, avoiding misinformation and discriminatory outputs.
- Emphasizing on ethics and Federated Learning for privacy.

AD-AutoGPT: Implementation Results

AutoGPT installation was completed by using source code file from the official github repository. Open AI and Google API keys were set in the `.env` file which was modified from `.env.template` file. AutoGPT installation along with Google Search Engine creation was completed. Further steps would to utilise AutoGPT in a langchain framework to replicate the AD-AutoGPT framework.



```
Terminal
(base) bonesh@Boneshwars-MacBook-Air: AutoGPT-0.3.1 % docker-compose build auto-gpt
WARN[0000] /Users/bonesh/Downloads/AutoGPT-0.3.1/docker-compose.yml: the attribute 'version' is obsolete, it will be ignored, please remove it to avoid potential confusion
[+] Building 127.3s (10/12)
=> [auto-gpt internal] load build definition from Dockerfile
=> => transferring dockerfile: 1.33kB
=> WARN: FromAsCasing: 'as' and 'FROM' keywords' casing do not match (line 28)
=> WARN: FromAsCasing: 'as' and 'FROM' keywords' casing do not match (line 34)
=> [auto-gpt internal] load metadata for docker.io/library/python:3.10-slim 3.4s
=> [auto-gpt auth] library/python:pull token for registry-1.docker.io 0.0s
=> [auto-gpt internal] load .dockerignore 0.0s
=> => transferring context: 175B 0.0s
```

 7: AutoGPT installation using docker

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT
- 5 AD-AutoGPT
- 6 Med-PaLM**
- 7 Conclusion

Med-PaLM: Understanding Med-PaLM

- **Summary:** Flan-PaLM is an enhanced version of Google's PaLM model, fine-tuned to follow instructions more effectively and perform a broad range of tasks with exceptional accuracy. It particularly excels in multi-step reasoning, making it well-suited for handling complex fields such as medicine. With its release, Flan-PaLM has set new benchmarks in natural language processing, especially in scenarios where only minimal or no prior examples (few-shot and zero-shot learning) are provided.
- Building on Flan-PaLM's capabilities, Med-PaLM takes things a step further by tailoring the model specifically for medical use. Through specialized prompt tuning focused on healthcare needs, Med-PaLM ensures safer, more reliable responses aligned with the standards expected from medical professionals.

Med-PaLM: Understanding Med-PaLM

- **Research Question:** How well can large language models (LLMs) encode and apply clinical knowledge, and how can these models be aligned to perform safely and effectively in medical question-answering tasks?
- **Proposed Solution:** Introducing a new medical benchmark that integrates both professional and consumer medical queries—MultiMedQA & HealthSearchQA.
- Developing Med-PaLM, a specialized instruction-tuned model aligned with medical needs through prompt tuning.
- Addressing the critical question of trust, bias, and safety in deploying LLMs for healthcare applications—an area largely unexplored in previous research.

Med-PaLM: Pipeline of Med-PaLM

- **Human Evaluation Framework:** Proposes a pilot human evaluation framework where clinicians and non-experts (Lay users) assess model responses along dimensions such as factual alignment, bias, and potential harm.
- The goal of this evaluation was to assess if there are any crucial information being omitted that should not be missed out in the first place.

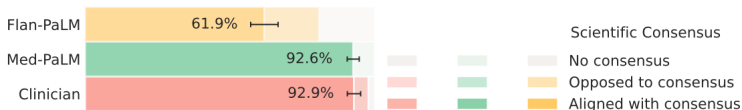


图 8: Human Evalutaion Comparison

Med-PaLM: Performance of all State-of-the-art LLMs

Model	Dataset	Parameters	Accuracy (%)	Notable Findings
Med-PaLM	MedQA (USMLE)	540B	67.6	Comparable with clinicians in terms of scientific consensus (92.6%). Reduces harmful outputs to 5.8%. Specialized using instruction prompt tuning with
Flan-PaLM	MedQA (USMLE)	540B	67.6	Surpassed PubMedGPT by 17%. Achieved state-of-the-art in multiple-choice questions .
Flan-PaLM	MedMCQA	540B	57.6	Outperformed Galactica (120B) by 4.7%. Effective on Indian medical exam questions .
Flan-PaLM	PubMedQA	540B	79.0	Matches human-level accuracy (78%). Outperforms BioGPT by 0.8%.
PubMedGPT	MedQA (USMLE)	2.7B	50.3	Trained on biomedical abstracts . Focused on biomedical-specific reasoning.
Galactica	MedMCQA	120B	52.9	High performance on scientific reasoning tasks . Comparable to Flan-PaLM on certain datasets.
DRAGON	MedQA (USMLE)	360M	47.5	Earlier state-of-the-art before Flan-PaLM. Shows good multi-step reasoning capabilities.
BioLinkBERT	MedQA (USMLE)	340M	45.1	Focused on biomedical information retrieval .
GPT-Neo	MedQA (USMLE)	2.7B	33.3	Limited to basic multiple-choice performance .
PubMedBERT	MedQA (USMLE)	100M	38.1	Optimized for biomedical abstracts retrieval , but underperforms on multi-task benchmarks.

图 9: Summarising models and their results

Med-PaLM: Future Directions and Potential Improvements mentioned in the paper

- The authors suggest to perform fairness and equity check on the training data which might affect the model's performance.
- Dynamically evolving datasets and pipeline to include that into the pipeline would avoid the outdated outputs reflected by the model. Also expanding the benchmark to datasets involving medical records or real-time patient data to test model performance in clinical workflows could be one more add-on to the whole pipeline enhancing the model's performance.
- Handling Uncertainty and Trust in LLM Outputs The authors suggest that LLMs need better mechanisms to communicate uncertainty when providing answers, especially in safety-critical situations. Further research could focus on uncertainty quantification methods to help clinicians decide when to trust a model's response or defer to human expertise.

Med-PaLM: Future Directions and Potential Improvements-Others

- **Explainability and Transparency of Model Outputs:**
Future research could focus on building interpretable models that provide clear reasoning for each answer, similar to a clinical justification.
- **Multilingual Medical Models and Domain Specific datasets:** This method could potentially be used all over the world and help technicians and doctors of a particular domain as well.
- **Knowledge Graphs and abstracted data storage for better reasoning:** Incorporating KGs could potentially enhance the interpretability and the model's reasoning. Adding RAGs or Retrieval Augmented Generation on top of KGs could potentially be overkill but for larger datasets it could prove to be handy.

- 1 Introduction
- 2 Literature Survey Overview
- 3 DALK
- 4 BioMedGPT
- 5 AD-AutoGPT
- 6 Med-PaLM
- 7 Conclusion**

Conclusion

Each paper contributes essential elements toward building a domain-specific LLM for Alzheimer's Disease (AD):

- **Agentic Capabilities (AD-AutoGPT):** Leverages autonomous agents for multi-step reasoning, refining outputs with sources like PubMed or EMRs.
- **Multimodal Knowledge Integration (BioMedGPT):** Combines text, clinical notes, cognitive tests, and imaging data (MRI, PET) for diagnosis through multi-task learning.
- **Instruction Prompt Tuning (MedPaLM):** Ensures safe outputs using AD-specific datasets (e.g., MMSE) aligned with medical standards.
- **Contextual Reasoning with Knowledge Graphs (DALK):** Tracks disease progression with dynamic graphs, adjusting recommendations over time.

Conclusion: Approaches

After reviewing these papers and analyzing their core concepts, the following one-liner encapsulates my proposed approach for developing a novel pipeline for Alzheimer's Disease.

- **Multimodal Pipeline with Separate Models for Imaging and Text, Coordinated by an LLM Agent:** Here we use separate models supervised by a LLM agent to overcome the domain specific issue faced by generalist models like BioMedGPT. Using such separate models aid in making the model to understand very complex data like scRNA seq data.
- **RAG-Enabled LLM with Real-Time Clinical Trial Retrieval and Personalized Recommendations:** The core idea is to use Knowledge Graph that are evolving dynamically inspired by DALK and RAG to retrieve the relevant triples, URLs, images & other multi-modal data and reason behind it.

Conclusion: Approaches

- **Two Stage Pipeline for Early Detection and Advanced Disease Management:** Two stage design ensures accurate early detection and robust long-term management.
Stage 1: MRI analysis focuses on early-stage diagnosis using Vision Transformers to detect subtle neurodegenerative changes.
Stage 2: The focus shifts to text-based tracking of symptoms, behavior, and medication adherence as the disease progresses, with LLMs managing personalized care plans.

References

- [1] Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sunkwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, et al.
Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature.
arXiv preprint arXiv:2405.04819, 2024.
- [2] Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al.
Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks.
arXiv preprint arXiv:2305.17100, 2023.
- [3] Haixing Dai, Yiwei Li, Zhengliang Liu, Lin Zhao, Zihao Wu, Suhang Song, Ye Shen, Dajiang Zhu, Xiang Li, Sheng Li, et al.
Ad-autogpt: An autonomous gpt for alzheimer's disease infodemiology.
arXiv preprint arXiv:2306.10095, 2023.
- [4] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al.
Large language models encode clinical knowledge.
arXiv preprint arXiv:2212.13138, 2022.
- [5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al.
Palm: Scaling language modeling with pathways.
Journal of Machine Learning Research, 24(240):1–113, 2023.
- [6] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le.
Finetuned language models are zero-shot learners.
In International Conference on Learning Representations.