Hita Kambhamettu*
hitakam@seas.upenn.edu,
Yidi Huang*
Yidi.Huang@pennmedicine.upenn.edu,
Kevin Johnson
Kevin.Johnson1@pennmedicine.upenn.edu,
Angela Bradbury
Angela.Bradbury@pennmedicine.upenn.edu

Knowledge-Grounded Medical Dialogue Generation for Genetic Counseling Regarding Alzheimer's Risk

January 12, 2024

Springer Nature

*Indicates equal contribution

Contents

Part I
Manuscript

Chapter 1
Knowledge-Grounded Medical Dialogue Generation

Abstract With the introduction of the first FDA-approved treatment for Alzheimer's disease (AD), genetic testing for APOE, a major genetic risk factor for AD, has become a critical step to assess treatment eligibility. In order to address the increasing volume of APOE testing, tools to help patients understand genetic risk factors and their implications are urgently needed. Conversational agents powered by large language models (LLMs) can help triage patients and supplement human counselors. However, deploying such agents poses challenges: institutional barriers prevent the input of clinical data, including protected health information (PHI), into commercial LLMs, LLMs potentially hallucinate critical medical facts, and LLMs should mimic the communication style of clinicians in order to be trusted supplements. We introduce a dual-method approach to enhance LLMs' accuracy and clinical communication effectiveness. First, we build a knowledge bank of recorded patient-provider genetic counseling sessions and leverage an open-source LLM to extract and summarize relevant information. We leverage this knowledge bank to develop a retrieval-augmented system for answering patient questions. We find that responses generated from our pipeline are more readable and better resemble human responses compared to those directly from GPT-4, suggesting that this pipeline enhances both accuracy and a clinician-like tone of communication.
Keywords: Alzheimer's risk, large language models, genetic counseling, retrieval-augmented generation

1.1 Introduction

Alzheimer's disease (AD) represents a significant global challenge from medical, economic, and societal perspectives (Wimo et al. 2017). Genetic testing for risk-modifying variants, particularly at the APOE locus, can reveal an individual's innate risk for developing AD, and is usually undergone in tandem with genetic counseling (Thambisetty and Howard 2023). As AD awareness improves and new experimental therapies are developed, the demand for genetic testing will only increase, outpacing

supply and necessitating innovative ways of offering genetic counseling services. Conversational agents (CA) show promise as an interactive informational resource to address the increasing demand for genetic testing and the limited workforce of genetic providers (Walton et al. 2023, Zhou and Bickmore 2022, Al-Hilli et al. 2023). We emphasize that CAs are not intended to replace providers, but rather to triage patients who have unique needs and could benefit from a visit with a genetic provider from those who have common questions which could be sufficiently answered by a digital tool. Recent advances in large language models (LLMs) have led to massive leaps in the quality of computer-generated text, enabling the development of fluidly conversing CAs that exhibit higher-order reasoning abilities. While popular CAs such as ChatGPT are instructed to adopt the "friendly assistant" persona by default, prompt engineering can be used to control the tone, style, or content of responses.

For a conversational agent to effectively inform a patient of the risks associated with APOE genotype results, two core requirements arise: determining a patient's numerical and relative risk and conveying this information in a manner akin to a practicing clinician. Genetic counselors use various techniques to effectively and empathetically communicate sensitive and complex risk information. For example, to help patients interpret the numerical risk for AD, a counselor might invert the probability to highlight the chance of not developing AD. Drawing from human-centered computing principles, and more specifically, reflexive thematic analysis (Braun and Clarke 2019), we propose a framework that grounds medical dialogue in data gathered from genetic counseling sessions without directly sharing PHI with proprietary LLMs. Consequently, we pose the question: can such an approach enable LLMs to generate medical text that retains both content integrity and clinical stylistic integrity?

In this work, we develop a question-answering system tailored to address patients' inquiries concerning APoE genetic testing and implications of individual results, drawing insights from a comprehensive thematic analysis of genetic counseling sessions. This is a foundational step towards the larger goal of developing a CA for Alzheimer's disease counseling and other medical contexts. Furthermore, we investigate the feasibility of learning from semantic patterns identified within qualitative data instead of learning from personal health information.

1.2 Related Work

With the advancements of natural language processing (NLP), there has been an increasing focus on pre-training transformers using task-specific biomedical datasets, including datasets for medical question-answering. Models such as BioBERT and ClinicalBERT have showcased high performances across QA tasks (Lee et al. 2020, Yan and Pei 2022). In the medical dialogue domain, Liu et al. (2020) released a high-quality Chinese medical dialogue dataset containing 12 types of common Gastrointestinal diseases named MedDG, with more than 17K conversations. Li et al. (2021) developed an end-to-end variational Bayesian generative strategy to

generate medical dialogue by approximating posterior distributions over patient states and physician actions. Lin et al. (2021) proposed a low-resource medical dialogue-generating system along with a Graph-Evolving Meta-Learning (GEML) framework that learns to evolve the commonsense graph for reasoning disease-symptom connections.

Lehman et al. (2023) have underscored the efficacy of compact, domain-specific language models over broader, general-purpose counterparts, even when fine-tuned with limited annotated data. However, despite these strides, exemplified by domain-specific models like BioGPT and Med-PaLM, the challenge of hallucinations and biases persists (Luo et al. 2022, Singhal et al. 2022). This underscores the intrinsic limitations and uncertainties tied to relying solely on large language models as inherent knowledge bases.

Rather than fine tuning language models or developing probabilistic methods for medical text generation, our approach leverages the language understanding and modeling capabilities of existing transformers. Several significant challenges exist when attempting to implement LLMs in practical clinical settings. Models specific to the medical domain often utilize comparatively smaller-scale LLMs, which may present less accurate and robust representations. Additionally, the fine-tuning of even these smaller LLMs is both computationally demanding. To address these challenges, we contribute to emerging work exploring retrieval-augmented generation in medical settings with a particular focus on synergizing open and closed source LLMs and using qualitative analysis to better inform the retrieval-augmented generation.

1.3 Methods

We give a brief overview of our reflexive thematic analysis and cover the architecture of our dual-method approach using Llama-2 and GPT-4.

1.3.1 Data

In order to simulate conversations between genetic counselors and patients, our system was developed using transcripts of genetic counseling sessions from the Alzheimer's Prevention Initiative Generation Program (Generation study) (Langlois et al. 2019). The Generation study consisted of standardized counseling and disclosure sessions about apolipoprotein E (APOE) results in the context of Alzheimer's disease prevention. The research study was approved by a university Institutional Review Board. Patients were recruited during 2017-2020 as part of screening for the generation study 1 and generation study 2 AD prevention trials, and enrolled into the Alzheimer's Prevention Initiative with consent for recording sessions and use in secondary research.

The dataset consists of recorded disclosure sessions between a patient and a genetic counselor (n=40 sessions). Each session lasted approximately one hour and followed the structure outlined by the Generation study which includes topics such as background information about the APOE gene, insurance-related information, and possible modifiers of risk. We transcribed audio recordings and performed speaker diarization using WhisperX, an open-source toolkit for performing speech-to-text transcription and speaker detection.(Bain et al. 2023).

1.3.2 Thematic Analysis

We conducted a qualitative thematic analysis of transcripts from genetic counseling sessions to taxonomize the techniques used by genetic counselors when answering patient queries. The content of the excerpts used in data collection is derived from the Generation Study (Langlois et al. 2019). For example, if a patient asked what the difference between mild cognitive impairment and dementia was, we observed that a genetic counselor would distill the terms into "take-home messages", a few words that described each ailment. These techniques were subsequently used in prompts to guide the LLM's responses.

The thematic analysis consisted of an initial theme development during which two authors separately analyzed a subset of five counseling sessions and created themes based on techniques used repeatedly by genetic counselors, an axial coding pass where all authors reviewed the emergent themes and accompanying excerpts, and a pruning pass during which two external researchers, one of whom is a genetic counselor, conducted a post-hoc audit of the analysis to remove duplicates or misclassified excerpts. The results from the thematic analysis are shown in 1.3.2. Despite its relatively low frequency, as indicated by the singular count in the 'Deferring to specialists' theme, we emphasize its critical importance, especially in the context of chatbot interactions. This theme underscores the necessity of integrating a safeguard mechanism within chatbot systems, ensuring that complex or sensitive inquiries are appropriately redirected to human specialists.

1.3.3 Generation Pipeline

Our response generation pipeline follows a retrieval-augmented generation pattern using LLMs (Lewis et al. 2020). A schematic of our approach can be seen in Figure 1.1. Briefly, we first construct a knowledge base from the session transcripts. When the system is queried with a user question, we retrieve relevant embeddings from the knowledge base to incorporate during prompting, allowing us to ground model responses in gold standard information from medical providers. In order to address privacy concerns, excerpts are summarized using an open-source LLM (Llama-2)

Table 1.1 Themes identified in counseling sessions

Theme	Description	Count
Relating to other participants	Link APOE biomarker risk patients to peers, highlighting shared lifestyle adaptations and coping strategies	25
Recognizing patients' prior knowledge	Anticipate patients' background knowledge	5
Distilling definitions into "take home messages"	Condense technical information	22
Acknowledging uncertainty	Highlight the statistical uncertainties of genetic risk estimates	49
Describing scope of risk	Rephrase risk in a more optimal way	72
Considering broader effects	Weigh holistic risk factors	38
Deferring to specialists	Defer to specialists when queries exceeded expertise	1

before being incorporated into the prompt. All embeddings below are computed using OpenAI's text-embedding-ada-002.

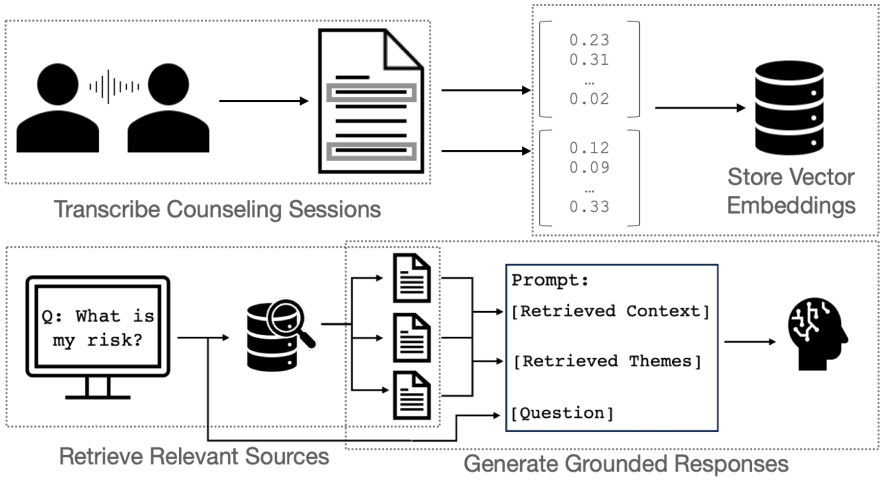

Fig. 1.1 Retrieval-augmented generation pipeline for generating answers to patient queries regarding Alzheimer's risk.

To construct our knowledge base, we first manually annotated transcripts to identify instances of the themes we previously identified. Each such instance is stored as a separate excerpt. We use Llama-2 to summarize all excerpts and store vector embeddings of these summaries, along with the annotated themes for each corresponding excerpt. This summarization step allows us to pass in the information from these transcripts without the need to share PHI with commercial LLM vendors. When the user asks a question Q, our system retrieves the three documents nearest to Q in the

embedding space. We incorporate these documents and their stored themes into a GPT-4 prompt as follows:

> You are a genetic counselor talking to a patient who has 2 APoE E4 genes. The patient asks the following question: [QUESTION] Answer this question as a genetic counselor. When answering the question, make sure you consider [THEME]. Be sure to use the following information in your answer when appropriate: [SUMMARY]

While large language models, such as GPT-4, can generate very convincing answers with implicit knowledge on a wide variety of topics, they are known to generate *hallucinatory content* (Lee et al. 2023). Though we cannot not entirely eliminate the potential for hallucination, in order to minimize the chances that the chatbot generates false information, we check the alignment of the final generated response to the retrieved context. The knowledge extraction stage of our pipeline is used to compute the semantic similarity between the generated answer and the extracted paragraph using the same sentence transformer as for extraction. If the computed similarity is below a certain threshold, we output that the chatbot is unable to answer the question. To further reduce the possibility of hallucination, the prompt specifies that the model should construct an answer by modifying the information from the retrieved excerpt rather than using its internal knowledge.

The structural development and implementation of the generation pipeline follows phases 0-2 detailed in Van De Sande et al. (2022). We evaluate available models (GPT and Llama), collect relevant data from genetic counseling sessions, and handle privacy by summarizing the retrieved knowledge before passing it into a closed-source LLM, and simulate results prospectively through an evaluation study detailed in the next section.

1.4 Evaluation

To evaluate the pipeline, we curated a dataset from transcripts of genetic counseling sessions. We reviewed and selected 20 questions that could be answered comprehensively with minimal additional patient information, and relied on one of the themes to be answered. For example, the question "What's the difference between dementia, mild cognitive impairment, and Alzheimer's disease?" would benefit from 'take-home messages' summarizing the three terms. A genetic counselor further reviewed and ultimately selected 10 representative questions based on an APOE disclosure framework. For each patient query in the evaluation dataset, we compared four types of responses: generated by our grounded pipeline using both LLaMA-2 and GPT-4 as response models, generated by plainly prompting GPT-4 without grounding (where the question is submitted without the theme and summary), and written by human genetic counselors. To evaluate the generated responses, we computed the ROUGE-L score, a well-established metric for evaluating QA performance based on n-gram similarity to a reference answer (human response). To compare the readability of the responses, we compute the Flesch Reading Ease, Gunning-Fog, and Coleman-Liau scores based on word and sentence lengths. These metrics also incorporate standards

based on the Patient Education Materials Assessment Tool (PEMAT) Shoemaker et al. (2014), specifically by focusing on elements like the use of everyday language and the definition of medical terms. The results are summarized in Table 2.

1.5 Results

Grounded GPT performed the best across all of our computed metrics. Responses from both of our grounded pipelines better matched those from human counselors than those from ungrounded GPT. We also find that grounding GPT improves the readability of the output, and that GPT produces more readable output than Llama. Comparing readability between human responses with those from grounded GPT yielded mixed results. Human responses scored slightly higher in the Flesch Reading Ease test, but slightly worse in the Gunning Fog and Coleman-Liau indices.

Table 1.2 Summarized text metrics

Response	ROUGE-L	Flesch Reading Ease	Gunning Fog	Coleman-Liau
Grounded Llama	0.180	48.198	12.527	12.077
Grounded GPT	**0.241**	**59.103**	**9.576**	**9.195**
Ungrounded GPT	0.160	45.684	13.566	13.027
Human Counselor	-	60.365	11.158	10.123

1.6 Discussion and Future Work

We present a pipeline to generate responses to patient questions regarding AD risk that emulate the style and content of genetic counselors. The resulting grounded responses better match human responses and are more readable than direct answers from GPT-4. Grounding the model in human conversations appears to calibrate the model to the level of jargon contained in the original conversation between genetic counselor and patient. For example, in response to the question "Do we know of other genes which cause Alzheimer's," ungrounded GPT discusses the molecular pathophysiology of several candidate AD genes, while Grounded GPT simply states that other AD genes exist but cumulatively have a small effect on risk. Responses generated by Llama-2 tended to be longer and more conversational (using greetings and signposting language), but tended to inject information that was not contained in the grounding text. Additionally, these responses naively incorporated the themes; for instance, when prompted to "acknowledge the uncertainty of risk," the Llama-2-generated response simply included the statement that "it's important to acknowledge that the risk estimate is an estimate and not a guarantee."

Counterintuitively, human responses scored worse on two of the three readability metrics. This may be because of their brevity, resulting in a higher density of complex words. Human counselors tended to respond more directly to the questions. For example, in answering "How common is the E4 type of APOE?" the human counselor gives a one sentence response providing the statistic, while both grounded and ungrounded GPT respond over several lines. The ungrounded GPT offers the frequencies of other APOE alleles, while grounded GPT reassures that the E4 allele is not a definitive marker. While this example illustrates our system working as intended–incorporating the theme of *acknowledging uncertainty* to generate a more empathetic response–it raises a question of whether and when patients prefer a direct answer. Prior literature in RLHF has shown that humans prefer longer answers in the general domain chatbot setting (Singhal et al. 2023). However, patient preferences for the verbosity of responses in medical chatbots have not been well established. Human counselors may vary their approach to answering questions based on setting, and future work could explore how conversational agents can dynamically adapt their response based on patient dispositions.

Overall, these results suggest a number of interesting directions for future research. Human evaluations of subjective quality and patient preference are in progress. One downstream avenue is to further optimize the interaction between open and closed-source LLMs, ensuring even greater reliability and coherence. In addition, there's a need to investigate other clinical scenarios where our approach can be adapted and employed. Addressing the institutional barriers surrounding clinical data access would also be pivotal, possibly through collaborations that ensure data security while granting models limited but essential data access.

1.7 Conclusion

We develop a system for answering patient queries about Alzheimer's risk using a retrieval-augmented generation pattern. The use of tailored responses to patient inquiries to counselors appears to be a feasible and promising strategy that circumvents the challenge of quoting prior counselor responses, some of which contain sensitive health information, to reply to other patients' needs. The dual-method approach we introduce brings together the advantages of both open and closed-source LLMs, capitalizing on their strengths while addressing their individual limitations.

References

Al-Hilli, Z., Noss, R., Dickard, J., Wei, W., Chichura, A., Wu, V., Renicker, K., Pederson, H. J. and Eng, C.: 2023, A randomized trial comparing the effectiveness of pre-test genetic counseling using an artificial intelligence automated chatbot and traditional in-person genetic counseling in women newly diagnosed with breast

cancer, *Annals of Surgical Oncology* **30**(10), 5990–5996.

Bain, M., Huh, J., Han, T. and Zisserman, A.: 2023, Whisperx: Time-accurate speech transcription of long-form audio, *INTERSPEECH 2023* .

Braun, V. and Clarke, V.: 2019, Reflecting on reflexive thematic analysis, *Qualitative research in sport, exercise and health* **11**(4), 589–597.

Langlois, C. M., Bradbury, A., Wood, E. M., Roberts, J. S., Kim, S. Y., Riviere, M.-E., Liu, F., Reiman, E. M., Tariot, P. N., Karlawish, J. and Langbaum, J. B.: 2019, Alzheimer's prevention initiative generation program: Development of an apoe genetic counseling and disclosure process in the context of clinical trials, *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **5**, 705–716. PMID: 31921963.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J.: 2020, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* **36**(4), 1234–1240.

Lee, P., Bubeck, S. and Petro, J.: 2023, Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine, *New England Journal of Medicine* **388**(13), 1233–1239. PMID: 36988602.

Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A. and Alsentzer, E.: 2023, Do we still need clinical language models?, *arXiv preprint arXiv:2302.08091* .

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D.: 2020, Retrieval-augmented generation for knowledge-intensive NLP tasks, *CoRR* **abs/2005.11401**. **URL:** *https://arxiv.org/abs/2005.11401*

Li, D., Ren, Z., Ren, P., Chen, Z., Fan, M., Ma, J. and de Rijke, M.: 2021, Semi-supervised variational reasoning for medical dialogue generation, *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 544–554.

Lin, S., Zhou, P., Liang, X., Tang, J., Zhao, R., Chen, Z. and Lin, L.: 2021, Graph-evolving meta-learning for low-resource medical dialogue generation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 13362–13370.

Liu, W., Tang, J., Qin, J., Xu, L., Li, Z. and Liang, X.: 2020, Meddg: A large-scale medical consultation dataset for building medical dialogue system (preprint).

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H. and Liu, T.-Y.: 2022, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* **23**(6), bbac409.

Shoemaker, S. J., Wolf, M. S. and Brach, C.: 2014, Development of the patient education materials assessment tool (pemat): a new measure of understandability and actionability for print and audiovisual patient information, *Patient education and counseling* **96**(3), 395–403.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. et al.: 2022, Large language models encode clinical knowledge, *arXiv preprint arXiv:2212.13138* .

Singhal, P., Goyal, T., Xu, J. and Durrett, G.: 2023, A long way to go: Investigating length correlations in rlhf, *arXiv preprint arXiv:2310.03716* .

Thambisetty, M. and Howard, R.: 2023, Lecanemab and apoe genotyping in clinical practice—navigating uncharted terrain, *JAMA neurology* **80**(5), 431–432.

Van De Sande, D., Van Genderen, M. E., Smit, J. M., Huiskens, J., Visser, J. J., Veen, R. E., Van Unen, E., Hilgers, O., Gommers, D. and van Bommel, J.: 2022, Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter, *BMJ Health & Care Informatics* **29**(1).

Walton, N., Graceffo, S., Sutherland, N., Kozel, B., Danford, C. and McGrath, S.: 2023, Evaluating chatgpt as an agent for providing genetic education, *bioRxiv* pp. 2023–10.

Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, A. M., Winblad, B., Jönsson, L., Liu, Z. and Prince, M.: 2017, The worldwide costs of dementia 2015 and comparisons with 2010, *Alzheimer's & Dementia* **13**(1), 1–7.

Yan, B. and Pei, M.: 2022, Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 2982–2990.

Zhou, S. and Bickmore, T.: 2022, A virtual counselor for breast cancer genetic counseling: Adaptive pedagogy leads to greater knowledge gain, *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.