Precision Grounding: Augmenting Large Language Models with Evidence-Based Databases for Trustworthy Genetic Variant Summarization

Xinsong Du[1,2]*, Anna Nagy[3]*, Michael F. Oates[3.4], Yifei Wang[5], Xinyi Wang[1,2], Joseph M. Plasek[1,2], Samuel J. Aronson[3,4], Matthew S. Lebo[3,6,7], Li Zhou[1,2]

[1] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School
[2] Department of Biomedical Informatics, Harvard Medical School
[3] Laboratory for Molecular Medicine, Mass General Brigham
[4] Digital Research Operations, Mass General Brigham
[5] Michtom School of Computer Science, Brandeis University
[6] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School
[7] Molecular and Population Genetics, Broad Institute of Harvard and MIT

*Equal Contribution

Corresponding author:
Xinsong Du, Ph.D.
Division of General Internal Medicine and Primary Care
Department of Medicine
Brigham and Women's Hospital and Harvard Medical School
399 Revolution Dr, Suite 777
Somerville, MA 02145
E-mail: xidu1@bwh.harvard.edu

Abstract

Backgrounds: Accurate interpretation of genetic variants is critical for precision medicine. While large language models (LLMs) show promise for summarization, they are prone to hallucinations. Many retrieval-augmented generation (RAG) techniques have been proposed to reduce hallucinations by incorporating external information, but generic web grounding-based forms of RAG retrieve information from the internet using the same techniques and internet resources across all domains and tasks, limiting the effectiveness of hallucination reduction. We thus propose Precision Grounding that augments LLMs with evidence-based, variant-specific information to improve summarization accuracy.

Methods: Unlike generic web grounding-based forms of RAG, which creates a vector database and then retrieves relevant information based on document embedding similarity between prompt and information in the vector database, Precision Grounding uses a query tool that integrates domain expert-selected resources and allows users to query relevant factual information using identifiers in databases. In our case, which is genetic variant summarization, we developed CATT (https://shorturl.at/pw81X), an open-source tool that harmonizes variant-specific information across ClinGen, ClinVar, and GenCC. Users can query factual information based on Variation ID for grounding LLMs. We evaluated our approach against a commonly used baseline–web grounding-based RAG–using 50 representative variants selected by genetic experts based on pathogenicity and information density. Accuracy and completeness were manually assessed using a 5-point Likert scale, with higher accuracy defined as lower clinical harmfulness. We then conducted error analysis and evaluated natural language processing (NLP) metrics, comparing them to human assessments via Spearman's correlation.

Results: Web grounding-based RAG significantly improved genetic variant summarization. At the time of evaluation, models lacking web-search grounding (O1, Claude 3.5 Sonnet, and DeepSeek R1) showed lower accuracy than GPT-4o, which incorporated web-based grounding. Using GPT-4o, Precision Grounding outperformed web-search grounding, achieving accuracy and completeness scores of 4.76 (+0.74) and 4.94 (+0.84), respectively. Error analysis revealed that Precision Grounding reduced clinically harmful hallucinations, such as incorrect pathogenicity classification. Automated NLP metrics correlated weakly with expert-rated accuracy; BLEU-4 showed strongest and significant negative correlation with expert-rated completeness ($\rho = -0.3321$, $p = 0.0185$).

Conclusion: Precision Grounding outperformed web-search grounding for genetic variant summarization. Our open-source tool, CATT, enables integration of curated, precise domain-specific knowledge and reduces hallucinations in LLM outputs. Weak correlations between NLP metrics and expert ratings underscore the need for improved automated evaluation metrics tailored to summarization tasks.

Keywords: Large Language Models**,** Genomics, Hallucination, Summarization, Retrieval Augmented Generation

1. Introduction

Genomic variation plays a significant role in human disease, with approximately 5 million variants present in each individual's genome.[1–3] Technological advances have made it routine to detect these variants, fueling progress in both basic and clinical research.[4] However, the ability to detect DNA variation has outpaced our capacity to interpret the clinical relevance of these variants.[5] Current practices focus on only a small subset of variants, limiting the potential impact of genetic testing and its contribution to research.[1] This gap is partly due to the high personnel costs required for the in-depth assessment and summarization of evidence related to specific genes and variants.[6] Fortunately, large language models (LLMs) hold significant potential to enhance clinical practice by automating the interpretation and summarization of genetic variants, thereby reducing costs and improving accessibility to genetic insights.

However, LLMs are prone to generating inaccurate or misleading information—a phenomenon known as hallucination.[7,8] While many retrieval-augmented generation (RAG) techniques have been proposed to reduce hallucinations by incorporating external factual information such as Wikipedia,[9] generic web grounding based forms of RAG retrieve information from the internet using the same techniques and the same internet resources across all domain and tasks.[10–13] This grounding approach may fail to retrieve specialized, domain-specific knowledge accurately, limiting its ability to ensure factual accuracy, especially in high-stakes domain.[12,13] In genomic medicine, such hallucinations will limit the ways these tools can be used, as incorrect variant interpretations may lead to incorrect pathogenicity classifications, flawed clinical summaries, and inconsistent evidence representation. These errors can have serious consequences in clinical decision-making, potentially leading to unnecessary medical interventions, misdiagnoses, or missed opportunities for timely treatment.

To address these challenges, we propose Precision Grounding that enhances LLM with publicly available evidence-based databases and resources identified by domain experts. Unlike traditional web grounding-based RAG which uses a vector database for querying, Precision Grounding uses a query tool which allows users to query relevant information based on identifiers in domain-specific databases (e.g., genetic variation ID). In the context of generic variant summarization, we developed CATT (ClinGen AI data Transformation Tool) as the query tool. Notably, CATT harmonized ClinGen,[1] ClinVar,[8] and GenCC[9] as targeted resources for retrieval augmentation and allows users to query factual information using Variation IDs in ClinVar to ground LLMs. By implementing Precision Grounding, we aim to improve the factual accuracy, consistency, and clinical relevance of LLM-generated interpretations. This work fills a critical gap in applying LLMs to genomic medicine and supports safer, more reliable, and evidence-informed use of generative models in clinical decision support.

2. Methods

2.1.CATT

We developed CATT to integrate and harmonize expert curated genetic databases, including ClinGen, ClinVar, and GenCC, which are developed to support the understanding of relationships between human genetic variation and health. CATT is open-source and publicly accessible via the ClinGen website (https://shorturl.at/pw81X) and GitHub (https://github.com/mgbpm/clingen-ai-tools). Data from these sources are linked using standardized identifiers such as HGNC gene IDs or ClinVar Variation IDs to ensure consistency across datasets. Users can easily query the integrated database and retrieve information for specified variants by using command-line and providing the variation ID in ClinVar. Then, CATT outputs queried genetic and clinical data into formats optimized for LLMs and other AI algorithms. The tool generates primarily two types of output: template-based natural language text and structured CSV files, both designed for direct use in LLM workflows. This dual-output design enables flexible incorporation of evidence-based knowledge into prompt engineering or downstream AI applications potentially in the future enabling more robust variant interpretation and/or clinical decision support. Notably, CATT's modular architecture supports future extensibility; new data sources can be incorporated by adding a subdirectory to the "sources" folder, with metadata and download specifications provided through a standardized set of files ("config.yaml", "dictionary.csv", and "mapping.csv"). Additionally, CATT supports batch processing, allowing querying multiple variants in a single run.

2.1.1. Template-Based Text Generation

The tool converts structured genetic and clinical data into natural language descriptions using Genshi Template (https://genshi.readthedocs.io/en/latest/upgrade.html#text-templates), which is a flexible, standardized, and well-maintained tool. This approach allows for the dynamic generation of text summaries that preserve critical contextual and clinical details. Each dataset field, such as gene name, variant classification, or clinical significance, is mapped to a corresponding narrative structure. Users can define custom templates to specify how these structured data points should be expressed in natural language.

During processing, the tool applies the predefined templates to individual records, generating text outputs. The templates incorporate conditional phrasing, allowing the generated text to adapt based on field values. This provides context for the content, particularly for categorical and numerical data. The final text representations can be saved as separate text files for each record or concatenated into a single document.

2.1.2. CSV-Based Data Representation

In addition to generating templated-based text files, CATT provides an option to structure genetic variant data into CSV format. This structured format allows LLMs to analyze tabular genetic information in conjunction with natural language queries, particularly in attachment-based input workflows where LLMs process structured data as supplementary input.

The tool allows users to specify which gene- and variant-level features should be included in the CSV output, such as gene symbol, variant ID, clinical significance, and pathogenicity classification. The column formatting is standardized to ensure consistency across records, making the data easily interpretable by LLMs. Categorical data can be converted into human-readable text to facilitate structured reasoning without requiring additional preprocessing. The CSV output can be used directly as an attachment for LLM queries, enabling models to extract insights from structured clinical and genomic records.

2.2.Precision Grounding to Augment LLMs to Increase Variant Summarization Trustworthiness

As illustrated in **Figure 1**, unlike the generic RAG method which first creates a vector database and then retrieves relevant information based on the similarity of document embeddings, our method first integrates expert-selected evidence-based resources then retrieves relevant information based on identifiers in databases. Unlike generic RAG method where a vector database is needed, Precision Grounding requires the creation of a domain-specific query tool that allows users to query information from integrated domain-specific databases. In our case (**Figure 2**), which is the genetic variant summarization task, CATT is the query tool. Our pipeline includes three steps: Step 1: Query CATT for user specified genetic variant information and get retrieved information in .txt and .csv files. Example .txt and .csv files are available in our GitHub repository (https://shorturl.at/tSQvK). Step 2: Upload these files as attachments into the LLM interface. Step 3: Prompt the LLM to generate a clinically useful variant summary.

2.3.Evaluation of The Effectiveness of Precision Grounding

The evaluation consisted of two phases: 1) LLM selection, 2) evaluation method selection, and 3) comparison and evaluation of Precision Grounding and web-search grounding. During LLM selection, we aim to select the most accurate LLM when Precision Grounding was not employed. Additionally, given there are a variety of evaluation approaches for assessing LLM performances,[14] we tested and selected the most appropriate evaluation method. In the model comparison phase, we assessed the performance of Precision Grounding using the selected LLM and evaluation method. We also evaluated our baseline method: web-search grounding, which has been integrated into many LLMs as a function and widely used in daily life. We compared Precision Grounding with web-search grounding.

A total of 50 representative genetic variants were selected for the purpose of testing and evaluation, with 10 variants used in the first and second phases and all 50 variants used in the third phase. Variant selection was designed to capture the diversity of existing variant spectrum, considering factors such as variant type (e.g., loss of function, missense, splicing), ClinVar pathogenicity classification (e.g., benign, pathogenic, risk allele), submitter concordance or discordance in ClinVar, and information density (measured by the number of gene and variant classification submissions).

2.3.1. LLM Selection

We first identified the most accurate LLM for genetic variant summarization by evaluating four widely used LLMs, including GPT-4o, O1, Claude 3.5 Sonnet, and DeepSeek R1. One-shot prompting was employed, as preliminary observations indicated that zero-shot prompting often produced lengthy and clinically less useful outputs. Each LLM was prompted to generate a summary intended to assist clinicians in making informed clinical decisions. The exemplary summary included in the prompt was crafted by a genetic expert to guide the LLMs toward producing clinically useful outputs.

2.3.2. Evaluation Method Selection

2.3.2.1.Human Evaluation

The evaluation guideline was developed through an expert consensus between two genetic experts (AN and MSL). To minimize subjective bias, LLMs' responses were anonymized and randomized during evaluation, ensuring the reviewers were blinded to the identity of the model that generated each response.

Accuracy was manually evaluated using a 5-point Likert-Scale, and a higher accuracy means less clinically harmful.[15] A score of 1 indicated very poor accuracy, characterized by major errors that could lead to harmful interpretation of the variant's information. A score of 2 reflected poor accuracy with notable errors impacting understanding. A score of 3 corresponded to moderate accuracy, with inaccuracies primarily affecting supporting

details rather than key findings. A score of 4 indicated minor inaccuracies or hallucinations that did not significantly alter interpretation. A score of 5 denoted perfect accuracy, devoid of any errors.

Completeness was similarly assessed using a 5-point Likert-Scale.[15] A score of 1 indicated a complete absence of relevant information from key sources (ClinVar, ClinGen, or GenCC). A score of 2 reflected significant omissions that impaired the interpretation of the variant and/or associated gene. A score of 3 indicated partial completeness, with information drawn from at least one of the sources but lacking key elements. A score of 4 denoted mostly complete responses, with only minor omissions. A score of 5 indicated full completeness, incorporating all relevant variant and gene information from ClinVar, ClinGen, and GenCC, including American College of Medical Genetics and Genomics (ACMG) codes,[16] pathogenicity, and gene- or disease-level details.

2.3.2.2.Automated NLP Evaluation

Although human evaluation is considered the gold standard for assessing LLM performance on clinical tasks,[17,18] it is costly, labor intensive, and time-consuming (i.e., not scalable). Therefore, we assessed the effectiveness of traditional NLP metrics in evaluating the LLM-generated summaries for genetic variants. We calculated the correlation between automated NLP metric scores and human evaluation scores using Spearman's rank correlation coefficient, which was selected due to its widespread application in previous studies examining the relationship between human assessments and automated NLP metrics.[17,19]

2.3.3. Comparison and Evaluation of Precision Grounding and Web-Search Grounding

Following model selection, we augmented the best performing model with precision grounding and web-search grounding separately. During Precision Grounding, the CATT-generated .CSV and .TXT files were uploaded as attachments, and the model was informed of their presence via the additional instruction section of the prompt. Regarding web-search grounding, we turned on the online-search function when prompting the LLM. Using the selected evaluation method, we then assessed the accuracy and completeness of the LLM-generated variant summaries of the two RAG methods.

2.4.Error Analysis

During error analysis, we categorized errors into two types based on their potential impact on clinical interpretation: 1) Clinically significant errors: these errors could alter the variant classification in a way that may lead to inappropriate clinical decisions. This category includes factual inaccuracies or hallucinations that misrepresent the variant's pathogenicity when compared to the current knowledge (e.g., from ClinVar). For example, if the LLM incorrectly states that a benign variant is pathogenic, this could result in unnecessary interventions or anxiety for patients. 2) Clinically negligible errors: these are minor inaccuracies or misstatements that do not affect the overall interpretation or classification of the variant. Examples of these errors include numerical errors or suggesting a variant classification code that is inappropriate while maintaining the correct overall pathogenicity. Such errors are unlikely to mislead a clinician or influence medical decision-making, as the core pathogenicity assessment remain aligned with current evidence.

3. Results

3.1.Evaluation of The Effectiveness of Precision Grounding

Characteristics of the 50 included genetic variants are illustrated in **Supplementary Table S1**, which includes their pathogenicity information, and their information density in databases.

3.1.1. LLM Selection

The evaluation result is illustrated in **Table 2**. GPT-4o demonstrated the highest average accuracy (4.7), followed by DeepSeek R1 (4.1), OpenAI o1 (3.5), and Claude 3.5 Sonnet (3.5). For completeness, DeepSeek R1 scored the highest (4.9), followed by GPT-4o (4), o1 (3.8), and Claude 3.5 Sonnet (3.2). Since GPT-4o achieved the highest accuracy, it was selected to test the effectiveness of CATT. Detailed prompt and responses from LLMs during this process are illustrated in **Supplementary Table S2**. **Supplementary Table S3** contains notes and scores of each variant from our genetic expert when evaluating the LLM's accuracy. Gold standard genetic summaries written by our genetic expert (AN) are provided in the **Supplementary Table S4**.

3.1.2. Evaluation Method Selection

We found that all automated NLP metrics exhibited low Spearman correlations with human evaluations (**Figure 3**). For accuracy, the highest (though weak) positive correlation was observed with ROUGE-2 ($\rho = 0.1295$, $p = 0.3702$), followed by AlignScore-base[20] ($\rho = 0.1490$, $p = 0.3018$); however, neither correlation was statistically significant, indicating a weak and unreliable association with human-rated accuracy. In contrast, for completeness, several metrics showed statistically significant negative correlations with human ratings: BLEU-2 ($\rho = -0.3435$, $p = 0.0146$), BLEU-4 ($\rho = -0.3321$, $p = 0.0185$), METEOR ($\rho = -0.3134$, $p = 0.0267$), and ROUGE-L ($\rho = -0.3087$, $p = 0.0292$). These findings indicate that none of the automated metrics reliably captured human judgements of summary accuracy. Moreover, the significant inverse correlations with completeness suggest that, in the genetic domain, lower BLEU, METEOR, and ROUGE-L scores may correspond to higher human-perceived completeness—possibly reflecting limitations of these metrics in evaluating abstractive, semantically rich summaries. Therefore, due to the weak correlation between NLP metrics with gold standard human evaluation, we decided to use human evaluation.

3.1.3. Comparison and Evaluation of Precision Grounding and Web-Search Grounding

Table 3 shows the performance of the best-performing model (GPT 4o) on 50 variants with Precision Grounding and web-search grounding. We found Precision Grounding could largely improve the accuracy from 4.02 to 4.76 and completeness from 4.10 to 4.94. In addition, Precision Grounding eliminated clinically harmful hallucination such as summarization of the wrong variant or misinterpretation of the pathogenicity of a variant. Detailed prompt and responses from LLMs during this process are illustrated in **Supplementary Table S5**. **Supplementary Table S6** contains notes and scores of each variant from our genetic expert during the evaluation.

3.2.Error Analysis

Based on the evaluation of the ten variants during the LLM selection phase, as illustrated in **Supplementary Table S7**, we found O1 could misclassify the genetic variant pathogenicity and miss information from ClinVar when summarizing genetic variant. Claude 3.5 Sonnect could classify the pathogenicity incorrectly, report a wrong number of entries, and miss information from ClinVar. DeepSeek R1could also incorrectly classify the pathogenicity and report wrong number of entries. The most accurate model GPT-4o could misclassify the pathogenicity as well. Additionally, GPT 4o's summarization sometimes misses disease information, information from ClinVar, or minor interpretation information such as ACMG code. In addition to aforementioned errors and

hallucinations, during the phase of testing the effectiveness of Precision Grounding using 50 variants, we found GPT-4o + web-search grounding could overstate the pathogenicity and sometimes summarized the wrong variant (**Table 3**). Precision Grounding reduced hallucination regarding pathogenicity classification, and made the summary more complete, which further improved its trustworthiness.

Discussion

LLMs have demonstrated remarkable capabilities in natural language understanding and generation; however, they are prone to producing content that appears factual but is ungrounded—a phenomenon known as hallucination. In genomic medicine, such inaccuracies can lead to misinterpretations of genetic variants, potentially resulting in erroneous clinical decisions. To reduce this risk, various RAG techniques have been proposed for enhance LLM outputs by incorporating external information.[9] Nevertheless, generic web grounding-based forms of RAG retrieve information from the internet using the same techniques across all domains, limiting domain relevance and the effectiveness of hallucination reduction. In this study, we proposed Precision Grounding approach, which retrieves context-specific evidence from expert selected evidence-based databases as resources to guide LLM outputs when summarizing genetic variants. We developed CATT, an open-source tool that integrates authoritative variant databases and enables users to query and retrieve relevant information for specific variants. When applied to LLM summarization tasks, CATT-based Precision Grounding significantly improves both accuracy and completeness while reducing clinically harmful hallucinations. This improvement supports potential the clinical utility of targeted Precision Grounding offering and important steps toward providing healthcare professionals more trustworthy and actionable variant interpretations.

Precision Grounding substantially improves the accuracy and completeness of LLM outputs for genetic variant summarization. Among the evaluated models, GPT-4o achieved the highest accuracy. This is partially attributed to its ability to perform live web searches, unlike O1, Claude 3.5, and DeepSeek R1, which did not have the function of web-search grounding in our setting at the time of evaluation. Notably, O1, Claude 3.5 Sonnet, and DeepSeek R1 occasionally misclassified pathogenicity of a variant, which could lead to harmful consequences in clinical practice, underscoring the need for factual grounding in LLM outputs. Despite GPT-4o's superior performance, hallucination persisted, including misclassifying pathogenicity and summarizing the wrong variant. However, when we replaced web-based retrieval with CATT-generated files as background context, these hallucinations were eliminated. Regarding completeness, our Precision Grounding approach helps models ground in comprehensive domain specific data from curated databases such as ClinVar, ClinGen, and GenCC. This enables LLMs to receive critical annotations such as ACMG codes and disease association information, which were frequently missed by using web-search grounding. These results demonstrated the effectiveness of our targeted Precision Grounding strategy and CATT toolkit in improving the reliability and comprehensiveness of LLM-generated summaries in the genetic domain.

The low correlation observed between automated NLP evaluation metrics and human assessments highlights a key limitation in current approaches to evaluating the quality of AI generated clinical text. NLP metrics, such as BLEU,[21] ROUGE,[22] and METEOR,[23] primarily focus on surface-level similarity between generated and reference texts, under the assumption that good-quality outputs will have substantial lexical or structural overlap with human-written references. However, these metrics often fail to account for semantically correct paraphrasing, factual correctness, and clinical relevance – criteria that are critical to human evaluators. While human evaluation provides richer and more context-sensitive judgements, it is time-consuming, labor-intensive and subject to inter-rater variability, limiting its scalability.[24] These challenges underscore the urgent need for more sophisticated automated evaluation methods that better align with human judgment in assessing clinical utility. Such methods would enable more efficient, consistent, and scalable evaluations of NLP models, facilitating their integration into clinical workflows where accuracy and reliability are paramount.

This study has several limitations. Due to the high cost of expert evaluation, our analysis was limited to a sample of 50 variants. Broader validation of CATT using a larger and more diverse set of variants will be necessary to comprehensively assess its effectiveness. The current version of the pipeline is not fully automated and does not include a graphical user interface, which may limit its broader adoption in clinical settings. In addition, while ClinVar and GenCC represent some of the most comprehensive publicly available resources for genetic interpretations, they are not fully curated. As a result, they may contain inaccuracies and conflicting submissions

from different contributors. A comprehensive risk analysis would be required to assess the appropriateness of this tool for any clinical use. This risk analysis would have to consider current performance, bias, privacy, non-determinism as well as drift hazards and validation requires for each specific workflow.

Future work should focus on extending Precision Grounding to additional domains by developing integrated, expert-curated databases tailored to each domain. Tools for querying and retrieving information from these domain-specific resources are also needed to support effective grounding. To improve usability and scalability, we plan to develop a fully automated pipeline that eliminates manual steps—such as extracting files from CATT and manually attaching them to LLMs via web interfaces—and enhances the software's reproducibility and FAIRness (Findability, Accessibility, Interoperability, and Reusability).[25,26] We also aim to implement a graphical user interface to improve CATT's accessibility for non-technical users. The development of an automated, NLP-based evaluation framework is essential for scalable assessment of genetic variant summarization. Such a system would enable large-scale evaluation of Precision Grounding in the genomic domain. Additionally, methods to automatically detect hallucinations in LLM-generated variant summaries are warranted.

Conclusions

Accurate genetic variant summarization is essential for timely and reliable clinical decision-making. Although LLMs show promise for this task, hallucinations limit their clinical applicability. We proposed Precision Grounding and developed CATT, an open-source tool that retrieves relevant information from evidence-based, domain-specific databases. Our results show that the CATT-assisted Precision Grounding approach significantly reduces LLM hallucinations and outperforms web-search-based grounding. Weak correlations between NLP metrics and expert ratings underscore the need for improved automated evaluation metrics tailored to summarization tasks.

Author Contributions

Conceptualization and design: XD, AN, MFO, SJA, MSL, LZ. Programming and software development: XD, MFO. Prompt design for large language models: XD, AN, XW, LZ. Large language model generated response preparation: XD, XW. Gold standard variant summary preparation: AN. Genetic variant summary evaluation guideline preparation and manual evaluation: AN, MSL. Calculation of correlation between natural language processing metrics and expert evaluation results: YW. Results interpretation: XD, AN, MFO, JMP, YW, SJA, LZ. Original draft preparation: XD, AN. Significant contribution to paper revision: XD, AN, SJA, LZ. Funding Acquisition: XD, MSL, LZ. Supervision: LZ.

Acknowledgement

Research was supported by NIH-NHGRI (U24HG006834-10S1), NIH-NIA (R01AG080429), NIH-NLM (R01LM014239), and OpenAI Researcher Access Program.

Tables

Table 1. Manual Evaluation of LLM Performance on Genetic Variant Summarization During Model Selection (n =10 variants)

Evaluation	GPT-4o + Web Search Grounding	o1	Claude 3.5 Sonnet	DeepSeek R1
Average accuracy	4.7	3.5	3.5	4.1
Average completeness	4.1	4.0	3.3	4.8

Table 2. Manual Evaluation of LLM (GPT-4o) Performance Using Web-Search vs. Precision Grounding (n=50 variants)

Variant ID	GPT-4o + Web Search Grounding	GPT-4o + Precision Grounding Using CATT
Average Accuracy	4.02	4.76
Average Completeness	4.10	4.94

Table 3. Error Analysis

Measures	Error Category	Specific Error	Example Summary	
			GPT-4o + Web-Search Grounding	GPT-4o + Precision Grounding
Accuracy	Clinically significant errors	Incorrect classification of pathogenicity	X	
		Summarized wrong variant	X	
	Clinically negligible errors	Overstated pathogenicity	X	X
		Incorrect population frequency	X	X
		Incorrect codes	X	X
		Report incorrect number of submissions in databases	X	
Completeness		Missing disease information	X	
		Missing information of entries in a database	X	
		Missing ID information (e.g., ACMG code)	X	

"X" means that the model made the specific type of error(s)

Figures

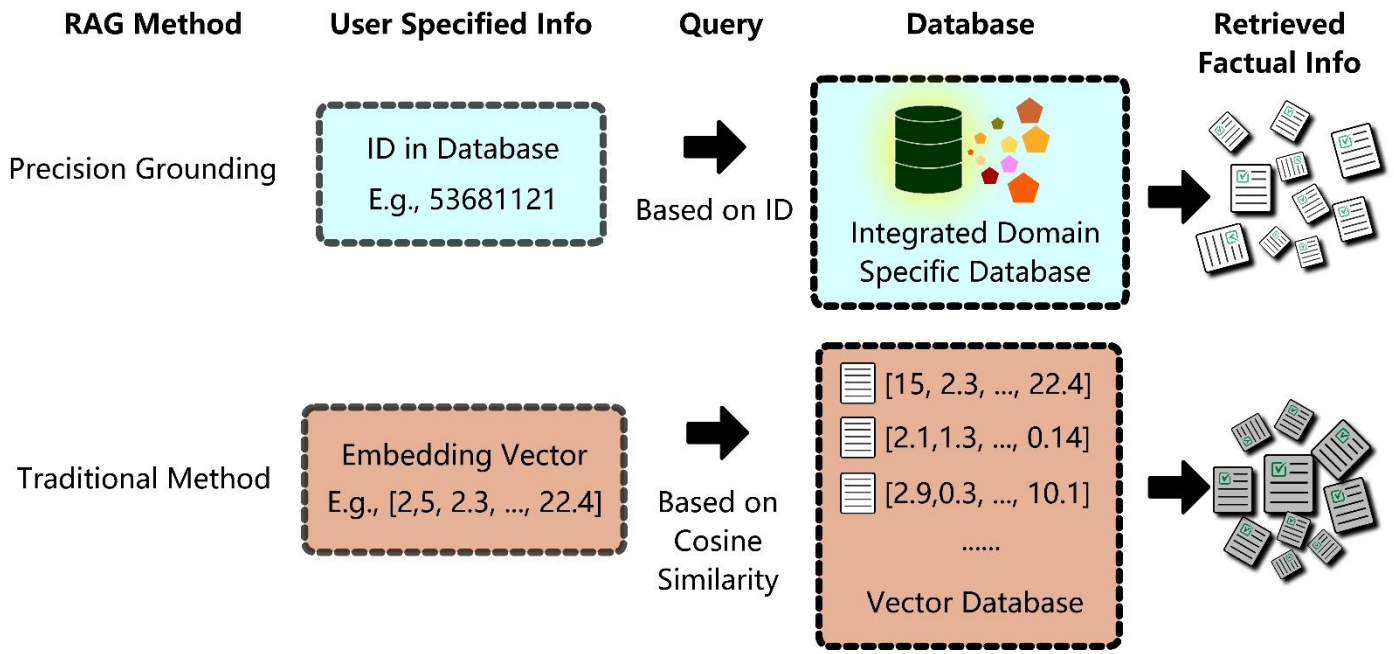


Figure 1. **RAG Methods Description**. Unlike traditional RAG methods which usually create a vector database and retrieve factual information based on cosine similarity, our proposed precision grounding method creates an integrated domain specific database, and retrieves relevant factual information based on identifiers in databases.

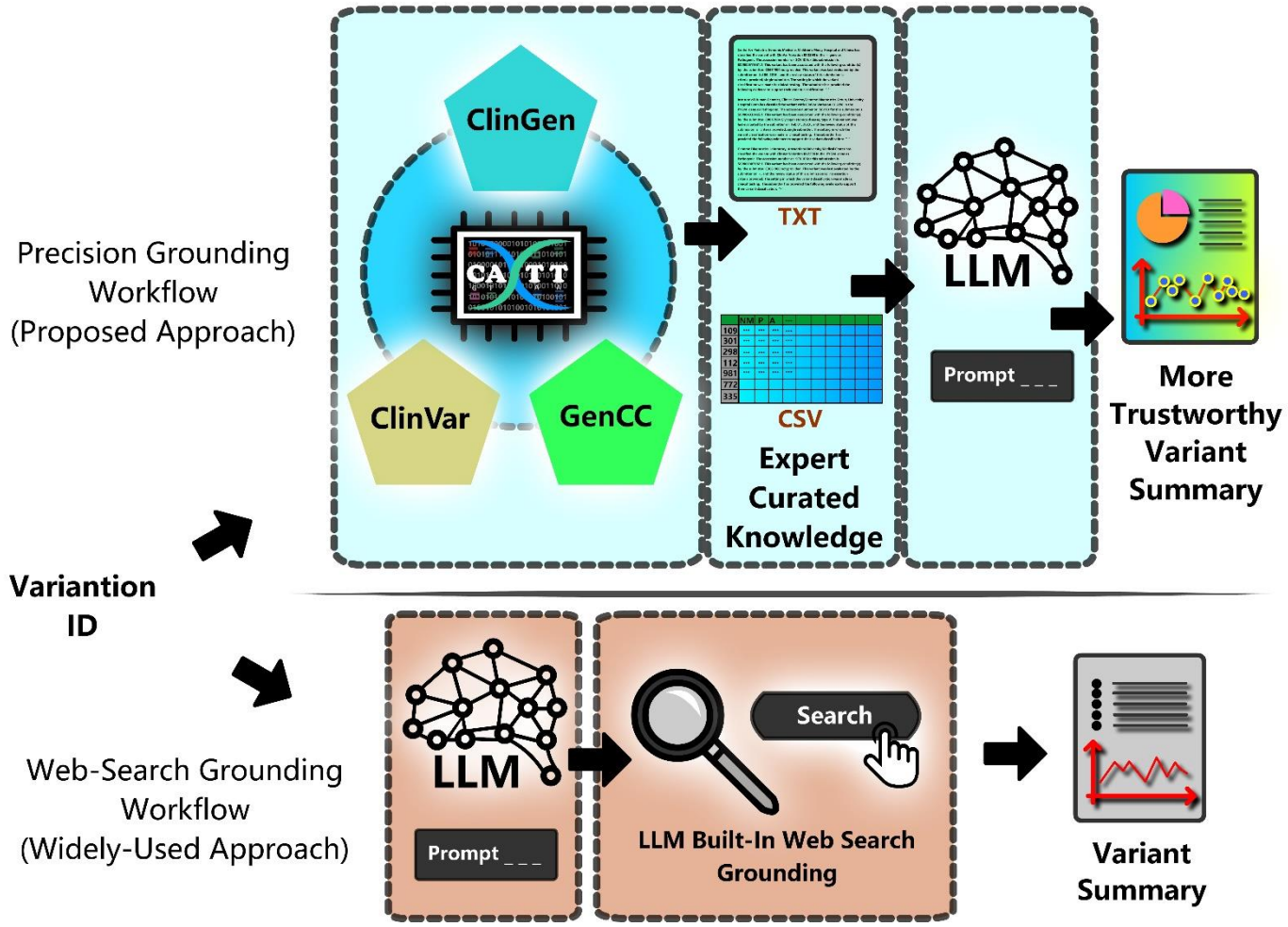


Figure 2. Precision Grounding vs Web-Search Grounding for Augmenting LLMs in Genetic Variant Summarization. In the traditional method, users directly prompt the LLM with web-search functionality enabled. Our precision grounding approach involves three steps: (1) generating curated variant-specific information in .txt and .csv formats using the CATT tool, (2) attaching these files to the LLM as background context, and (3) prompting the LLM to summarize the variant based on the grounded evidence.

Please summarize the following genetic variant as comprehensive as possible, and make the summary useful for a clinician: Variation ID: 2298 NM_005609.4(PYGM):c.148C>T (p.Arg50Ter)

Instruction

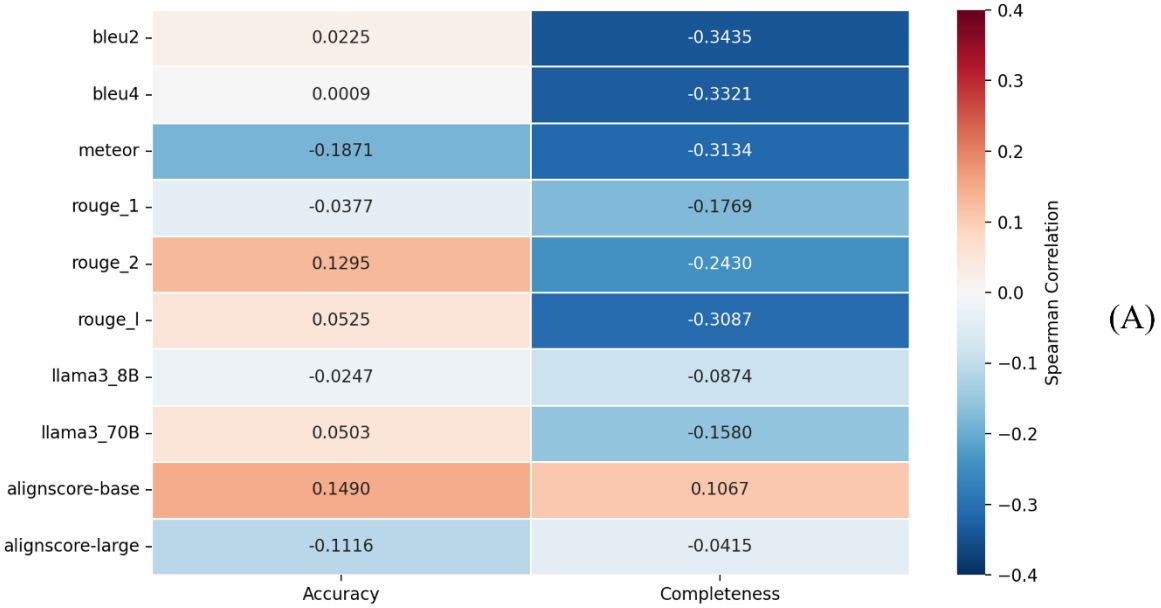
Example:
1) Summarization for the variant with ID 562 NM_000140.5(FECH):c.315-48T>C: The NM_000140.5:c.315-48T>C variant in the FECH gene is a splice variant. The FECH gene has been definitively associated with autosomal recessive erythropoietic protoporphyria by ClinGen. In addition, the FECH gene has been associated with autosomal recessive erythropoietic protoporphyria by 5 submitters in GenCC with the majority classifying this as a strong or definitive association. In ClinVar, this variant has conflicting classifications of pathogenicity based on 18 submissions. The majority of submissions classify this variant as Likely Pathogenic or Pathogenic, however 2 submissions which were assessed in 2016 classified this as a Variant of Uncertain Significance, and 1 submission classifies this as Pathogenic with Low Penetrance. Based on submission summaries in ClinVar, this variant has been reported in the literature in individuals with erythropoietic protoporphyria. Functional studies have also demonstrated aberrant splicing and a mild reduction in enzyme activity. Multiple submissions note that this variant may not cause disease or may have lower penetrance in the homozygous state, and may only be pathogenic when found with a loss-of-function or deleterious allele. This variant has been found in population databases such as gnomAD. ACMG codes which have been applied by submitters include: PM3_VeryStrong, PP1_Strong, PS3_Moderate, PM3, PP6, BS2.

Example(s)

Now, please summarize variant with ID 2298 and make is as useful as possible for a clinician.

Additional Instruction

Figure 2. Prompt Structure. Our crafted prompt includes three sections: instruction, example(s), and additional instruction. In the instruction section, we asked the LLM to make a clinically useful summary for the specific variant, and specified the variation ID and name, since we observed that the LLM always fabricates the variant name if we only include the ID but did not include the name in the prompt. In the example(s) section, we included a gold standard summary written by our genetic expert, as we found that the LLM usually outputs a very lengthy summary with redundant information without an example. In the additional instruction section, we rephrase the task and notify the LLM to look at attachments (if there are), since we found notifying the LLM about the attachment could slightly improve the accuracy.



(B)

P-values of Spearman Correlations

	Accuracy	Completeness
bleu2	0.8767	0.0146
bleu4	0.9951	0.0185
meteor	0.1932	0.0267
rouge_1	0.7947	0.2191
rouge_2	0.3702	0.0891
rouge_l	0.7173	0.0292
llama3_8B	0.8648	0.5461
llama3_70B	0.7287	0.2733
alignscore-base	0.3018	0.4609
alignscore-large	0.4402	0.7746

Figure 3. (A) shows the Spearman correlation coefficients between human ratings (accuracy and completeness) and automated NLP metrics; (B) presents the corresponding p-values indicating statistical significance. Overall, the correlations with human-judged accuracy were weak and non-significant across all metrics (e.g., BLEU-2: $\rho = 0.0225$, $p = 0.8767$; AlignScore-base: $\rho = 0.1490$, $p = 0.3018$). In contrast, several metrics exhibited statistically significant correlations with completeness, though all were negative: BLEU-2 ($\rho = -0.3435$, $p = 0.0146$), BLEU-4 ($\rho = -0.3321$, $p = 0.0185$), METEOR ($\rho = -0.3134$, $p = 0.0267$), and ROUGE-L ($\rho = -0.3087$, $p = 0.0292$).

References

1. Harrison SM, Biesecker LG, Rehm HL. Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. *Curr Protoc Hum Genet*. 2019;103(1):e93. doi:10.1002/cphg.93
2. Green RC, Shah N, Genetti CA, et al. Actionability of unanticipated monogenic disease risks in newborn genomic screening: Findings from the BabySeq Project. *Am J Hum Genet*. 2023;110(7):1034-1045. doi:10.1016/j.ajhg.2023.05.007
3. Machini K, Ceyhan-Birsoy O, Azzariti DR, et al. Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project. *Am J Hum Genet*. 2019;105(1):177-188. doi:10.1016/j.ajhg.2019.05.017
4. Diao JA, Kohane IS, Manrai AK. Biomedical informatics and machine learning for clinical genomics. *Hum Mol Genet*. 2018;27(R1):R29-R34. doi:10.1093/hmg/ddy088
5. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. *Genome Med*. 2021;13(1):153. doi:10.1186/s13073-021-00965-0
6. McInnes G, Sharo AG, Koleske ML, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am J Hum Genet*. 2021;108(4):535-548. doi:10.1016/j.ajhg.2021.03.003
7. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. *Nature*. 2024;630(8017):625-630. doi:10.1038/s41586-024-07421-0
8. Du X, Novoa-Laurentiev J, Plasek JM, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *EBioMedicine*. 2024;109:105401. doi:10.1016/j.ebiom.2024.105401
9. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Curran Associates Inc.; 2020:9459-9474.
10. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. *npj Digit Med*. 2025;8(1):1-5. doi:10.1038/s41746-025-01536-y
11. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artif Intell Rev*. 2024;57(11):299. doi:10.1007/s10462-024-10921-0
12. Li X, Zhao R, Chia YK, et al. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. Published online February 21, 2024. doi:10.48550/arXiv.2305.13269
13. Kenthapadi K, Sameki M, Taly A. Grounding and Evaluation for Large Language Models: Practical Challenges and Lessons Learned (Survey). In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’24. Association for Computing Machinery; 2024:6523-6533. doi:10.1145/3637528.3671467
14. Du X, Zhou Z, Wang Y, et al. Adapting Generative Large Language Models for Electronic Health Record Applications: A Systematic Review of Methodologies, Evaluation, and Hallucinations. Published online May 10, 2025:2024.08.11.24311828. doi:10.1101/2024.08.11.24311828
15. Albaum G. The Likert Scale Revisited. *Market Research Society Journal*. 1997;39(2):1-21. doi:10.1177/147078539703900202
16. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. doi:10.1038/gim.2015.30
17. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30(4):1134-1142. doi:10.1038/s41591-024-02855-5

18. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit Med*. 2024;7(1):1-20. doi:10.1038/s41746-024-01258-7
19. Tie X, Shin M, Pirasteh A, et al. Personalized Impression Generation for PET Reports Using Large Language Models. *J Digit Imaging Inform med*. 2024;37(2):471-488. doi:10.1007/s10278-024-00985-3
20. Zha Y, Yang Y, Li R, Hu Z. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. Published online May 26, 2023. doi:10.48550/arXiv.2305.16739
21. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, eds. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2002:311-318. doi:10.3115/1073083.1073135
22. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*. Association for Computational Linguistics; 2004:74-81. Accessed March 2, 2025. https://aclanthology.org/W04-1013/
23. Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein J, Lavie A, Lin CY, Voss C, eds. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics; 2005:65-72. Accessed March 2, 2025. https://aclanthology.org/W05-0909/
24. Croxford E, Gao Y, Pellegrino N, et al. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Syst*. 2025;2(1):1-13. doi:10.1038/s44401-024-00011-2
25. Du X, Dobrowolski A, Brochhausen M, Garrett TJ, Hogan WR, Lemas DJ. Nextflow4MS-DIAL: A Reproducible Nextflow-Based Workflow for Liquid Chromatography-Mass Spectrometry Metabolomics Data Processing. *J Am Soc Mass Spectrom*. Published online January 5, 2025. doi:10.1021/jasms.4c00364
26. Du X, Dastmalchi F, Ye H, et al. Evaluating LC-HRMS metabolomics data processing software using FAIR principles for research software. *Metabolomics*. 2023;19(2):11. doi:10.1007/s11306-023-01974-3