

Pilot Evaluation of Seventeen Large Language Models on Supplement-Associated Disease Recommendations Using Literature-Based Evidence

Tim Chu
Ctim1007@gmail.com
Forest Hills Central High School
Grand Rapids, MI 49546

Derek Chen
derekchen0404@gmail.com
City High Middle School
Grand Rapids, MI 49505

Xiaodan Zhang
zhan2064@msu.edu
Michigan State University
East Lansing, MI 48824

Bin Chen
chenbi12@msu.edu
Michigan State University
East Lansing, MI 48824

Abstract

Despite the widespread use of dietary supplements, they do not undergo the same regulatory review as FDA-approved therapeutics. Currently, there are no authoritative resources to guide the selection of supplements for diseases. With the increasing popularity of large language models (LLMs), many individuals are turning to systems like ChatGPT for supplement recommendations. No prior studies have systematically evaluated LLMs in this context. This study presents an exploratory benchmark of 17 LLMs for supplement-disease relationship extraction using a corpus derived from 10 commonly used dietary supplements. An LLM-assisted, literature-derived reference dataset of supplement-disease relationships for 10 dietary supplements was initially generated from PubMed using GPT and subsequently evaluated through manual review of a set of randomly selected entries. The 17 LLMs were then prompted to independently extract supplement-disease relationships for the same supplements, and their performance was evaluated using both exact-match and semantically adjusted schemes. Semantically adjusted evaluation yielded significantly higher performance estimates than exact-match evaluation. Precision-recall trade-off analyses identified multiple Pareto-optimal models, indicating that no single model simultaneously maximized precision and recall. Across the 17 evaluated LLMs, supplement-disease relationship extraction consistently demonstrated moderate-to-high precision but low recall, suggesting that current LLMs are better suited for generating high-confidence candidate relationships than for comprehensive knowledge extraction.

Keywords: Large Language Model, Supplement-Disease Relationship Extraction, Supplement Recommendation, Entity Extraction.

Pilot Evaluation of Seventeen Large Language Models on Supplement-Associated Disease Recommendations Using Literature-Based Evidence

Tim Chu, Derel Chen, Xiaodan Zhang and Bin Chen

1. INTRODUCTION

Dietary supplement use is widespread in the United States. Data from the 2017-2018 National Health and Nutrition Examination Survey (NHANES) shows that 57.6% of adults used at least one supplement in the past 30 days, with prevalence exceeding 70% among adults aged 60 years and older, and supplement use has increased across all age groups over the past decade (Mishra et al., 2021). The results from a recent national survey indicate that 90% of users believe dietary supplements are essential for maintaining their health (Ipsos, 2023). Exploring the potential role of supplements in managing chronic diseases has also become an integral component of the U.S. government's commitment to "Make America Healthy Again" (U.S. Department of Health & Human Services, 2025).

Despite their widespread use, dietary supplements do not undergo the same regulatory review as FDA-approved therapeutics. As a result, there are no authoritative, easily accessible resources to guide the general public in selecting supplements for specific health conditions. With the increasing popularity of large language models (LLMs), many individuals are turning to systems like ChatGPT for supplement recommendations. Recent polling shows that approximately 25% of adults under age 30 seek AI-generated health information at least monthly, yet trust remains low: only 29% of adults and 36% of AI users report confidence in AI as a reliable source of health guidance (Presiado et al., 2024). The disbelief may rise as LLM adoption accelerates if no additional effort is made to tackle the unique challenges in health guidance.

The purpose of this study is to conduct an exploratory benchmark of multiple LLMs for supplement-disease relationship extraction and prediction. We constructed an LLM-assisted, literature-derived reference dataset, hereafter referred to as the "reference table," applied rigorous disease-terminology standardization, and evaluated 17 leading LLMs for their ability to reproduce these associations and correctly assign directional labels (improves or worsens). Model performance is assessed using classification metrics, including accuracy, precision, recall, and F1 score. As a pilot study, we focused on 10 commonly used dietary supplements. This work establishes an initial benchmark for evaluating LLMs in biomedical information extraction and provides a foundation for future comparative studies in this area.

2. RELATED WORK

LLMs are already widely used in health-related applications, including diagnosis (Tripathi et al., 2024), drug discovery (Chakraborty et al., 2023), and medical question answering (Wu et al., 2024), and their performance has been extensively benchmarked. However, the evaluation of LLMs for identifying supplement-disease relationships remains limited. Much of the training data for general-purpose LLMs comes from the open internet, where biased, incomplete, or conflicting information is common. Furthermore, many modern LLMs do not disclose their training corpora, making it difficult for the public to assess the reliability of generated outputs. Independent assessments of health-focused LLMs have documented safety-critical failure modes.

In oncology, for example, models have generated incorrect or clinically inappropriate treatment recommendations, deviated from evidence-based breast cancer guidelines (Stalp et al., 2024), and produced inaccurate advice regarding immunotherapy-related toxicities (Liang et al., 2025). Broader evaluations have revealed hallucinated clinical explanations (Zhou et al., 2025a), fabricated citations (Walters & Wilder, 2023), and reduced accuracy on free-text medical questions (Zhou et al., 2025b), raising persistent concerns about reliability even in newly released model versions. In response, some LLM providers like Claude have begun integrating curated scientific literature, including PubMed, to improve domain-specific performance (Claude Academy, n.d.).

Emerging research highlights the potential of LLMs in nutrition and supplement science. Wu et al. (2024) introduced DIETNERD, an LLM-based system designed to translate complex scientific findings into accessible nutrition guidance, while Hou et al. (2025) developed Dietary Supplement Knowledgebase 2.0 to integrate authoritative supplement data and improve reliability. A recent systematic review (Fazel, 2024) evaluated GPT models for generating nutritionally accurate dietary recommendations and found that integrating external data sources (e.g., NutritionIX API) through Retrieval-Augmented Generation can substantially improve performance.

While these studies demonstrate early promise, none have systematically evaluated a broad set of LLMs against a standardized evidence-based reference for supplement-disease associations. To address this gap, this study benchmarks 17 LLMs using a curated reference dataset and evaluates their performance in supplement-disease relationship extraction.

3. METHOD

We developed an evaluation framework for assessing the fidelity of large language model-generated outputs to published biomedical evidence. The four major stages include identifying sample supplements, generating an LLM-assisted, literature-derived reference dataset for supplement-disease relationships from PubMed, utilizing LLMs to create their individual supplement-disease relationship table, and evaluating the performance of those LLMs by comparing their individual table to the reference table. Fig. 1 illustrates the evaluation framework.

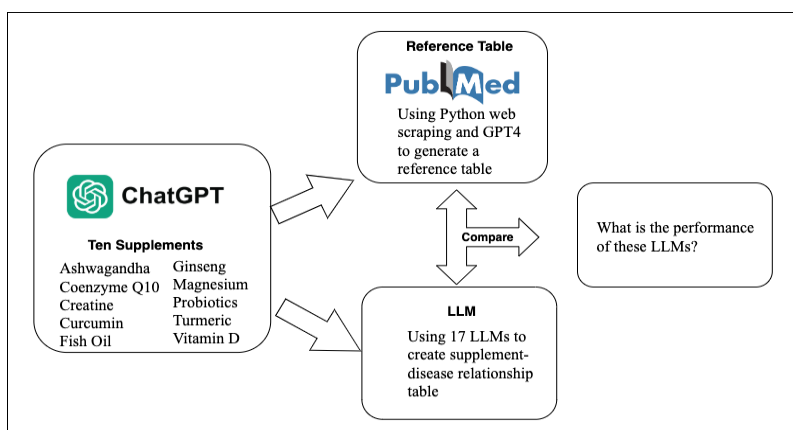


Figure 1. Evaluation Framework

3.1 Data Construction

Selection of Supplements

We identified ten commonly used dietary supplements via a ChatGPT query: ashwagandha, coenzyme Q10, creatine, curcumin, fish oil, ginseng, magnesium, probiotics, turmeric, and vitamin D. These ten supplements, chosen for their popularity, are used as the base for creating a reference table.

Reference Table Generation

We used Python scraping on PubMed, querying each supplement name individually to collect our data, which consisted of all articles returned by each query. Each article includes Title, Author(s), Journal, Abstract, Date (date the article was published), and PMID (PubMed ID). Then GPT-3.5 Turbo was used to filter down the articles, keeping only those clearly discussing supplement disease relationships in their abstracts. Next, GPT-4 was used to generate the supplement-disease relationships from the PubMed abstracts using the prompt shown in Appendix A. The Specific Supplement, Disease Mentioned, Disease Relationship, Effect on Disease, and Human Subject are further extracted from each abstract. Disease Relationship is a one-word summary of whether the supplement helps to treat the disease—either improves, worsens, or no effect, while Effect on Disease is a one to two sentence explanation of exactly how the supplement interacts with the disease.

After the initial data table was created, GPT-5 was used to refine it, addressing the issues with standardization and multiple supplements in a single supplement box. In this process, first, the rows where the supplement box already contained one of the ten supplements were skipped. For the rows where GPT-4 had identified a supplement not part of the 10 sample supplements, GPT-5 would read that row's abstract, being extra keen for mentions of supplement-disease relationships with one of the ten supplements, and regenerate the columns with the data from the abstract, conforming to the strict standards: being that the specific supplement column had to have only one supplement and that one supplement had to be one of the 10 sample supplements. If GPT-5 could not fill the columns according to data strictly from the abstract, the row was removed. Upon completing this refinement process, the final data table was constructed in which each row corresponds to one PubMed article, and has the columns Title, Authors, Journal, Abstract, Date, PMID, Specific Supplement, Disease Mentioned, Disease Relationship, Effect on Disease, and Human Subject.

Manual Quality Assessment

A random sample of 100 extracted supplement-disease relationships was manually reviewed against the source PubMed abstracts for quality assessment. The 100 relationships were jointly reviewed by an external researcher and one of the authors. Each relationship was assessed for correctness based on the available evidence in the corresponding abstract. Eighty-two of the 100 relationships were verified, corresponding to a validation accuracy of 82%. Most errors resulted from incorrect supplement or disease assignments or reversed relationship directions. Therefore, the reference table represents an automatically generated dataset that underwent human quality assessment to evaluate labeling consistency and identify potential errors.

3.2 Individual LLM Relationship Extraction Tables

The 17 of the most popular and top-performing LLMs on OpenRouter include DeepSeek-3.1, DeepSeek-R1, Gemini-2.0-Flash, Gemini-2.5-Flash, Gemini-2.5-Pro, GPT-4, GPT-4.1, GPT-5, Grok-3, Grok-4, Llama-3.3-70B-Instruct, Llama-4-Maverick, Opus-4, Opus-4.1, Qwen3-235B-A22B-2507, Qwen3-30B-A3B, and Qwen-Max. To generate each table, each model was accessed in Python using the OpenRouter API. Then each model was given the same prompt (as shown in Appendix B) to generate the Supplement, Disease, Relationship, and Confidence, and to identify as many evidence-backed supplement-disease relationships as possible. To mimic how the general public would use the model, we did not intend to optimize prompts for performance improvement.

3.3 Data Processing

Disease Term Standardization

To enable consistent comparison across models, we standardized all raw disease terms in the reference table and in the LLM-generated tables, as described in (Zhang et al., 2026). The raw outputs contained substantial variability, including differences in capitalization, hyphenation, abbreviations, pluralization, synonyms, and other formatting inconsistencies. Without harmonization, these variations would artificially inflate mismatches and undermine the validity of the performance evaluation. For this reason, all disease strings from the reference dataset and the model outputs were first combined into a single list, and a frequency table of unique entries was generated to identify duplicates and variant forms. A structured post-processing workflow was applied to harmonize these terms using GPT-4o. An internally developed set of standardization rules provided consistent criteria for normalizing disease names across sources. These guidelines included expanding abbreviations to complete diagnostic terms, consolidating synonymous or hierarchically related conditions, unifying terminology for tumors and cancers, and enforcing consistent formatting for capitalization, hyphenation, spacing, and plurality. The rules also specified procedures for handling multi-condition strings, incorporating severity and status modifiers, aligning viral infection terminology, and representing clinically recognized subtypes such as HER2-positive cancers. Informal and otherwise non-standard clinical terminology were corrected during this step. Although established biomedical vocabularies such as MeSH and UMLS provide standardized biomedical terminologies, GPT-assisted normalization was used due to variability and non-standard phrasing in extracted supplement-disease relationships.

After the automated raw disease terms processing, two authors conducted a manual review to ensure accuracy and clinical validity. During this review, we resolved ambiguous mappings and removed entries

that did not represent human disease entities. Excluded items included procedures (e.g., dialysis, hysterectomy), physiological parameters (e.g., blood pressure, glucose metabolism), functional states (e.g., disability in activities of daily living, slowness in gait speed), clinical outcomes (e.g., mortality, all-cause mortality), symptoms or non-specific descriptors, population labels (e.g., wrestlers), and non-human conditions. Composite strings were separated into their component diagnoses so that each disease concept was mapped one-to-one to a standardized label. The final standardized terminology set was used in all downstream matching procedures, accuracy calculations, and model-level comparisons.

Data Cleaning and Label Harmonization

After the standardization, further refinements were made to both the reference and LLM-generated tables. First, all rows with missing data in the disease row (e.g., a blank cell or "N/A") were removed. Then, duplicate supplement-disease relationships were removed. For example, if multiple rows identified the relationship that ashwagandha improves depression, only one of these rows would be kept, and all others would be deleted. For the LLM-generated tables, only duplicates generated under the same model were removed. When duplicates came from different models, each model retained only one instance of that relationship.

The reference table was further refined to improve consistency. Since the LLM-generated tables had only "Improves" or "Worsens" in the "Relationship" column, we removed all rows in the reference table where the Relationship column had a value that was not either "Improves" or "Worsens" (e.g., "Mixed", "Unclear", "No Effect") to ensure fair comparison. This decision was made to focus the evaluation on relationships with a clearly identifiable direction of association and to reduce ambiguity in determining agreement between model predictions and reference labels. However, this simplification does not fully capture the complexity and uncertainty of biomedical evidence and should be considered when interpreting the findings.

3.4 Performance Evaluation

Evaluation Metrics

Precision, recall, and F1-score were computed based on true positive, false positive, and false negative counts obtained from comparisons between model outputs and the reference dataset:

- True Positive (TP): The model listed a disease and a relationship that also appeared in the reference table for that same supplement. This represents a correct match.
- False Positive (FP): The model listed a disease and a relationship for a supplement that did not appear in the reference table. This represents a predicted association not supported by the reference.
- False Negative (FN): The model did not list a disease and a relationship that was present in the reference table. This represents a missed association.

True negatives were not used because the number of possible no-relationship supplement-disease pairs is extremely large and uninformative, which would inflate accuracy without reflecting meaningful model performance. Focusing on TP, FP, and FN ensures the metrics evaluate only the model's ability to generate correct, relevant associations. Therefore, accuracy was not used as an evaluation metric in this study.

The 95% confidence intervals for precision, recall, and F1-scores were estimated using nonparametric bootstrap resampling (2,000 iterations) at the relationship level, with metrics recomputed under the specified evaluation scheme in each iteration. This approach was used to account for sampling variability and provide robust uncertainty estimates without relying on parametric assumptions.

Evaluation Schemes

Two evaluation schemes were employed to assess model performance under different matching assumptions: an exact match evaluation and a semantically adjusted evaluation. In the exact match evaluation, a prediction was considered correct only if it exactly matched a reference supplement-disease relationship. Any deviation in wording or label structure was considered a mismatch. In the semantically adjusted evaluation, semantically equivalent relationships expressed using different wording were treated as matches. To be specific, all model outputs labeled as false positives were manually reviewed by two authors to identify cases where predictions were semantically equivalent to reference relationships but differed due to wording variation, partial matches, or ontology-related

inconsistencies (e.g., cognitive decline vs cognitive dysfunction). The reviewers jointly assessed each case, with knowledge of the LLM that generated the prediction, and discussed the available evidence to reach a consensus. Because the cases were reviewed jointly rather than independently, inter-rater agreement was not calculated. Relationships confirmed as representing the same clinical concept were reclassified as true positives. Precision, recall, and F1-score were computed separately under both evaluation schemes. The exact match evaluation provides a conservative estimate of model performance, whereas the semantically adjusted evaluation provides a relaxed estimate that accounts for semantic equivalence.

Model Comparison and Stratified Analysis

Cross-model performance was evaluated across the 17 LLMs by comparing precision, recall, and F1-score under the defined evaluation schemes. Models were ranked based on F1-score, and precision-recall trade-offs were evaluated using Pareto front analysis to characterize relative extraction performance across models. Models were also grouped by LLM family (Gemini, Opus/Claude, Qwen, Grok, Llama, DeepSeek, GPT) for comparison. The family precision is calculated as the arithmetic mean of the precisions of the member models. A subgroup analysis was conducted comparing models in the first half versus the second half of the evaluation set to assess potential differences in performance, with the hypothesis that models in the first half would demonstrate higher precision. Anytime an LLM generated an odd number of supplement-disease relationship pairs for a particular supplement, the remaining pair would go to the second half.

3.5 Post-hoc Error Analysis

In addition to quantitative evaluation, a post-hoc qualitative analysis was conducted to examine representative false positive predictions. Selected cases were reviewed using advanced large language models (e.g., GPT-5.1 and Gemini 3 thinking modes) to assist in interpreting potential reasons for incorrect or ambiguous predictions. This step was used solely for exploratory error analysis and was not included in the formal evaluation or performance comparison of models.

4. RESULTS

We retrieved around 70,000 articles and their abstracts using PubMed retrieval. Title, Author(s), Journal, Abstract, Publication Date, and PMID were extracted. After the first round of filtering with GPT-3.5 Turbo, 6,683 articles were retained. Then, using GPT-4 to create an initial reference table with the articles and refinement with GPT-5, this yielded a reference table with 4,017 rows and 2,550 unique supplement-disease pairs. After aggregation, the 17 LLMs generated data with the reference set; all raw disease strings were combined and deduplicated, producing 2,349 unique disease entries prior to harmonization. There are 1,005 unique standardized disease labels in the final set, which are used in downstream matching and model evaluation.

In the final standardized reference table, there are 1,912 unique supplement-disease relationships. Figure 2 displays the distribution of supplement-disease relationships in the final standardized reference table. The number of unique supplement-disease relationships studied varied for each supplement in the reference table. For example, the supplement with the most disease relationships was vitamin D, with 478. The supplement with the fewest disease relationships was creatine, with 51. The final output includes a total of 1,447 rows from all tables combined, representing 358 unique supplement-disease relationships.

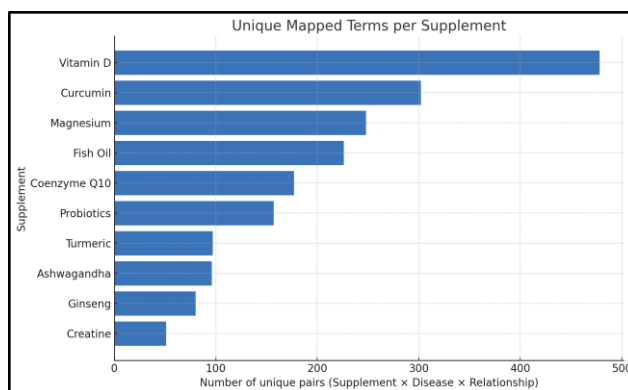


Figure 2. Unique Mapped Terms per Supplement in the Final Standardized Reference Table. Each bar represents the number of unique pairs after mapping.

Table 1 summarized the model performance across 17 LLMs under both exact match and semantically adjusted evaluations. Under the exact match evaluation, F1-score ranged from 0.033 to 0.11, precision ranged from 0.754 to 0.897, recall ranged from 0.017 to 0.059. Under the semantically adjusted evaluation, F1-score ranged from 0.035 to 0.116, precision ranged from 0.778 to 0.956, recall ranged from 0.018 to 0.062. A paired one-tailed t-test showed that F1-scores under the semantically adjusted evaluation scheme were significantly higher than those under the exact-match evaluation across the 17 LLMs ($p = 4.71 \times 10^{-7}$), indicating a consistent increase in measured performance when semantically equivalent relationships were accounted for through manual adjudication.

Model	Exact Match Evaluation			Semantically Adjusted Evaluation			$\Delta F1$
	P	R	F1 (95% CI)	P	R	F1 (95% CI)	
DeepSeek-r1	0.754	0.026	0.050 (0.037 - 0.063)	0.801	0.027	0.053 (0.040 - 0.067)	0.003
DeepSeek-3.1	0.815	0.034	0.066 (0.052 - 0.082)	0.89	0.038	0.072 (0.056 - 0.088)	0.006
Gemini-2.0-flash	0.808	0.027	0.052 (0.039 - 0.065)	0.856	0.028	0.055 (0.041 - 0.068)	0.003
Gemini-2.5-flash	0.889	0.034	0.065 (0.051 - 0.081)	0.945	0.035	0.068 (0.054 - 0.084)	0.003
Gemini-2.5-pro	0.823	0.041	0.079 (0.062 - 0.095)	0.885	0.044	0.084 (0.068 - 0.100)	0.005
GPT-4	0.888	0.017	0.033 (0.022 - 0.044)	0.944	0.018	0.035 (0.024 - 0.047)	0.002
GPT-4.1	0.765	0.055	0.102 (0.083 - 0.121)	0.778	0.055	0.103 (0.086 - 0.121)	0.001
GPT-5	0.757	0.047	0.089 (0.072 - 0.105)	0.797	0.049	0.093 (0.075 - 0.111)	0.004
Grok-3	0.85	0.03	0.058 (0.044 - 0.073)	0.897	0.032	0.061 (0.047 - 0.077)	0.003
Grok-4	0.814	0.048	0.090 (0.073 - 0.108)	0.839	0.049	0.093 (0.075 - 0.110)	0.003
Llama-3.3-70b-instruct	0.807	0.026	0.051 (0.037 - 0.064)	0.823	0.027	0.052 (0.039 - 0.065)	0.001
Llama-4-maverick	0.848	0.029	0.057 (0.043 - 0.071)	0.879	0.03	0.059 (0.045-0.073)	0.002
Opus-4	0.865	0.043	0.083 (0.065 - 0.100)	0.897	0.045	0.086 (0.069-0.103)	0.003
Opus-4.1	0.82	0.059	0.110 (0.093 - 0.129)	0.868	0.062	0.116 (0.099-0.136)	0.006
Qwen-Max	0.897	0.032	0.061 (0.047 - 0.077)	0.956	0.034	0.065 (0.050-0.080)	0.004
Qwen3-30b-a3b	0.781	0.026	0.050 (0.037 - 0.064)	0.877	0.029	0.057 (0.043-0.071)	0.007
Qwen3-235b-a22b-2507	0.784	0.028	0.055 (0.041 - 0.068)	0.812	0.029	0.056 (0.043-0.070)	0.001

Table 1. Comparison of the Large Language Models

Figure 3 presents the semantically adjusted F1-scores and their 95% bootstrap confidence intervals. Opus-4.1 achieved the highest overall performance (0.116, 95% CI: 0.099 – 0.136), followed by GPT-4.1 (0.103, 95% CI: 0.086 – 0.121). Bootstrap confidence intervals were generally narrow, indicating stable performance estimates across models. Despite variability in precision, all models exhibited consistently low recall, suggesting that current LLMs identify only a small fraction of true supplement-disease relationships.

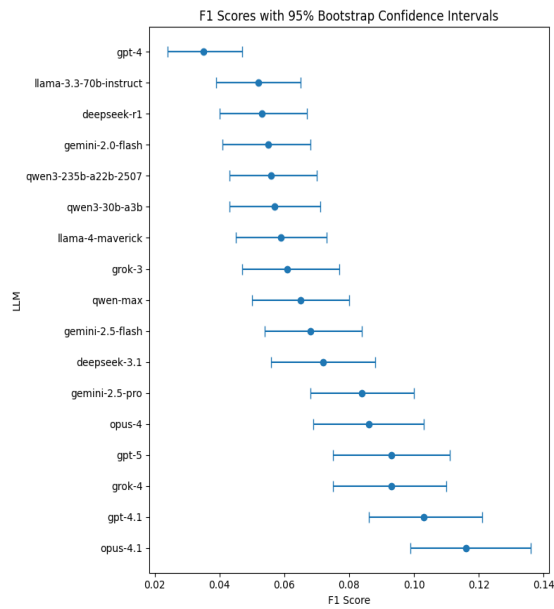


Figure 3. F1-Scores with 95% Bootstrap Confidence Intervals in the Semantically Adjusted Evaluation

As shown in Figure 4, precision–recall trade-offs varied substantially across models. The precision–recall analysis revealed two non-dominated operating regimes. The Qwen-Max model achieved both the highest precision and one of the highest recall values, representing the best overall balance across all evaluated systems. GPT-4 achieved slightly lower precision than the Qwen-Max model but the lowest recall among all models, indicating an extremely conservative extraction strategy. All remaining models were strictly dominated in at least one dimension.

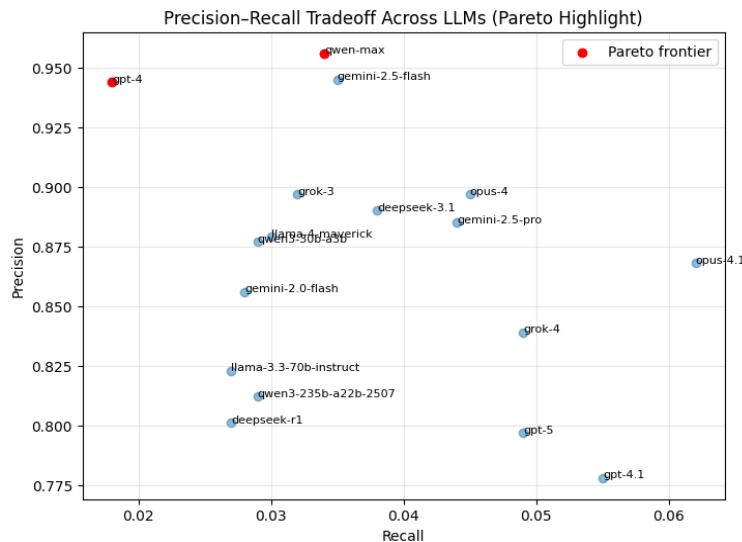


Figure 4. Precision-Recall Tradeoff across 17 LLMs in the Semantically Adjusted Evaluation. Each point represents one LLM, and highlighted points indicate LLMs on the Pareto frontier.

The families ranked in order of precision (calculated by averaging the precisions across the models in each family) are Google’s Gemini series with 0.896, followed by Claude’s Opus series with 0.886, Alibaba Cloud’s Qwen series with 0.881, xAI’s Grok series with 0.868, Meta’s Llama series with 0.851,

Precision vs. Family

Legend: Precision (Blue Bar), Average Precision (Red Line)

LLM Family	Precision	Average Precision
Gemini	~0.85	~0.82
Opus	~0.85	~0.82
Owen	~0.85	~0.82
Grok	~0.85	~0.82
LLaMa	~0.85	~0.82
DeepSeek	~0.85	~0.82
GPT	~0.85	~0.82

To examine whether extraction performance varied according to the order in which relationships were generated, precision was compared between the first and second halves of the outputs. Across the 17 LLMs, precision was consistently higher for the first half of the generated relationships than for the second half, with 16 of the 17 models exhibiting this pattern (see Figure 6). A paired one-tailed t-test shows that the difference was statistically significant ($p = 1.02 \times 10^{-5}$). The only model that did not have a precision fall off was Gemini 2.5 Flash, with second-half precision increased slightly (+0.009). On the other hand, the model with the largest drop off in precision was GPT-4.1 (-0.288). This suggests that LLMs tend to prioritize high-confidence recommendations, while it remains unclear whether the false positives in the second half arise from weaker underlying associations or from hallucinations.



©2026 ISCAP (Information Systems and Computing Academic Professionals)
<https://iscap.us/proceedings/>

recommendations arose; they likely reflect hallucinations that occur when the model generates recommendations sequentially. These findings suggest that additional reasoning and verification steps are necessary to confirm individual recommendations.

Supplement	Disease	Relationship	Label PubMed	Frequency	Appearances
Creatine	Parkinson's Disease	Improves	FP	12	deepseek-3.1, gemini-2.5-pro, gpt-4, gpt-4.1, grok-3, grok-4, llama-3.3-70b-instruct, llama-4-maverick, opus-4, opus-4.1, qwen-max, qwen3-30b-a3b
Creatine	Huntington's Disease	Improves	FP	6	deepseek-3.1, gemini-2.5-pro, gpt-4, grok-3, grok-4, llama-4-maverick
Fish Oil	Prostate Cancer	Worsens	FP	6	gemini-2.0-flash, gemini-2.5-pro, gpt-4.1, opus-4.1, qwen-max, qwen3-235b-a22b-2507
Creatine	Amyotrophic Lateral Sclerosis	Improves	FP	5	gpt-4.1, grok-3, grok-4, llama-4-maverick, qwen3-235b-a22b-2507

Table2. Common False Positives in LLM-Generated Tables

5. DISCUSSION

5.1 Discussion on the Findings

This study found that across 17 LLMs, supplement–disease relationship extraction consistently operated in a high-precision, low-recall regime under both evaluation schemes. The semantically adjusted evaluation shows a consistent increase of F1-score across the models. This finding suggests that exact-match evaluation systematically underestimates model performance by penalizing semantically equivalent predictions expressed using different terminology. Although some models achieved modest improvements in recall, no model simultaneously achieved high precision and high recall, indicating that current LLMs are better suited for generating high-confidence candidate relationships than for comprehensive biomedical knowledge extraction.

A notable finding is that the model achieving the highest F1-score (Opus-4.1) was not necessarily part of the Pareto-optimal set (Qwen-Max and GPT-4) in the precision–recall space. While F1-score identifies the best overall balance between precision and recall, Pareto analysis highlights models that are non-dominated in terms of trade-offs. In our results, Pareto-optimal models exhibited distinct precision–recall behaviors and differed from the top-F1 models, indicating that different evaluation criteria yield different notions of “best” model.

Despite strong precision, recall, and F1-scores were substantially lower across models. This was expected, given that the PubMed-derived reference contains more unique supplement-disease pairs being studied; models rarely attempt to list the full universe of known relationships. Instead, they provide a much smaller subset that tends to be high precision but incomplete. Therefore, high precision in this context may be more indicative of reliable extraction, whereas lower recall reflects the incomplete and conservative nature of automated inference rather than poor model performance.

A notable trend was observed when comparing model generations within each family. Newer model versions generally demonstrated improved performance over earlier versions, particularly in F1-score. This pattern holds in 5 out of the 7 families (DeepSeek, Gemini, Grok, Llama, and Opus). The two exceptions, GPT and Qwen, have reasonable explanations. GPT-4.1 and GPT-5 show very similar F1-scores, both substantially improved over GPT-4. For Qwen, all three models were released within a short timeframe, and Qwen-Max remains the flagship model, making its slightly better performance expected. This cross-generation improvement provides additional insight beyond pairwise model comparisons and highlights the importance of temporal model evolution in biomedical information extraction tasks.

When performance is grouped at the LLM family level, the differences between families were surprisingly small, with the standard deviation being only 0.0218. Furthermore, the precision difference between the best-performing family, Gemini, and the worst-performing family, GPT, was only 0.0549, highlighting the closeness in performance between all the models' families; no model family stood out as significantly better or worse than the other families. This suggests that all new models (the newest models being the top performers) all have impressive, near-equal performance, regardless of family. This is likely because the health-related corpora used to train these models are highly similar.

One notable finding was that 16 of the 17 LLMs exhibited significantly higher precision in the first half of their outputs compared to the second half. Although LLMs usually don't prioritize accuracy based on output position, a possible explanation is error accumulation during generation. Since later predictions are based on the original prompt plus the model's own previous outputs, an inaccurate earlier prediction could potentially influence subsequent predictions. As the model generates more predictions, the opportunity for such error propagation may increase. This provides a possible explanation for the observed decline in precision across output positions and warrants further investigation.

It is important to distinguish between the benchmarking objectives of this study and clinical applicability. This work evaluates LLMs in a controlled research setting for supplement-disease relationship extraction and should not be interpreted as a system designed for clinical decision-making or public health guidance. Although the models demonstrate varying levels of precision and recall, the presence of false positive predictions indicates that outputs may include incorrect or unsupported relationships. Therefore, these systems are not suitable for unsupervised use in health-related decision support without expert validation.

5.2 Implications and Limitations

This study serves as an exploratory benchmark for LLM performance in the task of supplement-disease relationship extraction, offering a comparative assessment of model performance in a controlled research setting. The results may be useful for researchers investigating LLM applications in biomedical information extraction by highlighting relative differences in precision and recall across models and identifying current limitations in automated relationship extraction. However, the findings are intended for research and methodological evaluation purposes and should not be interpreted as clinical guidance or recommendations for public use.

This study has several limitations. First, the LLM-assisted reference table may introduce annotation noise and model-dependent measurement bias. The reference table was constructed through several LLM-assisted steps and may therefore reflect model-specific patterns that could influence the evaluation of other LLMs. A manual review of a random sample of 100 relationships showed 82% agreement, indicating generally good but imperfect quality. Some errors were not fully propagated through the dataset prior to evaluation, so minor residual labeling noise may remain in the reported results and may affect performance estimates and comparisons among models, particularly when differences in performance are small. Thus, the reference table should not be interpreted as a fully independent gold standard. Second, each LLM was evaluated using a single generation per prompt; although bootstrap analysis was used, variability from multiple stochastic generations was not explicitly modeled. Third, the reference table included only supplement-disease relationships with clearly defined directions of association (improves or worsens), which simplified the nuanced and sometimes conflicting nature of biomedical evidence. Additionally, because the reference dataset was derived primarily from PubMed abstracts, it may lack clinical nuance present in full-text articles and thus represents a literature-level benchmark rather than a clinically validated ground truth for medical decision-making. Finally, because LLMs evolve rapidly, the models evaluated in this study represent only a snapshot in time, and newer releases may exhibit different performance.

To improve LLM reasoning abilities, future work should also explore using LLMs to identify less-well-known supplement-disease relationships. Moreover, future work may focus on common chronic conditions, in addition to top supplements. For example, given one chronic condition, we will compare the recommended supplements across LLMs. Future work may also incorporate LLM-generated confidence scores to support threshold-based evaluation and more nuanced ranking of predictions.

6. CONCLUSION

This study represents an exploratory benchmark of LLMs for supplement–disease relationship extraction. The performance of 17 LLMs was evaluated using a reference table derived primarily from PubMed abstracts through LLM-assisted extraction and human review. The results demonstrated substantial variation in model performance and showed that evaluation methodology significantly influences performance estimates, with semantically adjusted evaluation consistently producing higher scores than exact-match evaluation. Although several models achieved relatively high precision, recall remained modest, suggesting that current LLMs are better suited for generating high-confidence candidate relationships than for comprehensive knowledge extraction. These findings underscore both the promise and current limitations of LLMs for biomedical relationship extraction and emphasize the need for systematic and semantically aware evaluation methodologies. Future work should expand the benchmark by incorporating independently curated datasets, repeated model evaluations, and expert review to further validate these findings and support more definitive conclusions regarding LLM performance in biomedical information extraction.

7. REFERENCES

- Chakraborty, C., Bhattacharya, M., & Lee, S.-S. (2023). Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Molecular Therapy—Nucleic Acids*, 33, 866–868. <https://doi.org/10.1016/j.omtn.2023.08.009>
- Fazel, S. V. S. (2024). *Investigating GPT models for nutritional accuracy: A systematic review and performance enhancement* [Master's thesis, University of Gloucestershire]. ResearchGate. <https://doi.org/10.13140/RG.2.2.16297.51041>
- Hou, Y., Bishop, J. R., Liu, H., & Zhang, R. (2025). Improving dietary supplement information retrieval: development of a retrieval-augmented generation system with large language models. *Journal of Medical Internet Research*, 27, e67677. <https://doi.org/10.2196/67677>
- Mishra, S., Stierman, B., Gahche, J., & Potischman, N. (2021). Dietary supplement use among adults: United States, 2017–2018 (NCHS Data Brief No. 399). National Center for Health Statistics. <https://doi.org/10.15620/cdc:101131>
- Ipsos. (2023, October 13). *Nine in ten dietary or nutritional supplement users agree that dietary supplements are essential to maintaining their health*. <https://www.ipsos.com/en-us/nine-ten-dietary-or-nutritional-supplement-users-agree-dietary-supplements-are-essential>
- Presiado, M., Montero, A., Lopes, L., & Hamel, L. (2024, August 15). *KFF health misinformation tracking poll: artificial intelligence and health information*. KFF. <https://www.kff.org/public-opinion/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/>
- Stalp, J. L., Denecke, A., Jentschke, M., Hillemanns, P., & Klapdor, R. (2024). Quality of ChatGPT-generated therapy recommendations for breast cancer treatment in gynecology. *Current Oncology*, 31(7), 3845–3854. <https://doi.org/10.3390/curroncol31070284>
- Liang, S., Zhang, J., Liu, X., Huang, Y., Shao, J., Liu, X., Li, W., Wang, G., & Wang, C. (2025). The potential of large language models to advance precision oncology. *eBioMedicine*, 115, 105695. <https://doi.org/10.1016/j.ebiom.2025.105695>
- Tripathi, S., Sukumaran, R., & Cook, T. S. (2024). Efficient healthcare with large language models: Optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association*, 31(6), 1436–1440. <https://doi.org/10.1093/jamia/ocad258>
- Claude Academy. (n.d.). *Using the PubMed Connector in Claude*. <https://academy.claude.com/tutorials/using-the-pubmed-connector-in-claude>

- U.S. Department of Health & Human Services. (2025). *Make America Healthy Again Initiative*. <https://www.hhs.gov/maha/index.html>
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, Article 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- Wu, S., Yacub, Z., & Shasha, D. (2024). DietNerd: A nutrition question-answering system that summarizes and evaluates peer-reviewed scientific articles. *Applied Sciences*, 14(19), 9021. <https://doi.org/10.3390/app14199021>
- Zhang, X., Paithankar, S., Pu, J., Murtaza, M. S., Shankar, R., Leshchiner, D., Koirala, S., Palmer, Z., Nault, R., Li, X., Xie, Y., & Chen, B. (2026). Automating scientific annotations for open transcriptomic profiles via multi-stage agents. *bioRxiv*. <https://doi.org/10.64898/2026.08.19.745739>
- Zhou, H., Liu, F., Wu, J., Zhang, W., Huang, G., Clifton, L., Eyre, D., Luo, H., Liu, F., Branson, K., Schwab, P., Wu, X., Zheng, Y., Thakur, A., & Clifton, D. A. (2026). A collaborative large language model for drug analysis. *Nature Biomedical Engineering*, 10, 870–881. <https://doi.org/10.1038/s41551-025-01471-z>
- Zhou, S., Xu, Z., Zhang, M., Xu, C., Guo, Y., Zhan, Z., Fang, Y., Ding, S., Wang, J., Xu, K., Xia, L., Yeung, J., Zha, D., Cai, D., Melton, G. B., Lin, M., & Zhang, R. (2025). Large language models for disease diagnosis: A scoping review. *npj Artificial Intelligence*, 1, Article 9. <https://doi.org/10.1038/s44387-025-00011-z>

APPENDIX A

Prompt Used to Generate the Supplement-Disease Relationships from the PubMed Abstracts

"You are a medical research assistant. Review the following abstract and determine whether it is relevant to a study on supplements' effects in humans. Please think step by step through each criterion below. For each, analyze the information and determine if it is explicitly supported by the abstract."

"Be accurate, cautious, and double check your reasoning before giving your final answer."

"Determine inclusion based on the following conditions:"

- "- It studies a specific supplement (e.g., ashwagandha, magnesium)"
- "- The supplement is the intervention (not a background variable)"
- "- The subjects are humans (not animals or in vitro)"
- "- It tracks real health outcomes (e.g., anxiety, inflammation)"
- "- The abstract clearly describes methods and results"
- "- Or if it discusses the supplement's relationship with a disease"

"For each question, reply with:"

- "- 'Yes' and include the specific detail (e.g., supplement name, population)"
- "- 'No' if not present"
- "- 'NA' if not applicable or unclear"

"Output Format:"

- "1. Supplement: <specific supplement name> / <No> / <NA>"
- "2. Intervention: <Yes> / <No> / <NA>"
- "3. Human_Sample: <Yes> / <No> / <NA> "
- "4. Outcomes: <specific outcome> / <No> / <NA>"
- "5. Method_Result: <Yes> / <No> / <NA>"
- "6. Relationship: <specific disease> / <No> / <NA>"
- "7. Decision - Is this abstract relevant?: <YES or NO>"
- "8. Reasoning: <Concise one-sentence explanation>"

APPENDIX B

Prompt Used to Generate the Supplement-Disease Relationships for Each LLM

As a medical researcher, generate a table of supplement-disease relationships.

For each supplement:

1. List all human clinical evidence relationships you are aware of, including preliminary or mixed findings. Do not stop at 3–4 per supplement unless there are truly no more known relationships, continue until no more exist. Do not omit low confidence ones, give all of them, the table should be comprehensive.

2. Relationship type ("improves" or "worsens")

3. Confidence level ("High", "Moderate", or "Low")

Format as CSV with these EXACT columns: "Supplement","Disease","Relationship","Confidence"

Rules:

- Use medically validated relationships
- Be specific with disease names
- Include both positive and negative relationships
- Assign confidence based on evidence strength:
 - * High = multiple clinical studies
 - * Medium = some clinical evidence
 - * Low = preliminary evidence

Example row: "magnesium","migraine","improves","High"

Output ONLY the CSV data with header, no commentary.