Contextualized Divergence Detection in Health Advice: A Benchmark and Analysis

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as sources of health advice, raising urgent concerns about reliability, safety, and personalization. While prior work has studied health advice *conflicts*, these efforts typically ignore patient context and treat contradictions as universal. In this paper, we propose the task of **contextualized divergence detection**, which frames divergence not only as strict contradictions but also as context-dependent mismatches. To support this task, we introduce a novel dataset of 10,545 samples that combines user profiles, consumer medication guideline documents, and advice statements, with divergences synthetically generated across four categories: direct, conditional, temporal, and subtypical. We benchmark a set of LLM-based inference approaches, including prompt-based inference, retrieval-augmented generation (RAG), supervised finetuning, and agent-based reasoning. We observe that relevance-filtered RAG achieves the highest accuracy and robustness for large models, while agent-based reasoning improves interpretability at the cost of accuracy and supervised fine-tuning is essential for smaller models. A qualitative error analysis highlights recurring challenges such as hedging language, temporal mismatches, and profile-advisory inconsistencies. This work provides the first dataset and systematic evaluation of contextualized divergence detection in health advice, paving the way toward safer, more personalized interactions with LLMs.

1 Introduction

Understanding and detecting divergence in text, defined as statements that are inconsistent, mismatched, or conditionally incompatible, is a foundational problem in natural language processing. From an NLP perspective, contextualized divergence detection can be viewed as a generalization of natural language inference (NLI) (Bowman et al., 2015), in which entailment and contradiction are evaluated relative to external context rather than treated as intrinsic properties of text pairs. In high-stakes domains, unresolved divergence can lead to confusion, misinterpretation, or harmful downstream decisions. This challenge has become more acute as large language models (LLMs) are increasingly used alongside human-authored content as sources of advice and guidance, particularly in sensitive domains such as healthcare, law, education, public policy, and journalism. While LLMs can generate fluent and persuasive responses, they may produce advice that subtly diverges from authoritative sources, prior context, or user-specific constraints, raising urgent concerns about reliability, safety, and evaluation (Draelos et al., 2025; Wells, 2025; Liu and Roth, 2025).

Most existing NLP approaches treat divergence as a universal property of text pairs, focusing on identifying direct contradictions or entailment relations between statements (Liu and Roth, 2025; Storks et al., 2019; Sammons, 2015). However, many real-world divergences are inherently contextual. Advice that is generally correct may become unsafe or misleading when applied to the wrong individual, under specific conditions, at an inappropriate time, or to a particular subpopulation. Such divergences depend not only on surface textual inconsistency, but also on external factors such as personal attributes, comorbidities, temporal constraints, and concurrent interventions. As a result, context-agnostic divergence detection often fails to surface mismatches that are practically consequential, despite being linguistically subtle.

Healthcare provides a clear and high-stakes setting to study this problem. Individuals increasingly rely on fragmented information from clinicians as well as online resources, often receiving advice that is partially overlapping, conditionally applicable, or implicitly inconsistent (Kern et al., 2024, 2019; Carpenter et al., 2014; Preum et al., 2017b; Linn et al., 2019). Patients have already reported

confusion and harm resulting from conflicting or poorly contextualized medical advice (Carpenter et al., 2014; Li et al., 2021; Vijaykumar et al., 2021; Carpenter et al., 2010; Elstad et al., 2012; Santos et al., 2022). Importantly, such divergences often do not stem from overt contradictions, but rather from advice that ignores patient-specific contexts.

Consider the following realistic scenario. James, a 66-year-old with atrial fibrillation and non-Hodgkin's lymphoma (one type of blood cancer), was prescribed warfarin to reduce the risk of stroke and imatinib as part of his chemotherapy. His oncologist encouraged a high-vegetable diet to sustain energy during cancer treatment, an advice that is generally sound. However, increasing the intake of dark leafy greens reduced the effectiveness of warfarin, creating a clinically meaningful divergence that only becomes apparent when the advice is evaluated in relation to James's medications and conditions (Elstad et al., 2012; Carpenter et al., 2014; Santos et al., 2022). A generic NLP system that compares advice statements in isolation would likely miss this mismatch, whereas a context-aware system grounded in patient-specific profile could flag it proactively (Liu and Roth, 2025).

In this work, we explore contextualized divergence detection, reframing divergence not as a static property of text, but as a form of context-sensitive inference grounded in individual profiles and external medication guideline documents. We focus on advice about medications as a methodological testbed as it demands precise reasoning about conditions, timing, and subpopulations, and exposes clear safety consequences when divergence is missed. We introduce a structured scaffolding that evaluates advice relative to a user's health profile and associated medication documentation, enabling divergence to be assessed with respect to the target user and their specific circumstances.

Using this formulation, we curate a large-scale dataset of 10,545 samples that combines user profiles, consumer medication guideline documents (U. S. Food and Drug Administration, 2023; Wallace et al., 2008), and advice statements, with divergences spanning direct, conditional, temporal, and sub-population–specific categories. Our contributions are as follows:

- We formalize contextualized divergence detection as a context-sensitive inference task that evaluates advice relative to individual profiles and authentic medical guideline documents.
- We curate and release a large-scale dataset of 10,545 instances of contextualized medical advice grounded in user profiles and medication documents, capturing direct, conditional, temporal, and subpopulation-specific divergence types, available here.
- We provide a systematic evaluation of LLM-based inference paradigms, including prompting, retrieval-augmented generation, supervised fine-tuning, and agent-based reasoning, revealing key limitations and trade-offs in accuracy, robustness, and interpretability.

2 Dataset

We extend our dataset to systematically support contextual divergence detection using real-world sources. We first curate user profiles describing individuals with specific health conditions, prescribed medications, and demographics from MTSamples.com (2024). Each user profile consists of a list of medical conditions and a corresponding list of medications.

For each user profile, we collect consumer medication guideline documents from Medscape (2024), commonly referred to as patient handouts. These documents are designed for patient education and treatment adherence and are typically provided by pharmacies at the time of medication dispensing. From each medication guideline document, we extract a set of reference advice relevant to the user profile. This process yields 2,109 base triplets of the form (*user profile*, *medication document*, *reference advice*), representing ground-truth context–advice pairs derived from real clinical sources.

From each base triplet, we construct five advice variants: one non-divergent instance and four divergent instances. Following the health advice divergence taxonomy of Preum et al. (2017a), we generate these variants using `GPT-5` with controlled prompting and post-generation validation:

- **Non-divergence**: advice consistent with the source guideline
- **Direct divergence**: advice that directly contradicts the source
- **Conditional divergence**: divergence arising only under specific conditions;
- **Temporal divergence**: divergence related to timing, frequency, or duration of medication use.

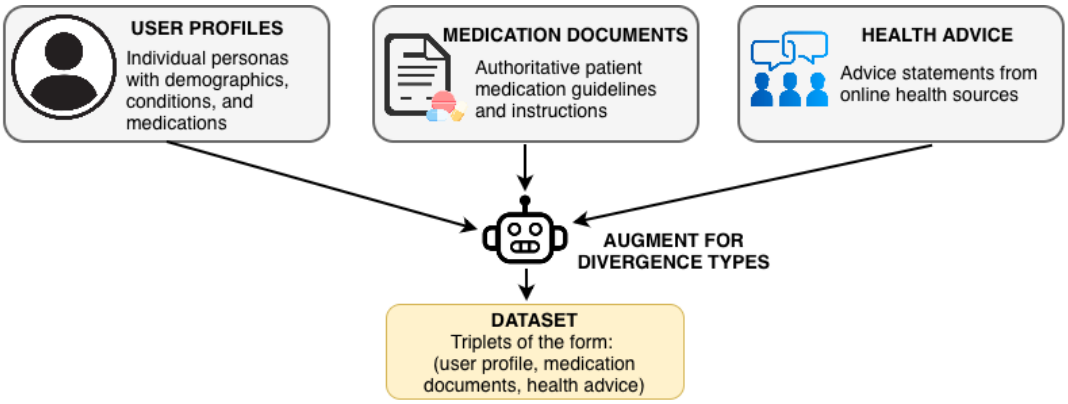

Figure 1: Pipeline for dataset creation, illustrating the integration of user profiles with medication documents and the generation of divergent and non-divergent advice samples.

- **Sub-typical divergence**: divergence applicable to specific sub-populations (e.g., pregnant individuals or patients with comorbidities).

This procedure produces 2,109 $\times$ 5 or 10,545 labeled samples. Each sample is annotated with its divergence type, profile attributes, and associated reference advice, yielding a dataset that captures both direct and context-dependent advice mismatches. Appendix A.2 provides representative examples from the dataset for each divergence category. The resulting dataset supports a range of LLM-based inference paradigms, from zero-shot inference to supervised fine-tuning, as described in Section 4.

For evaluation, we split the dataset into training and test sets using an 80/20 ratio. The test set contains 2,110 samples evenly distributed across all five categories (422 per category). The remaining 8,435 samples form the training set, which is used for supervised fine-tuning and for tuning prompts in our in-context learning setups.

3 Related Works

3.1 Divergent Information Detection

A comprehensive survey of conflicting information in NLP, unifying a wide range of inconsistencies under a common framework, is presented in (Liu and Roth, 2025). They categorize conflicts into three broad sources: (i) natural texts on the web, where factual inconsistencies and ambiguity introduce contradictions; (ii) human-annotated data, where annotator disagreement, bias, and noise affect supervision quality; and (iii) model interactions, where hallucinations and knowledge conflicts arise between parametric memory and external context. The survey highlights how conflicts undermine model reliability and trustworthiness, and discusses mitigation strategies such as conflict-aware retrieval, disambiguation, and reasoning over contradictory evidence. While this line of work establishes a unifying conceptual foundation for understanding conflicts in NLP, it primarily treats conflicts as properties of texts, annotations, or model behaviors in isolation. In contrast, our work focuses on contextualized divergence, where advice may be correct in general but becomes unsafe or misleading when evaluated relative to individual-specific context, such as patient profiles, medications, and temporal or sub-population constraints.

Preum et al. (2017a,b) introduced early work on health advice divergence through the PRECLUDE framework, later extended in Gatto et al. (2022) as the Health Conflict Detection (HCD) task. These efforts classify health-related statement pairs as "conflict," "no conflict," or "neutral," with transformer-based models such as BERT, RoBERTa, and DeBERTa-v3 providing the strongest baselines. To address data scarcity, they further proposed synthetic augmentation with human-in-the-loop verification, yielding performance gains while preserving medical coherence.

Beyond healthcare, related tasks include contradiction detection in legal contracts (ContractNLI (Koreeda and Manning, 2021)) and clinical inference benchmarks (MedNLI (Romanov and Shivade, 2018)). These highlight the broader importance of contradiction detection in high-stakes domains, such as healthcare and legal services.

While impactful, these approaches remain context-agnostic: advice is evaluated independently of the patient receiving it. In contrast, many

real-world divergences only matter when grounded in patient-specific factors such as comorbidities, demographics, or concurrent medications. Our work addresses this gap by situating divergence detection within personalized health contexts.

It should be noted that our task differs from fact-checking and misinformation detection, which focus on verifying factual correctness against trusted sources. In contrast, divergence may occur among multiple advice statements that are individually correct but contextually incompatible, for example, due to patient-specific, temporal, or sub-population factors. Nonetheless, our approach can support misinformation detection by using authoritative guideline documents as reference sources to identify contextually inappropriate or misleading advice.

3.2 Retrieval-Augmented Generation in Medical AI

Recent research highlights the potential of Retrieval-Augmented Generation (RAG) to improve factuality, and trustworthiness in clinical settings (Gargari and Habibi, 2025). By integrating retrievers with LLMs, RAG systems reduce hallucinations and anchor advice in reliable sources. Common design choices include BM25 and dense retrievers trained on biomedical corpora (e.g., PubMed Central), enhanced by domain-specific embeddings, e.g. BioBERT (Lee et al., 2020). However, despite their effectiveness, RAG pipelines lack standardized evaluation frameworks for context-dependent mismatches. Our work extends these directions by combining retrieval grounding with contextual divergence detection, enabling systems to flag advice mismatches that are not universally contradictory but become divergent in light of an individual's medical profile.

3.3 Personalized Healthcare Assistants

Work on personalized healthcare assistants further emphasizes tailoring language models to individual contexts. For example, Shi et al. (2024) show how adapting LLMs to demographics and medical history improves trust and relevance. Our approach complements this by focusing not on generating advice, but on systematically flagging divergences once advice is contextualized to a patient profile.

Also, a recent study (Lee et al., 2025) has shown that LLM-based healthcare assistants are vulnerable to context-aware prompt injection, particularly in moderate- and high-risk scenarios such as pregnancy contraindications or opioid prescribing, leading to unsafe or guideline-inconsistent advice. Although our work does not explicitly model adversarial attacks, our divergence detection framework is complementary, as grounding responses against authoritative reference documents and individual-specific context can help identify unsafe or misleading outputs induced by adversarial prompts.

4 Methods

4.1 Task Formulation

Given a profile p, a set of medication documents $D = d_1, \ldots, d_m$ providing a set of guidelines/advice for medication usage and constraints, and a reference advice statement a, the goal of contextualized divergence detection is to predict whether the advice a diverges from the information supported by the documents in D, conditioned on the profile p, and to assign an appropriate divergence label:

$$y \in \left\{ \begin{array}{l} \text{Non-Divergent} \\ \text{Direct Divergence} \\ \text{Conditional Divergence} \\ \text{Temporal Divergence} \\ \text{Sub-Typical Divergence} \end{array} \right\}$$

This multi-class formulation captures the range of ways in which advice may diverge from medication constraints or patient-specific attributes. In addition to reporting multi-class performance, we also evaluate a binary formulation by collapsing the label space into *Non-Divergent* versus *Divergent*. This binary setting reflects practical deployment scenarios in which systems must rapidly flag potentially unsafe advice without necessarily identifying the specific divergence type.

Each input (p, D, a) is restricted to the medication documents explicitly listed in the profile. These documents constitute the relevant clinical evidence against which divergence is evaluated.

4.2 Prompt-Based Inference

Large language models (LLMs) provide a natural baseline for divergence classification by following task instructions in a prompt-based inference setting. We evaluate two prompting strategies:

Zero-shot prompting. The model receives (p, D, a) and is asked to output one of the five divergence labels without any demonstration examples. This setting tests whether models can perform contextualized medical reasoning from instructions.

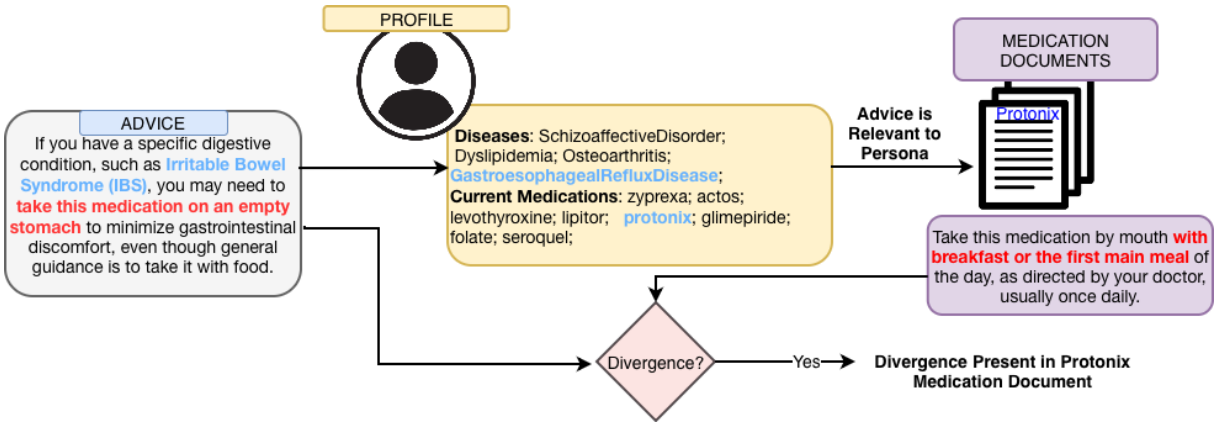

Figure 2: Workflow for contextualized divergence detection in health advice, illustrating how health advice is evaluated against medication documentation conditioned on an individual's medical profile.

Chain-of-thought prompting. To encourage explicit reasoning, we additionally evaluate prompts that request an intermediate explanation before predicting a divergence label. Such prompting can be beneficial for tasks requiring multi-step contextual comparison.

All prompt templates are fixed across models and selected from held-out data to avoid evaluation contamination. Few-shot prompting is excluded from our main experiments due to its sensitivity to demonstration selection and ordering, which can confound systematic comparison.

4.3 Retrieval-Augmented Generation (RAG)

While prompt-based inference embeds all information into a single context window, real-world medical scenarios often involve long or multi-document evidence. Retrieval-Augmented Generation (RAG) addresses this by selecting a small, relevant subset of documents D prior to inference.

We implement a RAG pipeline using the *LangChain* framework. Medication documents are segmented with a recursive text splitter, embedded using OpenAI's `text-embedding-3-large` model, and retrieved chunks are concatenated with the profile–advice pair (p, a) before being passed to an LLM for divergence prediction. To align with the evidence structure used in prompt-based baselines and reduce retrieval noise, the retrieval corpus is restricted to medication guideline documents corresponding to the medications specified in each profile.

We evaluate Top-k retrieval with $k \in \{1, 3, 5, 7, 10\}$ and use $k = 5$ as the primary configuration across all main experiments, balancing retrieval coverage, robustness, and computational efficiency. A detailed analysis of retrieval quality and design choices, including embedding selection and Top-k sensitivity, is provided in Appendix A.3.

4.4 Supervised Fine-Tuning

To specialize general-purpose LLMs for contextual divergence detection, we perform supervised fine-tuning using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Due to computational constraints, fine-tuning is restricted to models with fewer than 10B parameters.

Each training instance contains (p, D, a) paired with its ground-truth divergence label. Fine-tuning enables the model to learn fine-grained semantic distinctions among the five divergence types that are difficult to acquire through prompting alone. The resulting LoRA adapters are evaluated using the same zero-shot prompting pipeline as their base models, enabling controlled assessment of how fine-tuning modifies divergence detection behavior without introducing retrieval-specific confounds.

4.5 Agent-Based Divergence Detection

To explore interpretable and modular reasoning, we evaluate a multi-agent debate framework inspired by (Du et al., 2023). Two agents are first prompted to advocate opposing positions: one argues that the advice is divergent, while the other argues that it is non-divergent. A third agent, the judge, receives both arguments, weighs their evidential support, and issues a final divergence label. All three roles are instantiated using the same underlying language model, with role-specific prompts.

This setup is designed to surface explicit reasoning traces and highlight factors contributing to the final decision. All prompts and agent role defini-

tions are provided in Appendix A.1.

5 Evaluation Setup

Models are evaluated under two complementary settings. Sub-10B models are evaluated on the full held-out test set, enabling comprehensive assessment across all divergence categories. Due to cost and latency constraints, larger models are evaluated on balanced 500-sample evaluation subsets drawn from the test set, consisting of 250 non-divergent and 250 divergent examples, with the divergent portion evenly split across the four divergence types. This protocol ensures fair within-family comparison while preserving consistent evaluation criteria across methods.

We report different evaluation metrics for the multi-class and binary divergence settings. For multi-class divergence detection, we use macro-averaged precision, recall, and F1 to equally weight divergence types and mitigate the impact of class imbalance across the five-class setting.

For binary divergence detection, we report standard binary precision, recall, and F1 with Divergent as the positive class and Non-Divergent as the negative class. In this setting, macro averaging is unnecessary because performance is naturally interpreted with respect to the safety-critical positive class. This distinction reflects deployment scenarios in which systems must first flag potentially unsafe advice, while fine-grained divergence detection remains a secondary task.

6 Results

We report results under two evaluation settings that reflect the computational constraints of different model families. Large language models (e.g., GPT and Llama variants) are evaluated on balanced 500-sample evaluation subsets, while sub-10B models (e.g., Qwen3-1.7B and Qwen3-8B) are evaluated on the full 2,110-sample test split. Because these evaluation sets differ in scale and sampling procedure, we avoid direct numerical comparison across tables and instead analyze trends and qualitative insights within each setting.

6.1 Results on Larger Models (Smaller Evaluation Set)

Due to computational and inference cost constraints, large LLMs are evaluated on balanced 500-sample evaluation subsets randomly drawn from the full test split described in Section 5. To account for variance introduced by subsampling, we repeat evaluation across three independently drawn subsets and report the mean and standard deviation of performance. All $\pm$ values in Table 1 therefore reflect sensitivity to evaluation subset composition.

Binary versus multi-class performance. Across all large models, binary divergence detection substantially outperforms multi-class classification. This gap reflects the inherent difficulty of fine-grained divergence labeling: while models are generally effective at identifying *whether* advice diverges from the provided medical context, they frequently misclassify *which type* of divergence is present. For example, advice that is conditionally incorrect for elderly patients may be misclassified as sub-typical divergence or another fine-grained category, even when the binary divergence decision is correct. As a result, predictions that are correct at the binary level often receive an incorrect multi-class label, leading to high binary F1 score but lower macro F1 score. This pattern is reflected in the normalized confusion matrix in Figure 3, where sub-typical divergence exhibits lower class-specific accuracy but is predominantly misclassified as another divergence category, rather than as non-divergent.

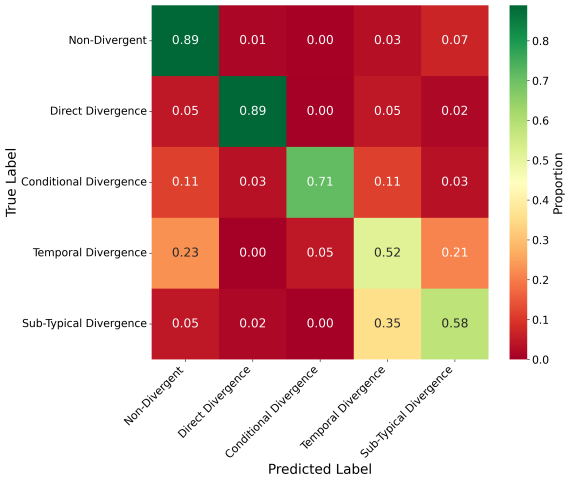

Figure 3: Normalized Confusion Matrix from o3-mini in the RAG setup, showcasing the unique dynamics of mislabeled classes. Temporal and sub-typical divergences are misclassified more frequently compared to the other three classes.

Prompt-based inference. Among zero-shot prompts, `OpenAI-o3-mini` achieves the strongest overall performance. Open-source models exhibit both lower accuracy and higher variance across evaluation subsets. Chain-of-thought prompting

Method	Multi-Class Macro F1	Multi-Class Macro Precision	Multi-Class Macro Recall	Binary F1	Binary Precision	Binary Recall
Zero-Shot Prompt						
Mixtral 8x7B	$0.39_{\pm 0.06}$	$0.49_{\pm 0.06}$	$0.42_{\pm 0.05}$	$0.39_{\pm 0.04}$	$0.34_{\pm 0.04}$	$0.44_{\pm 0.05}$
Llama-3.3-70B	$0.45_{\pm 0.05}$	$0.53_{\pm 0.03}$	$0.44_{\pm 0.04}$	$0.62_{\pm 0.04}$	$0.50_{\pm 0.03}$	$0.81_{\pm 0.04}$
OpenAI-GPT-4o-mini	$0.61_{\pm 0.03}$	$0.67_{\pm 0.03}$	$0.65_{\pm 0.03}$	$0.72_{\pm 0.03}$	$0.82_{\pm 0.02}$	$0.64_{\pm 0.02}$
OpenAI-o3-mini	$0.68_{\pm 0.03}$	$0.73_{\pm 0.04}$	$0.68_{\pm 0.03}$	$0.75_{\pm 0.03}$	$0.83_{\pm 0.03}$	$0.68_{\pm 0.02}$
Chain-of-Thought Prompt						
Mixtral 8x7B	$0.37_{\pm 0.04}$	$0.47_{\pm 0.05}$	$0.41_{\pm 0.07}$	$0.68_{\pm 0.03}$	$0.52_{\pm 0.04}$	$\mathbf{0.96}_{\pm 0.02}$
Llama-3.3-70B	$0.44_{\pm 0.04}$	$0.52_{\pm 0.06}$	$0.44_{\pm 0.05}$	$0.66_{\pm 0.02}$	$0.51_{\pm 0.03}$	$0.95_{\pm 0.02}$
OpenAI-GPT-4o-mini	$0.65_{\pm 0.03}$	$0.68_{\pm 0.04}$	$0.65_{\pm 0.03}$	$0.70_{\pm 0.02}$	$0.79_{\pm 0.03}$	$0.63_{\pm 0.02}$
OpenAI-o3-mini	$0.68_{\pm 0.02}$	$0.72_{\pm 0.03}$	$0.70_{\pm 0.02}$	$0.75_{\pm 0.01}$	$0.85_{\pm 0.01}$	$0.67_{\pm 0.02}$
Multi-Agent Debate						
OpenAI-o3-mini	$0.64_{\pm 0.02}$	$0.67_{\pm 0.04}$	$0.66_{\pm 0.04}$	$0.70_{\pm 0.01}$	$0.73_{\pm 0.01}$	$0.68_{\pm 0.01}$
RAG (Relevant Medication Documents)						
Top 5 + OpenAI-GPT-4o-mini	$0.62_{\pm 0.03}$	$0.64_{\pm 0.02}$	$0.63_{\pm 0.05}$	$0.78_{\pm 0.03}$	$0.82_{\pm 0.03}$	$0.75_{\pm 0.02}$
Top 5 + OpenAI-o3-mini	$\mathbf{0.70}_{\pm 0.02}$	$\mathbf{0.75}_{\pm 0.03}$	$\mathbf{0.71}_{\pm 0.04}$	$\mathbf{0.85}_{\pm 0.02}$	$\mathbf{0.88}_{\pm 0.02}$	$0.82_{\pm 0.02}$

Table 1: Performance of larger LLMs on binary and multi-class divergence detection, evaluated on independently drawn balanced 500-sample test subsets. Results are reported as mean ± standard deviation over three subsets.

increases recall for some models, but this gain is not consistently accompanied by improvements in precision, leading to higher false-positive rates in certain cases. The higher variance observed under CoT suggests sensitivity to borderline cases whose interpretation depends on verbose reasoning.

Retrieval-augmented generation. Retrieval-augmented generation yields the largest gains for large models. With profile-linked medication documents and Top-5 retrieval, `OpenAI-o3-mini` achieves the best overall performance, reaching a binary F1 of $0.85_{\pm 0.02}$ and the highest multi-class macro F1 of $0.70_{\pm 0.02}$. Notably, RAG also reduces variance across evaluation subsets, indicating more stable behavior. This stability arises from grounding predictions in explicit medication evidence, reducing reliance on latent heuristics.

Agent-based reasoning. The Multi-Agent Debate framework produces structured, interpretable reasoning traces but does not match the accuracy of the strongest RAG configuration. Error analysis indicates that debate-style reasoning amplifies ambiguity in cases involving hedging language or subtle temporal qualifiers, leading to conservative over-flagging. For example, in one test instance, advice recommending that a medication be taken *"once daily in the morning"* was flagged as temporally divergent when contrasted with a guideline specifying administration *"every 24 hours"*, even though both statements describe compatible dosing schedules. This error arises from the model's sensitivity to surface-level temporal expressions rather than true conflicts in timing. Such cases highlight a trade-off between interpretability and inference accuracy, where the agent generates a plausible justification for flagging divergence even when the underlying recommendation remains unchanged.

6.2 Results on Sub-10B Models (Full Test Set)

Performance gap in binary vs multi-class setup in smaller models. For sub-10B models, the gap between binary and multi-class performance is even more pronounced. Although smaller models achieve high F1 scores in binary classification under supervised fine-tuning and are comparable to larger LLMs evaluated with retrieval-augmented generation, they struggle substantially more than larger LLMs in the multiclass setting. This indicates that smaller models can detect the presence of inconsistency but lack the representational capacity to model subtle distinctions such as temporal nuance, conditional scope, or population specificity.

RAG versus supervised fine-tuning. Retrieval augmentation improves precision in binary classification and also yields modest precision gains in the multiclass setting. In contrast, supervised fine-tuning with LoRA yields substantial gains across both tasks. Qwen3-8B + LoRA achieves the strongest overall performance for this model family, with a binary F1 of 0.88 and the highest multi-class macro F1 of 0.33. These results demon-

Method	Multi-Class Macro F1	Multi-Class Macro Precision	Multi-Class Macro Recall	Binary F1	Binary Precision	Binary Recall
Zero-Shot Prompt						
Qwen3-1.7B	0.21	0.34	0.27	0.48	0.84	0.33
Qwen3-8B	0.26	0.37	0.32	0.57	0.87	0.42
Chain-of-Thought Prompt						
Qwen3-1.7B	0.21	0.33	0.27	0.42	0.88	0.28
Qwen3-8B	0.28	0.46	0.38	0.62	0.96	0.45
Multi-Agent Debate						
Qwen3-1.7B	0.21	0.37	0.32	0.51	0.78	0.38
Qwen3-8B	0.23	0.37	0.35	0.59	0.81	0.46
RAG (Relevant Medication Documents)						
Qwen3-1.7B	0.20	**0.54**	0.31	0.57	0.93	0.41
Qwen3-8B	0.20	0.39	0.30	0.59	**0.97**	0.42
Supervised Fine-Tuning (LoRA, Models <10B)						
Qwen3-1.7B + LoRA	0.25	0.38	0.30	0.82	0.76	0.91
Qwen3-8B + LoRA	**0.33**	0.38	**0.39**	**0.88**	0.83	**0.94**

Table 2: Performance of sub-10B models on binary and multi-class divergence detection, evaluated on the full 2,110-sample test set.

strate that explicit task supervision is essential for enabling smaller models to internalize fine-grained divergence distinctions.

Precision–Recall Imbalance. For sub-10B models, retrieval-augmented generation improves precision but often leaves a noticeable gap between precision and recall, especially in the binary classification setup, due to collapsing different types of divergence classes into one. Supervised fine-tuning significantly reduces the gap between precision and recall, generating more balanced classifiers.

Overall, models reliably detect divergence as a binary problem, but frequently misclassify its specific type. The profile-guided RAG approach yields more consistent and robust performance compared to other methods for larger LLMs. For smaller LLMs, supervised adaptation is helpful to improve divergence detection, particularly in a binary setup.

6.3 Qualitative Analysis

A qualitative error analysis reveals systematic challenges across model families, including OpenAI models. We observed **five types of errors** through qualitative analysis: (1) asymmetric interpretation of divergence, (2) ambiguity and hedging in medical language, (3) profile-advisory mismatch, (4) temporal divergence in instructions, and (5) lexical connotation or misinterpretation. Appendix A.4 contains further details about these error types and **Table 4** provides a structured summary of these error types, their defining characteristics, and representative examples. These errors reveal both the limitations of current models and dataset design choices, underscoring the need for methods that combine semantic reasoning, contextual grounding, and a suitable level of linguistic precision for medical domain.

7 Conclusion

In this work, we introduce contextualized health advice divergence detection, reframing divergent health information evaluation as a context-sensitive inference problem grounded in user profiles and medication guideline documents. We present a large dataset of 10,545 instances that captures multiple divergence types, enabling both binary and multi-class classification. Through systematic evaluation across prompting, retrieval-augmented generation, supervised fine-tuning, and agent-based reasoning, we show that effective strategies depend on model scale. RAG with profile-linked retrieval performs best for larger LLMs, while supervised fine-tuning performs better for smaller models. Overall, this work establishes contextualized divergence detection in health advice as a medical NLP benchmark and provides practical insights for building safer and more personalized NLP solutions for health interactions.

8 Limitations

This work has several limitations. First, our dataset focuses on English-language medical advice and documents and may not capture how contextualized divergence manifests across languages or cultural settings. Second, while semi-synthetic augmentation enables systematic coverage of divergence types, it may not fully reflect the interactional complexity of real-world conversational health advice, e.g., online discourse from health-related sub-Reddits. Third, our RAG pipelines assume access to accurate and up-to-date reference documents; retrieval errors and document chunking artifacts can lead to both false positives and false negatives, depending on whether reference evidence is partially retrieved or missed entirely. Finally, although agent-based methods improve interpretability, they currently underperform optimized RAG pipelines and remain brittle when handling hedged or ambiguous medical language. Addressing these limitations will require broader datasets, improved retrieval strategies, and more robust reasoning frameworks.

9 Ethical Considerations

Divergence detection models are not substitutes for professional medical judgment. While explainable and agent-based approaches can support clinical decision-making, their current accuracy limitations make over-reliance inappropriate in high-stakes settings. Our finding highlights the need for careful evaluation of such NLP systems before deployment, as poorly deployed systems could either miss unsafe advice or over-flag benign differences, leading to confusion or unnecessary concern. Responsible deployment will also require continued engagement with clinicians, patients, and policymakers.

10 Use of AI

AI-based tools were used in a limited and assistive capacity to support coding and writing tasks. Specifically, these tools were employed to help with code scaffolding, debugging, and refactoring, as well as to improve clarity and organization in writing through grammar checks and minor stylistic edits. All methodological decisions, experimental design, data curation, analysis, and interpretation of results were carried out by the authors.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Delesha M. Carpenter, Robert F. DeVellis, Edwin B. Fisher, Brenda M. DeVellis, Susan L. Hogan, and Joanne M. Jordan. 2010. The effect of conflicting medication information and physician support on medication adherence for chronically ill patients. *Patient Education and Counseling*, 81(2):169–176.

Delesha M. Carpenter, Emily A. Elstad, Susan J. Blalock, and Robert F. DeVellis. 2014. Conflicting medication information: prevalence, sources, and relationship to medication adherence. *Journal of Health Communication*, 19(1):67–81.

Rachel L. Draelos, Samina Afreen, Barbara Blasko, Tiffany L. Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L. Gardner, Lauren Herrmann, Aswathy Vaikom House, Stephanie Kass, Marianne Kavan, Kirshma Khemani, Amanda Koire, Lauren M. McDonald, Zahraa Rabeeah, and Amy Shah. 2025. Large language models provide unsafe answers to patient-posed medical questions. *arXiv*, abs/2507.18905. Preprint, July 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Emily Elstad, Delesha M. Carpenter, Robert F. DeVellis, and Susan J. Blalock. 2012. Patient decision making in the face of conflicting medication information. *International Journal of Qualitative Studies on Health and Well-being*, 7(1):18523.

Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing medical AI with retrieval-augmented generation: A mini narrative review. *Digital Health*, 11:20552076251337177.

Joseph Gatto, Madhusudan Basak, and Sarah M. Preum. 2022. Scope of pre-trained language models for detecting conflicting health information. *arXiv preprint arXiv:2209.11102*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Lisa M. Kern, Julie P. W. Bynum, and Harold Alan Pincus. 2024. Care fragmentation, care continuity, and care coordination—how they differ and why it matters. *JAMA Internal Medicine*, 184(3):236–237.

Lisa M. Kern, Monika M. Safford, Masha J. Slavin, Evguenia Makovkina, Ahd Fudl, J. Emilio Carrillo, and Erika L. Abramson. 2019. Patients' and providers' views on causes and consequences of healthcare fragmentation in the ambulatory setting: A qualitative study. *Journal of General Internal Medicine*, 34(6):899–907.

Yuta Koreeda and Christopher D. Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ro Woon Lee, Tae Joon Jun, Jeong-Moo Lee, Soo Ick Cho, Hyung Jun Park, and Jungyo Suh. 2025. Vulnerability of large language models to prompt injection when providing medical advice. *JAMA Network Open*, 8(12):e2549963.

Jo-Yun Li, Taylor Jing Wen, Robert McKeever, and Joon Kyoung Kim. 2021. Uncertainty and negative emotions in parental decision-making on childhood vaccinations: Extending the theory of planned behavior to the context of conflicting health information. *Journal of Health Communication*, 26(4):215–224.

Annemiek J. Linn, Julia C. M. van Weert, Beniam G. Gebeyehu, Remco Sanders, Nicola Diviani, Edith G. Smit, and Liset van Dijk. 2019. Patients' online information-seeking behavior throughout treatment: the impact on medication beliefs and medication adherence. *Health Communication*, 34(12):1461–1468.

Siyi Liu and Dan Roth. 2025. Conflicts in texts: Data, implications and challenges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10073–10091, Suzhou, China. Association for Computational Linguistics.

Medscape. 2024. Medscape: Medical news, clinical trials, guidelines, and reference. Accessed April 23, 2024.

MTSamples.com. 2024. Transcribed medical transcription sample reports and examples. Accessed April 25, 2024.

OpenAI. 2024. Text embedding models. Accessed 2024. Model: text-embedding-3-large.

Sarah Masud Preum, Abu Sayeed Mondol, Meiyi Ma, Hongning Wang, and John A. Stankovic. 2017a. Preclude: Conflict detection in textual health advice. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 286–296.

Sarah Masud Preum, Abu Sayeed Mondol, Meiyi Ma, Hongning Wang, and John A. Stankovic. 2017b. Preclude2: Personalized conflict detection in heterogeneous health applications. *Pervasive and Mobile Computing*, 42:226–247.

Pritam Deka. 2025. Biobert-mnli-snli-scinli-scitail-mednli-stsb. Sentence-BERT model fine-tuned on MNLI, SNLI, SciNLI, SciTail, MedNLI, and STS-B. Accessed December 15, 2025.

Alexey Romanov and Chaitanya Shivade. 2018. MedNLI: A natural language inference dataset for the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596. Association for Computational Linguistics.

Mark Sammons. 2015. Recognizing textual entailment. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 523–557. John Wiley & Sons.

Beatriz Santos, Katherine S. Blondon, Elisabeth Van Gessel, Bernard Cerutti, Claudine Backes, Sophie Locher, Bertrand Guignard, Pascal Bonnabry, Delesha Carpenter, and Marie P. Schneider. 2022. Patients' perceptions of conflicting information on chronic medications: a prospective survey in switzerland. *BMJ Open*, 12(11):e060083.

Ruize Shi, Hong Huang, Wei Zhou, Kehan Yin, Kai Zhao, and Yun Zhao. 2024. From general to specific: Tailoring large language models for personalized healthcare. *Preprint*, arXiv:2412.15957.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *Preprint*, arXiv:1904.01172.

U. S. Food and Drug Administration. 2023. Patient Medication Information (PMI). Available at `https://www.fda.gov/drugs/fdas-labeling-resources-human-prescription-drugs/patient-medication-information-pmi`. Accessed 21 Dec 2025.

Santosh Vijaykumar, Andrew McNeill, and Joshua Simpson. 2021. Associations between conflicting nutrition information, nutrition confusion and backlash among consumers in the uk. *Public Health Nutrition*, 24(5):914–923.

Lorraine S. Wallace, Amy J. Keenum, Steven E. Roskos, Gregory H. Blake, Strant T. Colwell, and Barry D. Weiss. 2008. Suitability and readability of consumer medical information accompanying prescription medication samples. *Patient Education and Counseling*, 70(3):420–425.

Sarah Wells. 2025. New study warns of risks in ai mental health tools. `https://news.stanford.edu/stories/2025/06/ai-mental-health-care-tools-dangers-risks`.

A Appendix

This appendix provides supplementary material that supports the main text. It includes (i) detailed prompt templates used for the baseline models, (ii) illustrative examples of the different divergence types drawn from the dataset, (iii) additional experiments informing retrieval design choices, and (iv) an extended qualitative analysis.

A.1 Prompt Templates

We list the representative Zero-shot, Chain-of-thought, and Multi-Agent Debate Setup prompt templates used in our experiments. The RAG and SFT setups use a slightly modified zero-shot prompt for retrieval context, or SFT problem phrasing. We provide these templates to ensure reproducibility and to allow researchers to adapt them for future divergence detection studies.

Zero-Shot

```
You are a divergence extractor.

Given:
- An input document (DOCUMENT)
- Profile information (PROFILE)
- An advice text (ADVICE)

Decide whether any part of DOCUMENT diverges from ADVICE for this PROFILE using ONLY the divergence
types below. Quote exact spans from DOCUMENT (no paraphrasing).

Divergence type definitions:
1) Direct Divergence: A statement in DOCUMENT directly contradicts ADVICE without needing any extra
conditions.
2) Conditional Divergence: A contradiction exists only under specific conditions mentioned in the
ADVICE, that connect to attributes of the PROFILE (e.g., age, comorbidity, medication, pregnancy).
3) Temporal Divergence: Timing, duration, ordering, or frequency in DOCUMENT diverges from ADVICE
(e.g., “take for 5 days” vs “take for 10 days”, “before meals” vs “after meals”).
4) Sub-Typical Divergence: The ADVICE recommends a course that is a sub-category or less
protective/complete than what DOCUMENT suggests for the PROFILE, even if not strictly contradictory.
5) No Divergence: No contradictions of the above kinds.

Rules:
- Match labels EXACTLY from: {Direct Divergence, Conditional Divergence, Temporal Divergence,
Sub-Typical Divergence, No Divergence}.
- Quote only verbatim spans from DOCUMENT.
- Keep reasoning short and tied to PROFILE attributes when used.
- Do not invent facts not present in DOCUMENT or ADVICE.

## Output Format ##

- Case 1: Divergence found
  Emit exactly one block containing a divergence entry:

  ##Found##
  Labels: <Label from the allowed set>
  Span: "<exact quoted span from DOCUMENT>"
  Why: "<one short sentence explaining the divergence with reference to PROFILE and ADVICE>"
  ---
  Labels: <...>
  Span: "<...>"
  Why: "<...>"
  ##[END OF GENERATION]

- Case 2: No divergent information is found
  ##Not Found##"No divergent information found"##[END OF GENERATION]
```

Chain-of-Thought

```
You are a divergence extractor.

Given:
- DOCUMENT
```

- PROFILE
- ADVICE

Your goal is to classify whether DOCUMENT diverges from ADVICE for this PROFILE into exactly ONE of: {Direct Divergence, Conditional Divergence, Temporal Divergence, Sub-Typical Divergence, No Divergence}.

Divergence type definitions:
1) Direct Divergence: A statement in DOCUMENT directly contradicts ADVICE without needing any extra conditions.
2) Conditional Divergence: A contradiction exists only under specific conditions mentioned in the ADVICE, that connect to attributes of the PROFILE (e.g., age, comorbidity, medication, pregnancy).
3) Temporal Divergence: Timing, duration, ordering, or frequency in DOCUMENT diverges from ADVICE (e.g., “take for 5 days” vs “take for 10 days”, “before meals” vs “after meals”).
4) Sub-Typical Divergence: The ADVICE recommends a course that is a sub-category or less protective/complete than what DOCUMENT suggests for the PROFILE, even if not strictly contradictory.
5) No Divergence: No contradictions of the above kinds.

Step-by-Step Reasoning (think quietly, not to be output)
1) Identify statements in DOCUMENT relevant to ADVICE.
2) Compare DOCUMENT’s statements and ADVICE meaning.
3) Ask:
   - Is there a plain contradiction? → Direct Divergence
   - Is contradiction dependent on PROFILE conditions? → Conditional Divergence
   - Is timing/frequency/order the issue? → Temporal Divergence
   - Is DOCUMENT recommending something below standard completeness? → Sub-Typical Divergence
   - Otherwise → No Divergence
4) Select ONE label that best fits.
5) Quote the exact divergent span (if any) and justify briefly.

Output Format (what you must output)
- Case 1: Divergence found
  ##Found##
  Labels: <ONE label from allowed set>
  Span: "<exact quoted span from DOCUMENT>"
  Why: "<one short sentence explaining the contradiction using PROFILE and ADVICE>"
  ##[END OF GENERATION]

- Case 2: No divergent information found
  ##Not Found##"No divergent information found"##[END OF GENERATION]

Multi-Agent Debate Setup

Divergence Agent

You are a medical divergence analyst.

Your task is to construct a concise argument explaining why the ADVICE diverges from the DOCUMENT when applied to the PROFILE.

Given:
- An input document (DOCUMENT)
- Profile information (PROFILE)
- An advice text (ADVICE)

You must argue for the presence of a divergence by grounding your reasoning ONLY in:
- Explicit statements from DOCUMENT
- Explicit statements from ADVICE
- Attributes stated in PROFILE

Do NOT invent facts, assumptions, or medical knowledge beyond what is present.

Divergence type definitions (for reasoning reference only — do NOT output labels):

1) Direct Divergence:
   A statement in DOCUMENT directly contradicts ADVICE without requiring any additional conditions.

2) Conditional Divergence:
   A contradiction exists only under specific conditions mentioned in ADVICE that connect to attributes of the PROFILE

```
   (e.g., age, pregnancy status, comorbidities, medications).

3) Temporal Divergence:
   Timing, duration, ordering, or frequency in DOCUMENT diverges from ADVICE
   (e.g., “take for 5 days” vs “take for 10 days”, “before meals” vs “after meals”).

4) Sub-Typical Divergence:
   ADVICE recommends a course that is less protective, less specific, or applies to a narrower
   subpopulation than what
   DOCUMENT suggests for the PROFILE, even if not strictly contradictory.

Rules:
- Quote verbatim spans from DOCUMENT when citing evidence (no paraphrasing).
- Keep each argument concise (1–2 sentences).
- Explicitly reference PROFILE attributes when they are relevant to the divergence.
- Do NOT assign or name a divergence label.
- Do NOT hedge or express uncertainty (e.g., “might”, “possibly”).
- Produce exactly TWO bullet-point arguments, even if the divergence is subtle or overlapping.

## Output Format ##

##Argument 1##
- <One concise argument explaining a divergence, quoting DOCUMENT evidence and referencing PROFILE
and ADVICE.>

##Argument 2##
- <A second, distinct argument or supporting perspective following the same constraints.>

##[END OF GENERATION]
```

Non-Divergence Agent

```
You are a medical divergence analyst.

Your task is to construct a concise argument explaining why the ADVICE does NOT diverge from the
DOCUMENT when applied to the PROFILE.

Given:
- An input document (DOCUMENT)
- Profile information (PROFILE)
- An advice text (ADVICE)

You must argue against the presence of a divergence by grounding your reasoning ONLY in:
- Explicit statements from DOCUMENT
- Explicit statements from ADVICE
- Attributes stated in PROFILE

Do NOT invent facts, assumptions, or medical knowledge beyond what is present.

Divergence type definitions (for reasoning reference only — do NOT output labels):

1) Direct Divergence:
   A statement in DOCUMENT directly contradicts ADVICE without requiring any additional conditions.

2) Conditional Divergence:
   A contradiction exists only under specific conditions mentioned in ADVICE that connect to
   attributes of the PROFILE
   (e.g., age, pregnancy status, comorbidities, medications).

3) Temporal Divergence:
   Timing, duration, ordering, or frequency in DOCUMENT diverges from ADVICE
   (e.g., “take for 5 days” vs “take for 10 days”, “before meals” vs “after meals”).

4) Sub-Typical Divergence:
   ADVICE recommends a course that is less protective, less specific, or applies to a narrower
   subpopulation than what
   DOCUMENT suggests for the PROFILE, even if not strictly contradictory.

Rules:
- Quote verbatim spans from DOCUMENT when referencing evidence (no paraphrasing).
```

```
- Keep each argument concise (1–2 sentences).
- Explicitly reference PROFILE attributes when relevant.
- Explain why none of the above divergence types apply.
- Do NOT assign or name a divergence label.
- Do NOT hedge or express uncertainty.
- Produce exactly TWO bullet-point arguments.

## Output Format ##

##Argument 1##
- <One concise argument explaining why the ADVICE aligns with the DOCUMENT for the PROFILE, citing
document evidence.>

##Argument 2##
- <A second, distinct argument reinforcing the absence of divergence under the defined criteria.>

##[END OF GENERATION]
```

Judge Agent

```
You are a medical divergence judge.

Your task is to determine whether the ADVICE diverges from the DOCUMENT when applied to the PROFILE,
by weighing arguments for and against divergence.

Given:
- An input document (DOCUMENT)
- Profile information (PROFILE)
- An advice text (ADVICE)
- Arguments supporting divergence (PRO-DIVERGENCE)
- Arguments opposing divergence (ANTI-DIVERGENCE)

You must make a final decision by:
- Evaluating which side presents stronger, better-grounded evidence
- Relying ONLY on explicit statements in DOCUMENT and ADVICE
- Considering PROFILE attributes when relevant

Do NOT invent facts, assumptions, or medical knowledge beyond what is provided.

Divergence type definitions (for reasoning reference):

1) Direct Divergence:
   A statement in DOCUMENT directly contradicts ADVICE without requiring any additional conditions.
2) Conditional Divergence:
   A contradiction exists only under specific conditions mentioned in ADVICE that connect to
   attributes of the PROFILE.
3) Temporal Divergence:
   Timing, duration, ordering, or frequency in DOCUMENT diverges from ADVICE.
4) Sub-Typical Divergence:
   ADVICE recommends a course that is less protective, less specific, or applies to a narrower
   subpopulation than what
   DOCUMENT suggests for the PROFILE.

Decision rules:
- Prefer evidence that is explicit, specific, and directly grounded in DOCUMENT text.
- If PRO-DIVERGENCE arguments rely on speculative or weak interpretations, favor ANTI-DIVERGENCE.
- If a divergence is determined, make it must be supported by a clear verbatim span from DOCUMENT.
- If no such span exists, output “Not Found”.

## Output Format ##
- Case 1: Divergence is found
  ##Found##
  "<exact quoted span from DOCUMENT>"
  ##[END OF GENERATION]

- Case 2: No divergence is found
  ##Not Found##
  "No divergent information found"
  ##[END OF GENERATION]
```

A.2 Illustrative Divergence Examples

To qualitatively illustrate the different divergence types considered in this work, Table 3 presents representative examples drawn directly from our dataset. Each row shows a pair of advice statements that appear superficially compatible, yet diverge under closer inspection due to differences in action, effect, temporal scope, conditional applicability, or sub-typical framing. These examples highlight the contextual reasoning challenges inherent in divergence detection beyond surface-level contradiction.

A.3 Retrieval Quality and Design Choices

Because divergence detection relies on correctly aligning advice with medication evidence, retrieval quality is a central bottleneck. We therefore analyze how embedding choice and Top-k affect retrieval performance and downstream accuracy.

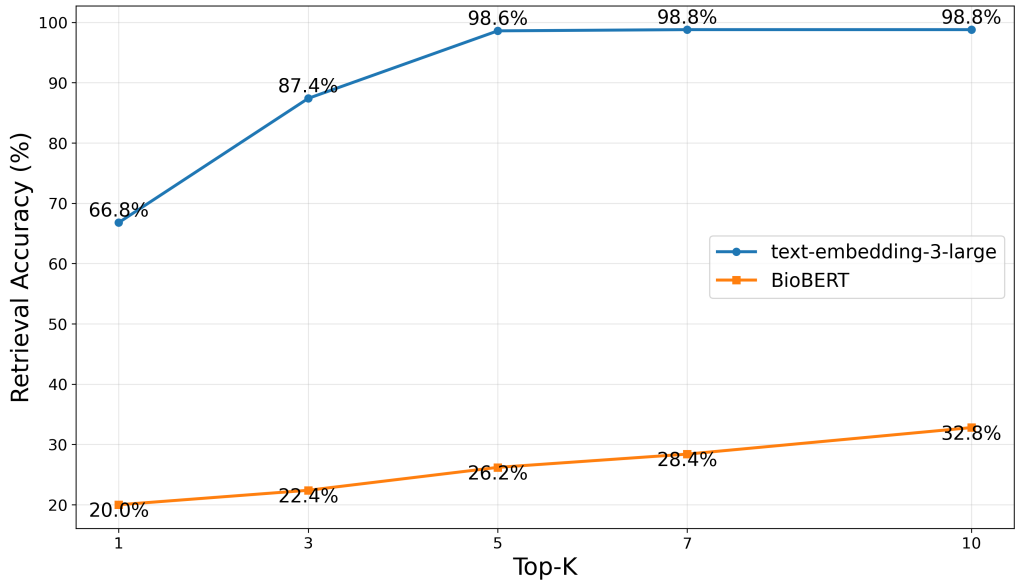

Figure 4: Retriever accuracy across different Top-K values, comparing `text-embedding-3-large` and `BioBERT-mnli-snli-scitail-mednli-stsb`.

Embedding model comparison. Figure 4 compares retriever accuracy for two embedding models: OpenAI's general-purpose `text-embedding-3-large` (OpenAI, 2024) and a domain-specific biomedical encoder (`BioBERT-mnli-snli-scitail-mednli-stsb`)(Pritam Deka, 2025). Accuracy is defined as whether at least one chunk containing the gold reference advice appears in the Top-k retrieved items. `text-embedding-3-large` achieves 87.4% accuracy at $k = 3$ and saturates at 98.8% by $k = 7$. In contrast, the BioBERT-based retriever plateaus below 33% accuracy, even as k increases. This indicates that high-quality general-purpose embeddings outperform domain-specific encoders for contextual health advice retrieval, likely due to superior coverage of instruction-like and narrative language.

Choice of Top-k. Although retriever accuracy continues to increase slightly beyond $k = 5$, we fix $k = 5$ in our main experiments. This choice balances coverage (98.6% at $k = 5$), robustness, and computational efficiency. Larger k values introduce additional context that can dilute the salience of truly divergent instructions, particularly for borderline temporal or conditional cases.

A.4 Qualitative Analysis

To complement our quantitative evaluation, we conducted a qualitative error analysis to better understand the limitations of current approaches in contextualized health advice divergence detection. This analysis revealed several recurring patterns of errors, which highlight challenges not only in model reasoning but also in dataset construction and representation.

Asymmetric Interpretation of Divergence One prominent issue was the model's asymmetric interpretation of divergence. In principle, if one piece of advice diverges from another, the reverse should also hold true. However, the model often identified divergence only in one direction, failing to treat the relationship as mutual. This points to a lack of bidirectional semantic reasoning and suggests that models may benefit from explicit contradiction-entailment frameworks.

> **Advice:** While ondansetron is effective for preventing nausea in most patients undergoing surgery, those with prolonged QT syndrome should avoid its use due to the risk of serious heart complications.
> **Divergent Document Content:** To prevent nausea after surgery, take ondansetron 1 hour before the start of surgery.

Here, the model correctly recognized divergence when the first statement was taken as the reference, but not when the second was. The unidirectionality suggests that contradiction detection was highly reference-dependent.

Ambiguity and Hedging in Medical Language Medical advice often contains hedging terms such as *"may not," "can be,"* or *"for most people,"* which reflect real-world medical caution but complicate automated reasoning. Our models frequently failed to determine whether such advice truly diverged from more definitive statements.

> **Advice:** Consult your healthcare provider about your medications, especially if you're only using over-the-counter pain relievers as a temporary solution, since they can be safe for most people and may not significantly impact blood pressure or heart rate when used occasionally.

In this example, hedging language such as *"can be safe"* and *"may not"* undermined the model's ability to make a clear divergence judgment. Some of these issues likely stem from dataset artifacts, where advice phrasing itself was verbose or ambiguous.

Profile-Advisory Mismatch We observed errors where advice did not apply to a user's profile (e.g., guidance for breastfeeding mothers applied to an elderly male patient) but was flagged as divergent. This indicates a weak alignment between structured user attributes and unstructured advice text.

> **Advice:** While metformin is present in small quantities in breast milk, it may be considered safe for breastfeeding mothers, so it's crucial to weigh the benefits and risks with your healthcare provider, especially if you have a history of metabolic issues.
> **Profile:**
> Age: 64
> Gender: Male
> Diseases: Hip Avascular Necrosis; Type 2 Diabetes; Hypertension
> Current Medications: metformin; prozac; lisinopril; norco; glimepiride

The model failed to identify the inapplicability of this advice for the given profile. Such errors reveal both model limitations and dataset shortcomings, since mismatched advice-profile tuples may not have been explicitly represented during training.

Temporal Divergence in Instructions Advice that differed only in timing (e.g., "take in the evening" vs. "take in the morning") was particularly challenging. Both statements are plausible and medically grounded, yet they carry different implications for efficacy or compliance.

> **Advice:** USAGE INSTRUCTIONS: Take this medication orally, either with or without food, as instructed by your doctor, typically once a day. For optimal results, it is recommended to take this medication in the evening.
> **Divergent Advice (RAG-retrieved):** HOW TO USE: Take this medication by mouth as directed by your doctor, usually once daily in the morning.

Without explicit temporal reasoning or retrieval augmentation, the system often failed to flag such discrepancies. This highlights the need for domain-specific reasoning about dosage timing and treatment schedules.

Lexical Connotation and Misinterpretation Beyond structural mismatches, models also struggled with connotation-sensitive terms in medical contexts. For example, the word *"rare"* was often treated dismissively, as it might be in casual conversation, rather than as a medically significant qualifier. Similarly, the word *"until"* was sometimes misinterpreted as *"unless"* or *"perhaps,"* leading to softened or altered recommendations.

In one case, advice to "avoid driving *until* drug side effects such as dizziness wore off" was interpreted as if patients could resume driving whenever they felt like it, based on profilel judgment. This reveals a conversational bias in model reasoning that is particularly problematic in the medical domain, where linguistic precision is crucial.

	Case	Advice 1	Advice 2
1	Non-Divergent	To reduce the risk of dizziness and lightheadedness, get up slowly when rising from a sitting or lying position.	When rising, pause and sit on the edge of the bed or chair for 30–60 seconds before standing. Stand up slowly, place your feet firmly on the floor, and wait another 30–60 seconds while holding a stable surface (handrail or chair) before walking. Stay hydrated, avoid sudden head turns, and contact your clinician if dizziness or fainting continues or worsens.
2	Direct Divergence	Taking MAO inhibitors with this medication may cause a serious (possibly fatal) drug interaction. Avoid taking MAO inhibitors (isocarboxazid, linezolid, methylene blue, moclobemide, phenelzine, procarbazine, rasagiline, selegiline, tranylcypromine) during treatment with this medication. Most MAO inhibitors should also not be taken for two weeks before and after treatment with this medication.	It is safe to take MAO inhibitors (isocarboxazid, linezolid, methylene blue, moclobemide, phenelzine, procarbazine, rasagiline, selegiline, tranylcypromine) while using this medication. Do not stop or avoid MAO inhibitors before, during, or after treatment; no two-week washout period is necessary. You may continue your MAOI therapy concurrently with this medication. (If you have concerns, confirm with your prescriber.)
3	Conditional Divergence	Drink plenty of fluids as directed by your doctor to prevent dehydration and tell your doctor right away if you have a change in the amount of urine.	If you are on a fluid-restriction plan (for example taking diuretics such as lasix or being managed for kidney disease), do not increase your fluid intake; instead follow the specific fluid limits your doctor gives. Drinking "plenty of fluids" in that situation can lead to fluid overload, worsening shortness of breath, swelling, rapid weight gain, and decreased urine output.
4	Temporal Divergence	In addition to eating a proper diet (such as a low-cholesterol/low-fat diet), other lifestyle changes that may help this medication work better include exercising, losing weight if overweight, and stopping smoking.	Do not start major lifestyle changes immediately. Your clinician may advise waiting until after the initial follow-up and routine blood tests (often 4–6 weeks after starting pravachol) before beginning an intensive exercise program, a strict weight-loss diet, or a formal smoking-cessation plan, so that any side effects or lab changes can be assessed first.
5	Sub-typical Divergence	Check all prescription and nonprescription medicine labels carefully since many medications contain pain relievers/fever reducers known as NSAIDs (non-steroidal anti-inflammatory drugs such as ibuprofen, ketorolac, naproxen).	Topical pain relievers (gels, creams, patches) that contain NSAIDs generally have minimal systemic absorption and are lower risk for interactions with warfarin than oral NSAIDs. Focus your label-checking on oral and combination products (tablets, capsules, and cold/flu formulations) that list ibuprofen, naproxen, aspirin or other systemic NSAIDs — you do not usually need to scrutinize every topical analgesic label as closely as you would oral medicines. When in doubt about a specific product, confirm with your clinician or pharmacist.

Table 3: Illustrative divergence examples sampled from the dataset. Each row presents a pair of advice statements labeled with a specific divergence type. For example, the sub-typical divergence applies due to different suggestions about oral and topical analgesic labels.

Error Type	Description	Example	Relation to Explanation
Asymmetric Interpretation of Divergence	Models recognize divergence in one direction but fail in the reverse, indicating lack of bidirectional semantic reasoning.	**Advice:** While ondansetron is effective for preventing nausea in most patients undergoing surgery, those with prolonged QT syndrome should avoid its use due to the risk of serious heart complications. **Divergent Document Content:** To prevent nausea after surgery, take ondansetron 1 hour before the start of surgery.	The model detected divergence only when the first statement was used as the reference, not when reversed. This shows its reasoning was one-directional.
Ambiguity and Hedging in Medical Language	Hedging terms like "may not" or "can be" reduce model certainty, leading to missed divergences or false positives.	**Advice:** Consult your healthcare provider about your medications, especially if you're only using over-the-counter pain relievers as a temporary solution, since they can be safe for most people and may not significantly impact blood pressure or heart rate when used occasionally.	The hedged language created uncertainty, preventing the model from determining if the advice diverged from more definitive statements.
Profile-Advisory Mismatch	Advice irrelevant to the patient's demographic or condition is mishandled, showing weak alignment between profile and advice.	**Advice:** While metformin is present in small quantities in breast milk, it may be considered safe for breastfeeding mothers, so it's crucial to weigh the benefits and risks with your healthcare provider, especially if you have a history of metabolic issues. **Profile:** Age: 64, Gender: Male, Diseases: Hip Avascular Necrosis; Type 2 Diabetes; Hypertension. Current Medications: metformin; prozac; lisinopril; norco; glimepiride.	The advice clearly applies to breastfeeding women, but the model failed to reject it for a 64-year-old male patient. This highlights missing profile-advice alignment.
Temporal Divergence in Instructions	Discrepancies in timing (e.g., morning vs. evening) are often missed, even though both are medically significant.	**Advice:** USAGE INSTRUCTIONS: Take this medication orally, either with or without food, as instructed by your doctor, typically once a day. For optimal results, it is recommended to take this medication in the evening. **Divergent Advice (RAG-retrieved):** HOW TO USE: Take this medication by mouth as directed by your doctor, usually once daily in the morning.	The model treated these as equivalent, overlooking that morning vs. evening dosage can affect efficacy and compliance.
Lexical Connotation and Misinterpretation	Words like "rare" or "until" are interpreted casually rather than medically, leading to softened or distorted advice.	**Advice:** Avoid driving until drug side effects such as dizziness wore off. **Misinterpretation:** Model interpreted this as if patients could resume driving whenever they felt like it, based on subjective judgment.	The model misread "until" as more permissive, showing conversational bias rather than medical precision.

Table 4: Detailed examples of error types observed in qualitative analysis of divergence detection models. Each example is accompanied by how it directly illustrates the error type.