Contextual Divergence Detection in Health Advice: A Benchmark and Analysis

Reet Kothari
Dartmouth College
reet.y.kothari.gr@dartmouth.edu

Sirajum Munir
Bosch Research
sirajum.munir@us.bosch.com

Madhusudan Basak
University of Massachusetts Amherst
madhusudan.buet@gmail.com

Sarah Preum
Dartmouth College
sarah.masud.preum@dartmouth.edu

Abstract

Large language models (LLMs) are increasingly used as sources of health advice, raising concerns about reliability, safety, and personalization. While prior work has studied conflicting/divergent health advice, these efforts typically ignore patient context. In this paper, we propose the task of **contextual divergence detection**. We introduce a novel dataset of 10,545 samples that combines user profiles, relevant medication guideline documents, and advice statements, and four types of synthetically generated contextually divergent health advice: 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 supervised fine-tuning is essential for smaller models. A qualitative error analysis highlights recurring challenges such as hedging, temporal mismatches, and profile-advisory inconsistencies. This work provides the first large dataset and systematic evaluation of contextual 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, contextual 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 contextual. Advice that is generally correct may become unsafe or misleading when applied to the wrong individual, under another condition, or at an inappropriate time. Such divergences depend not only on surface-level 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 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

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39035–39060
October 24-29, 2026 ©2026 Association for Computational Linguistics

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, a generally sound advice. 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 a patient-specific profile could flag it proactively (Liu and Roth, 2025).

In this work, we explore contextual 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 contextual 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 contextual 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). From each medication guideline document, we extract a set of reference advice (direct chunks) relevant to the user profile. This yields 2,109 base triplets of the form (*user profile*, *medication document*, *reference advice*), representing grounded context–advice pairs. Figure 1 illustrates the dataset construction pipeline.

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-divergent**: advice aligned with the guideline
- **Direct**: explicit contradiction between a pair of advice
- **Conditional**: divergence that arises only when the advice is evaluated under a patient-specific condition represented in the profile, such as a comorbidity, concurrent medication, age, or pregnancy status.
- **Temporal**: timing/frequency-related divergence
- **Sub-typical**: divergence arising because the advice narrows, specializes, or treats differently a subtype, formulation, or subpopulation covered by a broader recommendation,

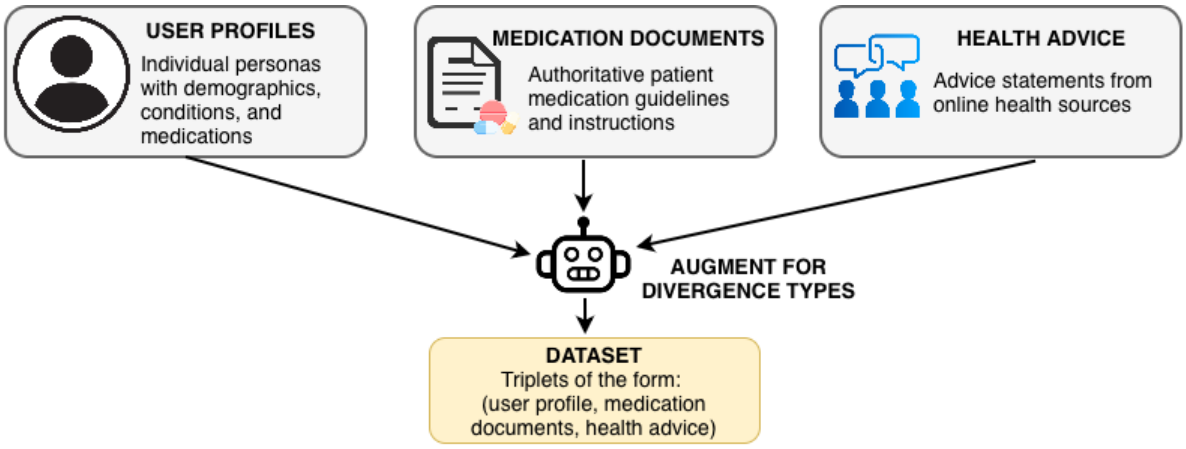


Figure 1: Pipeline for dataset creation, illustrating the integration of real user profiles with real medication documents and the generation of divergent and non-divergent advice samples.

e.g., non-dairy milk, pediatric or geriatric medication, population-specific therapy. Unlike conditional divergence, this distinction does not depend on an explicit patient-specific condition.

The base triplets are constructed through explicit medication-level grounding. Each clinical profile contains the patient's recorded conditions and prescribed medications. For each medication appearing in a profile, we retrieve the corresponding Medscape medication guideline document and extract reference advice from the document. Thus, the medication provides the direct link between the patient profile and source document, while the extracted reference advice is directly grounded in the same medication documentation. For profiles containing multiple medications, each medication is linked to its respective guideline before relevant advice is extracted.

The underlying profile–medication associations originate from the clinical notes, while medication documents and reference advice are linked through the medication-guideline matching procedure described above. Manual validation therefore focuses on the synthetically generated advice: each generated variant is checked for consistency with the corresponding profile and medication documentation and for correctness of the intended divergence label.

The dataset is grounded in real-world patient profiles, medication guideline documents, and real reference advice extracted directly from clinical notes and medication guidelines. Only the divergent variants are synthetically generated, as real divergence is rare, and using only real samples would prevent us from trying more data-intensive methods like finetuning. Each generated divergent advice undergoes a manual quality check to verify (i) consistency with the medication documents, and (ii) correctness of divergence labels. We also examine surface statistics such as advice length and lexical overlap to ensure that divergence categories cannot be trivially separated using shallow textual features.

To assess annotation reliability, two annotators independently labeled 100 stratified samples for binary and five-class divergence classification. We report Cohen's $\kappa = 0.97$ for the binary setting and $\kappa = 0.79$ for the multi-class setting. Class-wise agreement was highest for non-divergent (95%) and direct divergence (85%), and lower for conditional (70%), temporal (60%), and sub-typical (60%) divergence, reflecting the challenging nature of the context-dependent categories.

The overall 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 examples from the dataset for each divergence category.

For evaluation, we use an 80/20 train–test split. The test set contains 2,110 samples evenly distributed across all five categories (422 per class), while the remaining 8,435 samples are used for supervised fine-tuning and prompt development. We perform the train–test split at the level of the 2,109 base profile–document–reference-advice triplets, before constructing the five advice variants. All

five variants derived from a given base triplet are therefore assigned to the same split, preventing direct leakage of synthetically related instances between training and evaluation. Because the same medication guideline may be associated with multiple patient profiles or reference advice instances, base-triplet separation does not necessarily imply medication-document separation.

Train–test overlap audit. Because the same medication guideline can support multiple patient profiles and reference advice statements, we additionally audit train–test overlap at the document–reference-advice level. All medication documents represented in the test set also occur in the training set, and thus our split is not document-disjoint. However, overlap of the exact (document, reference advice) pair is rare: only 7 of the 422 held-out base triplets (1.7%) contain a document–reference-advice pair also observed in training. Thus, 98.3% of held-out base triplets evaluate a document–reference-advice relation not observed during training. This substantially limits the possibility that performance can be explained by memorization of specific document–advice mappings, although models may have encountered the underlying medication guideline during training.

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 contextual 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.

While PRECLUDE2 is the closest prior work conceptually, our formulation differs in several important ways. First, PRECLUDE2 detects conflicts primarily between pairs of textual health advice and between advice and activities, whereas our task evaluates an advice statement against document-level medication guidance conditioned on a patient profile. Second, PRECLUDE2 uses a semantic rule-based approach based on linguistic rules and external knowledge bases, while our benchmark evaluates learned LLM-based methods over substantially longer evidence. Third, applying PRECLUDE2 directly to our benchmark would require constructing a new rule base for the profile–document–advice setting rather than simply evaluating an existing trained model. We therefore view PRECLUDE2 as an important conceptual predecessor, but not a directly reproducible baseline for the proposed benchmark.

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 contextual 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 the multi-class formulation, we evaluate a binary divergence setting by collapsing all four divergence categories into a single *Divergent* label while retaining the *Non-Divergent* class. Importantly, the binary results are computed from the same labeled instances used in the multi-class evaluation rather than through a separate dataset or evaluation procedure.

This binary setting reflects practical deployment scenarios in which systems must rapidly flag potentially unsafe advice without necessarily identifying the specific type of divergence. Figure 2 illustrates the contextual divergence detection workflow.

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 contextual medical reasoning from instructions.

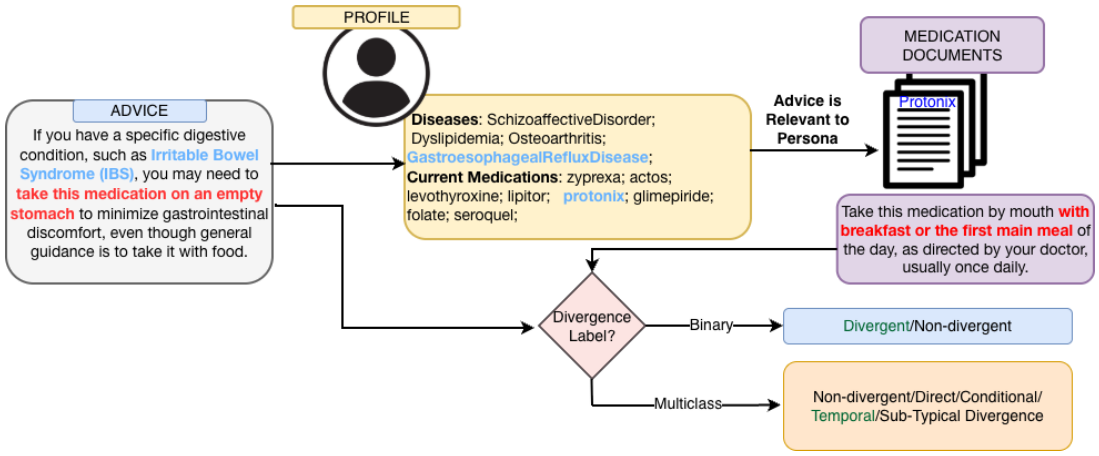


Figure 2: Workflow for contextual divergence detection in health advice, illustrating how health advice is evaluated against medication documentation conditioned on an individual's medical profile.

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, the structure used in prompt-based baselines and to reduce retrieval noise, the retrieval corpus is restricted to medication guideline documents corresponding to the medications specified in each profile.

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 support 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.

5 Evaluation Setup

Models are evaluated in two complementary settings. Sub-10B models are evaluated on the complete 2,110-sample held-out test set, enabling comprehensive assessment across all divergence cate-

gories. Due to cost and latency constraints, larger models are evaluated on three independently sampled, balanced 500-sample subsets drawn from the held-out test set. Each subset consists of 250 non-divergent and 250 divergent examples, with the divergent portion evenly distributed across the four divergence types. We report the mean and standard deviation across the three subsets for these experiments. Task formulation, labels, and evaluation metrics remain consistent in both settings, and comparisons are made primarily between methods evaluated under the same protocol.

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 score 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 evaluate the sub-10B models on the full 2,110-sample test split and large LLMs on a balanced 500-sample subset due to inference cost constraints. As these settings differ in scale and sampling procedure, we focus on trends within each setting rather than direct numerical comparison across them. We share the reproducibility plan in Section 9.

6.1 Results on Larger Models (Evaluation Set)

Large LLMs are evaluated on balanced 500-sample subsets from the full test split. To account for variance introduced by subsampling, we repeat evaluation across three independently drawn subsets and report the mean and standard deviation of performance in Table 1.

Binary versus multi-class performance. Across most models, binary divergence detection substantially outperforms five-way classification. Models are often able to recognize that advice conflicts with the medical context, but struggle to identify the precise divergence category. Figure 3 shows that direct and non-divergent cases are comparatively easier to classify, while temporal and sub-typical divergence remain the most challenging. These harder categories are frequently confused with other divergence types rather than with the non-divergent class, suggesting that models detect the presence of a mismatch but fail to localize its semantic source. Full per-class F1 breakdowns across methods are provided in Appendix A.5.

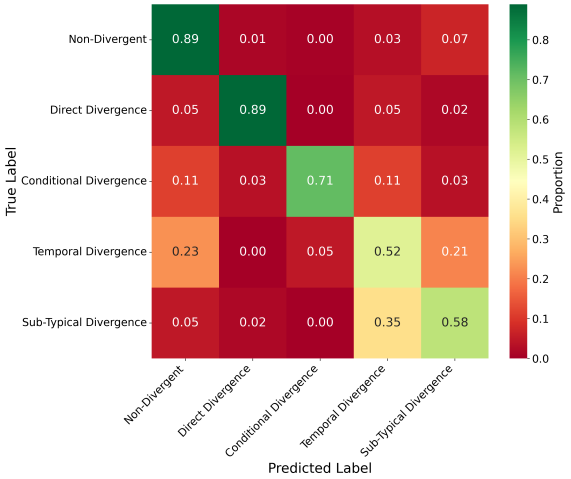


Figure 3: Normalized Confusion Matrix from o3-mini in the RAG setup, showcasing the unique dynamics of mislabeled classes. Temporal and sub-typical divergences are misclassified more frequently compared to the other three classes.

Prompt-based inference. Among prompt-only methods, `OpenAI-o3-mini` achieves the strongest overall performance, followed closely by `OpenAI-GPT-4o-mini`. Open-source models achieve substantially lower F1 than closed models across both settings.

For `Mixtral 8x7B` and `Llama-3.3-70B`, Chain-of-Thought (CoT) produces near-perfect recall but substantially lower precision in the binary setting, suggesting that explicit reasoning encourages conservative over-prediction of divergence in safety-sensitive settings.

Retrieval-augmented generation. RAG achieves the highest observed performance for large models. With profile-linked medication documents and Top-5 retrieval, OpenAI-o3-mini reaches a binary F1 of 0.85 ± 0.02 and a multi-class macro F1 of 0.70 ± 0.02. RAG particularly improves binary divergence detection relative to prompt-only inference, while improvements in multi-class performance are smaller. With

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.

profile-linked medication documents and Top-5 retrieval, OpenAI-o3-mini achieves the highest observed overall performance, reaching a binary F1 of 0.85 ± 0.02 and a multi-class macro F1 of 0.70 ± 0.02. Relative to prompt-only inference, retrieval grounding yields higher observed binary performance, while the differences in multi-class performance are smaller.

Agent-based reasoning. The Multi-Agent Debate framework produces more interpretable reasoning traces but underperforms the strongest RAG configuration. Error analysis suggests that debate-style reasoning amplifies ambiguity in cases involving hedging language or subtle temporal qualifiers.

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

Table 2 presents results for sub-10B models evaluated on the full 2,110-sample test set.

Binary versus multi-class performance in smaller models. For sub-10B models, the gap between binary and multi-class performance is even larger than for larger LLMs. Smaller models can often detect that divergence exists, but struggle to correctly identify the divergence category, particularly for temporal and sub-typical cases.

Under zero-shot prompting, smaller models exhibit high precision but low recall, indicating conservative prediction behavior. For example, MedGemma-4B achieves a binary precision of 0.97 but a recall of only 0.18. This imbalance is substantially reduced after supervised fine-tuning.

RAG and supervised fine-tuning. RAG improves binary precision for smaller Qwen models by grounding predictions in medication documentation when compared with zero-shot and Chain-of-Thought prompt. However, RAG did not yield consistent multi-class improvements for smaller models. Supervised fine-tuning with LoRA yields substantial improvements across both tasks in terms of F1, compared to zero-shot and Chain-of-Thought prompting. `Qwen3-8B + LoRA` achieves the strongest overall performance within this model family, reaching a binary F1 of 0.88 and a multi-class macro F1 of 0.33. Fine-tuning also dramatically improves recall, suggesting that smaller models benefit more from explicit task supervision than from retrieval grounding alone.

Although MedGemma benefits more noticeably from Multi-Agent Debate and RAG in the binary setting in terms of recall, its multi-class performance remains limited, indicating that domain specialization alone is insufficient for fine-grained contextual divergence reasoning.

6.3 Qualitative Analysis and Exploratory Evaluation on Reddit Data

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-advice mismatch, (4) temporal divergence in instructions, and (5) lexical connotation or misinterpretation. Appendix A.8 contains further details about these error types, and **Table 8** provides a structured summary of these error types, their defining characteristics, and rep-

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
MedGemma-4B	0.18	0.52	0.26	0.30	0.97	0.18
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
MedGemma-4B	0.21	0.46	0.25	0.58	0.82	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
MedGemma-4B	0.24	0.35	0.27	0.78	0.84	0.72
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
MedGemma-4B	0.26	0.36	0.29	0.77	0.82	0.74
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**
MedGemma-4B + LoRA	0.28	0.40	0.31	0.78	0.83	0.73

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

resentative 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 the medical domain.

While semi-synthetic augmentation enables systematic coverage of divergence types, it may not fully reflect the interactional complexity of conversational text. To investigate this, we conduct a small exploratory evaluation using 100 samples of patient-doctor discussion threads collected from Reddit health discussions on r/AskDocs in A.9. Real-world conversational datasets containing patient advice interactions are difficult to obtain due to privacy constraints. Although public forums such as Reddit contain related discussions, they require substantial pre-processing and annotation to ensure clinical validity. So we adopt a controlled benchmark setting as a practical compromise. The core challenge studied here remains invariant across structured and conversational settings, e.g., reasoning over profile-conditioned divergence between advice and medical guidance.

Computational Cost To provide an estimate of deployment cost, we measured token usage over 100 GPT-4o-mini inference calls. Each call processed an average of approximately 10,057 tokens, reflecting the relatively-retrieved *medication-document* context used by the task.

7 Conclusion

In this work, we introduce contextual 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 achieves the highest observed performance for larger LLMs, while supervised fine-tuning yields the strongest observed performance for smaller models. Overall, this work establishes contextual 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 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 when relevant evidence is partially retrieved or missed entirely.

Second, although agent-based methods improve interpretability, they currently underperform compared to optimized RAG pipelines and remain brittle when handling hedged or ambiguous medical language. Addressing these limitations will require improved retrieval strategies and more robust reasoning frameworks.

Future work can also extend this study to a large-scale analysis of contextual divergence in other platforms (e.g., social media) and other languages.

9 Reproducibility

For reproducibility, we release all resources required to replicate our experiments here. We provide prompt templates used for zero-shot, chain-of-thought, and agent-based inference (Appendix A.1). We also provide details of the retrieval configuration used in the RAG pipeline, including embedding model selection and Top-k retrieval settings (Appendix A.3).

10 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.

11 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, (iv) explanation of LoRA configurations, (v) per-method classwise F1 breakdown, (vi) cross-protocol evaluation for validity, (vii) ablation study, (viii) an extended qualitative analysis, and (ix) an exploration of the formulation in online health discourse.

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, 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##
    Labels: <ONE label from allowed set>
  "<exact quoted span from DOCUMENT>"
  ##[END OF GENERATION]

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

	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.

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. For example, general advice to "drink plenty of fluids" may become conditionally divergent for a patient with kidney disease who is under fluid restriction: the mismatch depends on the patient's profile. In contrast, if a guideline recommends avoiding NSAIDs broadly while the advice recommends avoiding only oral NSAIDs and treats topical NSAIDs differently, the divergence arises from a subtype distinction rather than from a patient-specific condition, and is therefore sub-typical.

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.

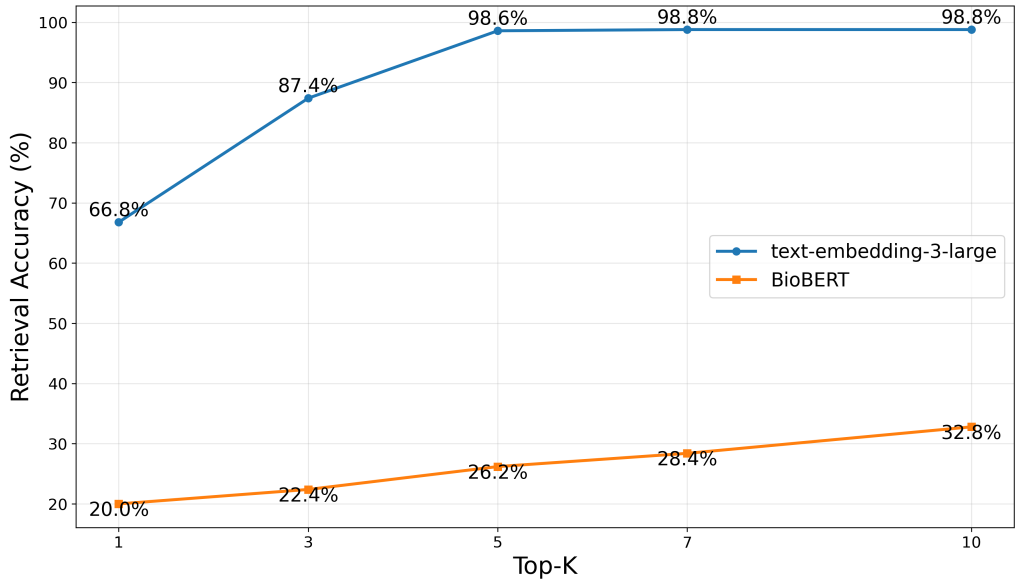


Figure 4: Retriever accuracy across different Top-K values, comparing `text-embedding-3-large` and `BioBERT-mnli-snli-scitail-mednli-stsb`.

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. Retrieval failures are rare (<2%) at $k = 5$, but when they occur they primarily affect conditional divergence cases that require precise grounding in medication documentation.

Chunking configuration. Medication documents are segmented using a recursive text splitter with a chunk size of 1000 and an overlap of 200. We keep this configuration fixed across retrieval experiments to ensure a consistent retrieval corpus when comparing embedding models and Top-k settings.

A.4 Supervised Fine-Tuning with LoRA

For supervised adaptation of sub-10B models, we use Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient fine-tuning method that injects small trainable low-rank matrices into selected layers of a pretrained model while keeping the original weights frozen. This approach substantially reduces the number of trainable parameters and memory requirements, enabling efficient fine-tuning while preserving the base model's general capabilities.

Adapter configuration. LoRA adapters are applied to both attention and feed-forward projections within each transformer block. Specifically, adapters are inserted into the query, key, value, and output projections of the attention mechanism as well as the gate, up, and down projections of the feed-forward network. This configuration allows LoRA to adapt both contextual reasoning and intermediate feature transformations.

We use a rank of $r = 8$ for the low-rank update matrices. This provides sufficient representational capacity while maintaining a small parameter footprint. The scaling factor is set to $\alpha = 16$, which controls the magnitude of the LoRA updates relative to the frozen base weights. A dropout rate of 0.05 is applied to the LoRA update path to reduce overfitting.

Bias parameters are not adapted during fine-tuning, following common practice in parameter-efficient LLM adaptation.

Training configuration. Fine-tuning is performed using supervised instruction tuning on the training split of the dataset. Due to memory constraints when adapting multi-billion-parameter models, training uses a per-device batch size of 1 with gradient accumulation over 8 steps, resulting in an effective batch size of 8 per device.

The learning rate is set to 2×10^{-4}, which is typical for LoRA-based supervised fine-tuning where only adapter parameters are updated. Training is performed for one epoch over the dataset, which we found sufficient for effective task adaptation without overfitting.

Mixed-precision training is used to reduce memory consumption and improve throughput.

Training procedure. Each training example consists of the tuple (p, D, a), where p denotes the user profile, D the relevant medication documents, and a the advice statement. The model is trained to predict the corresponding divergence label using the same prompt formulation used during inference. This ensures that the fine-tuned adapters specialize the model for contextual divergence classification while remaining compatible with the prompting pipeline used in evaluation.

A.5 Class-wise F1 scores

Method	Non-Divergent	Direct	Conditional	Temporal	Sub-Typical
Zero-Shot Prompt					
OpenAI-o3-mini	0.78	0.71	0.70	0.49	0.43
MedGemma-4B	0.37	0.08	0.14	0.24	0.06
Chain-of-Thought Prompt					
OpenAI-o3-mini	0.8	0.69	0.75	0.48	0.49
MedGemma-4B	0.31	0.22	0.30	0.19	0.02
Multi-Agent Debate					
OpenAI-o3-mini	0.73	0.74	0.76	0.48	0.42
MedGemma-4B	0.34	0.32	0.30	0.17	0.06
RAG (Relevant Medication Documents)					
OpenAI-o3-mini	0.89	0.90	0.81	0.48	0.55
MedGemma-4B	0.28	0.33	0.44	0.17	0.06
Supervised Fine-Tuning (LoRA, Models $<$10B)					
MedGemma-4B	0.32	0.34	0.44	0.16	0.14

Table 4: Per-class F1 score for divergence detection, evaluated on a 500 sample subset for o3-mini and full 2,110 test set for MedGemma

Table 4 reports per-class F1 scores across representative inference paradigms. Across methods, non-divergent, direct, and conditional divergence cases are consistently the easiest categories to classify, particularly for `OpenAI-o3-mini`, where RAG achieves F1 scores of 0.89, 0.90, and 0.81, respectively. In contrast, temporal and sub-typical divergence remain substantially more challenging across all inference settings. Even under the strongest RAG configuration, `OpenAI-o3-mini` achieves only 0.48 F1 scores on temporal divergence and 0.55 on sub-typical divergence. These categories frequently require implicit reasoning over timing constraints, subgroup applicability, or nuanced profile-specific conditions that are not always explicitly stated in the advice text.

The smaller `MedGemma-4B` model exhibits substantially lower per-class performance overall. Note that `MedGemma-4B` is evaluated on the full 2,110-sample test set, whereas OpenAI-o3-mini results are reported on a balanced 500-sample subset, making direct numerical comparison between the two inappropriate. Within `MedGemma-4B`'s own results, while supervised fine-tuning and retrieval grounding improve conditional divergence detection, temporal and sub-typical classes remain difficult, reinforcing the broader observation that smaller models struggle with fine-grained contextual reasoning even when domain-specialized.

			Multi-Class			Binary		
Model	**Method**	**Protocol**	**Macro F1**	**Macro Prec.**	**Macro Rec.**	**F1**	**Prec.**	**Rec.**
Qwen3-8B	Zero-Shot	3×500	$0.20_{\pm 0.04}$	$0.31_{\pm 0.05}$	$0.28_{\pm 0.05}$	$0.40_{\pm 0.04}$	$0.56_{\pm 0.04}$	$0.31_{\pm 0.04}$
Qwen3-8B	Zero-Shot	Full 2,110	0.26	0.37	0.32	0.57	0.87	0.42
Qwen3-8B	CoT	3×500	$0.21_{\pm 0.05}$	$0.35_{\pm 0.03}$	$0.28_{\pm 0.04}$	$0.38_{\pm 0.03}$	$0.52_{\pm 0.04}$	$0.30_{\pm 0.03}$
Qwen3-8B	CoT	Full 2,110	0.28	0.46	0.38	0.62	0.96	0.45
Mixtral 8×7B	Zero-Shot	3×500	$0.39_{\pm 0.06}$	$0.49_{\pm 0.06}$	$0.42_{\pm 0.05}$	$0.39_{\pm 0.04}$	$0.34_{\pm 0.04}$	$0.44_{\pm 0.05}$
Mixtral 8×7B	Zero-Shot	Full 2,110	0.43	0.45	0.45	0.49	0.41	0.61
Mixtral 8×7B	CoT	3×500	$0.37_{\pm 0.04}$	$0.47_{\pm 0.05}$	$0.41_{\pm 0.07}$	$0.68_{\pm 0.03}$	$0.52_{\pm 0.04}$	$0.96_{\pm 0.02}$
Mixtral 8×7B	CoT	Full 2,110	0.42	0.45	0.44	0.49	0.55	0.44

Table 5: Cross-protocol evaluation of Qwen3-8B and Mixtral 8×7B. The 3×500 results report mean $\pm$ standard deviation across three independently sampled balanced subsets, while the Full 2,110 results are computed deterministically over the complete held-out test set.

A.6 Cross-Protocol Evaluation

To assess whether the main findings are sensitive to differences in evaluation protocol, we compare Qwen3-8B and Mixtral 8×7B under both evaluation settings. Specifically, we compare performance on the three independently sampled balanced 500-example subsets with performance on the complete 2,110-example held-out test set. The task formulation, prompts, labels, and evaluation metrics remain unchanged across the two settings.

Across protocols, the multi-class results are broadly consistent. Qwen3-8B remains relatively weak on fine-grained classification under both settings, while Mixtral 8×7B consistently achieves higher macro F1. For example, Mixtral achieves a zero-shot macro F1 of 0.39 ± 0.06 under the sampled protocol and 0.43 on the full test set, while Qwen3-8B achieves 0.20 ± 0.04 and 0.26, respectively. These cross-protocol results suggest that the main multi-class trends are not solely attributable to differences in evaluation sample size.

Setting	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
No Persona	0.466	0.407	0.378	0.524	0.524	0.524
No Documents	0.317	0.292	0.270	0.540	0.540	0.540

Table 6: Contextual grounding ablation results. *No Persona* removes the user profile while retaining the medication documentation and advice; *No Documents* removes the medication documentation while retaining the user profile and advice.

Class	No Documents	No Persona
No Divergence	0.696	0.652
Direct Divergence	0.244	0.653
Conditional Divergence	0.415	0.319
Temporal Divergence	0.000	0.178
Sub-Typical Divergence	0.000	0.090

Table 7: Per-class F1 scores for the contextual grounding ablations.

A.7 Contextual Grounding Ablation

To evaluate whether divergence detection relies on contextual grounding, we ablate the two primary sources of context: the user persona and the medication documentation. In the *No Persona* setting, the model receives only the medication documentation and advice, while in the *No Documents* setting, it receives only the user persona and advice.

Table 6 reports the overall results. Removing the persona reduces macro F1 to 0.378, while removing the medication documentation results in a larger degradation to 0.270. These results indicate that both contextual sources contribute to fine-grained divergence detection, with medication documentation having a particularly strong effect.

The per-class results in Table 7 further reflect the information required by the definitions of the divergence categories. Direct Divergence remains relatively detectable without the persona (0.653 F1), as an explicit contradiction can often be identified by comparing the advice directly with the medication documentation. Conditional Divergence, in contrast, depends on conditions represented in the user profile, and its performance decreases to 0.319 F1 when that information is removed.

The complementary pattern is particularly clear for Temporal and Sub-Typical Divergence. When the medication documentation is removed, both categories fall to 0.000 F1. These categories require establishing how the advice differs from the underlying recommendation, such as its timing or the subtype of a recommendation, making the source documentation essential for determining the corresponding divergence. Conditional Divergence remains partially detectable without the documentation (0.415 F1), consistent with the persona retaining patient-specific conditions that can provide useful signals, although the source recommendation needed to establish the divergence is absent. No Divergence remains comparatively detectable under both ablations.

Takeaway. The class-level ablations broadly follow the information requirements of the divergence taxonomy. Conditional Divergence is particularly affected by removing patient context, while Temporal and Sub-Typical Divergence are especially dependent on the medication documentation needed to establish the corresponding discrepancy. Direct Divergence remains substantially more detectable when the persona is removed, consistent with explicit contradictions often being identifiable directly from the advice and source documentation. Overall, the results indicate that the two contextual sources play complementary roles in distinguishing among divergence types.

A.8 Qualitative Analysis

To complement our quantitative evaluation, we conducted a qualitative error analysis to better understand the limitations of current approaches in contextual 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 profile judgment. This reveals a conversational bias in model reasoning that is particularly problematic in the medical domain, where linguistic precision is crucial.

A.9 Illustrative Example from Online Health Discussions

To explore whether the contextual divergence formulation may generalize beyond the semi-synthetic benchmark introduced in this work, we conducted a small exploratory evaluation using 100 samples collected from Reddit health discussions on `r/AskDocs`. Unlike the structured benchmark, these samples contain naturally occurring conversational health advice and informal language.

For each sample, the original Reddit post contained the **persona and contextual profile**, since posts frequently describe the user's symptoms, medical history, medications, demographics, or actions already taken. The **reference advice** was derived from actions or recommendations explicitly mentioned within the post itself (e.g., medications taken, treatment decisions, or intended actions). Responses provided by verified or medically grounded commenters were treated as the **authoritative documents** to evaluate for contextual divergence.

We evaluated this setting using *o3-mini*, applying the same divergence reasoning framework introduced for the structured benchmark. After generating a divergence prediction, we manually reviewed each sample to determine whether the assigned label appropriately captured the relationship between the Reddit post context and the authoritative response.

The resulting label distribution was skewed toward non-divergent cases, with 63 samples labeled as *No Divergence*, 31 as *Direct Divergence*, 4 as *Temporal Divergence*, 1 as *Conditional Divergence*, and 1 as *Sub-Typical Divergence*. Across the 100 evaluated samples, the model achieved an overall accuracy of 89% relative to the manually validated labels.

The experiment aligned with our earlier assumptions regarding the distribution of divergence types in natural online health discussions, where most interactions tend to be non-divergent or directly divergent, while more nuanced categories such as conditional, temporal, and sub-typical divergences appear substantially less frequently.

While this experiment is preliminary and substantially smaller in scale than the primary benchmark, the results suggest that contextual divergence reasoning can extend to conversational health discussions despite the increased ambiguity, informal language, and variability present in real-world online interactions.

Example Post.

```
r/askdocs

Title: Pancreatitis with 30% necrosis on tail of pancreas evolving into neuropathy?

My friend (56F, 170cm, 70kg, european, having chronic pancreatitis diagnosed for 4 years but not
serious until a week ago, pretty good health habit[URL] approx. 2 coffees a day, 1glass of wine a
week, a lot of vegetable, all bio homemade, not too fatty, never smoked, no medication on regular
basis) was diagnosed with chronic pancreatitis after acute attack 3 years ago. At that time, she ended
up with minor necrosis (3-5%). Now she had huge attack, she also fell of while vomitting at night,
after dinner with a glass of wine and fatty pizza. Now she is in hospital for 1 week, CT showed 30%
necrosis after this attack. CRP values are lowering, was 300, now is 25. However, suddenly and very
quickly she stopped walking and strong neuropathy came. She is extremly sensitive to touch, sweating
heavily, does not eat and is in depression. Sher barely reacts to family, cries silently and is heavy
breathing. Due to covid situation, visits are still not allowed, so is there alone. She is on
diazepam, but IDK whether diazepam is to surpass the neuropathy or the neuropathy is caused by
diazepam (I do not know when she received first one).
So my question i[URL] can chronic pancreatitis cause such cognitive decline and strong neuropathy? Why
it could be that diazepam is used?

Family asked doctors, but it is hard to communicate with them and we do our bit of research to
understand situation.
```

Mapping to contextual Divergence Detection.

```
Age: 56
Gender: female
Medical Condition: Chronic pancreatitis with recent acute attack (30% necrosis); new severe
neuropathy and cognitive decline
Current Medication: diazepam
```

```
Reference Advice: Avoid alcohol and high-fat meals; request a comprehensive medication review
(including the indication and potential side effects of diazepam), a neurology consult for new
neuropathic symptoms, prompt screening and empiric treatment for nutritional deficiencies (e.g.,
thiamine before glucose, B12, folate) and electrolyte abnormalities, and a clear nutrition plan
(small low-fat meals, pancreatic enzymes, or enteral feeding if oral intake is poor); ensure pain
control, fall precautions with supervised ambulation/physiotherapy, mood support, and structured
communication by designating a family point person, scheduling regular updates with the team, and
using virtual visits if in-person visitation is restricted.
```

Differences from the Structured Benchmark. This exploratory experiment also highlights several challenges that distinguish real-world conversational health discussions from the controlled benchmark introduced in this work.

First, the **user profile is incomplete and inconsistently specified**. Some health-focused communities such as `r/AskDocs` often contain relatively rich contextual cues, including symptoms, medications, demographics, prior diagnoses, or actions already taken. However, many other online discussions provide substantially less structured information, and important contextual attributes may still be omitted even within medically oriented forums. Unlike our benchmark, where personas are explicitly structured and standardized, conversational data requires implicit extraction and reconstruction of relevant contextual information from noisy and highly variable user narratives.

Second, the **reference evidence is less clearly grounded**. In the benchmark dataset, advice statements are evaluated directly against authoritative medication guideline documents. In Reddit discussions, recommendations may originate from anecdotal experiences, partial recollections, or community-driven knowledge sharing. Determining which statements should serve as authoritative references therefore becomes substantially more ambiguous.

Third, the **advice is embedded within informal dialogue**. Reddit discussions frequently contain hedging, uncertainty, sarcasm, incomplete statements, or multiple overlapping recommendations. This creates additional difficulty in identifying the precise claim that should be evaluated for divergence.

Implications. These findings suggest that contextual divergence detection remains relevant in naturally occurring conversational health settings, but also highlight the additional challenges introduced by noisy and weakly grounded online discussions. The structured benchmark proposed in this work isolates the core reasoning problem under controlled conditions where contextual information and reference evidence are explicitly available.

The exploratory `r/AskDocs` evaluation provides preliminary evidence that the framework may transfer to broader conversational environments, though larger-scale studies with carefully validated annotations would be necessary to draw robust conclusions. We therefore view conversational health platforms as an important future direction for extending contextual divergence detection toward real-world deployment settings.

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 8: 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.