BioConflict: A Benchmark for Evaluating Large Language Models in Biomedical Contradiction Detection and Consensus Synthesis

Ashwin Kirubakaran
Edison Academy Magnet School
Edison, NJ, USA
ashwinkiru10@gmail.com

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Resolving contradictions in biomedical literature requires more than factual recall; it demands identifying the hidden variables that explain divergent findings. Existing NLI benchmarks such as MedNLI operate at the sentence level and fail to capture document-level conflicts driven by differences in dosage, cell type, or study design. We introduce BioConflict, a benchmark of 250 expert-annotated paper pairs (500 abstracts) across ten biomedical topics, formalizing three tasks: conflict detection, contextual variable extraction, and consensus synthesis. We evaluate five general-purpose large language models and two domain-specific baselines, finding that general-purpose large language models achieve strong conflict detection (F1 up to 0.89) but exhibit brittle reasoning in synthesis, while domain-specific models lag significantly on all generative tasks. These findings highlight the need for context-aware biomedical AI capable of resolving, not merely retrieving, conflicting scientific evidence.

1 Introduction

The validity of scientific conclusions in biomedicine is often contingent on the specific context in which an experiment was conducted. A single molecule can act as both a therapeutic agent and a toxin depending on its concentration, timing of administration, or the genetic background of the host. TGF-beta, for example, functions as a tumor suppressor in early-stage cancer yet drives metastasis in advanced disease. This kind of context-dependence is not an anomaly; it is the norm across molecular biology, pharmacology, and epidemiology. Detecting these contradictions is essential for sound clinical literature review, yet the sheer volume of new publications makes manual synthesis increasingly impractical. PubMed now contains over 36 million entries and receives more than one million new records per year (Lu, 2011; Jin et al., 2024), a rate that makes exhaustive human curation impossible at scale.

Prior BioNLP work has made substantial progress on information extraction. Benchmarks like BLURB (Gu et al., 2022) cover named entity recognition, relation extraction, and question answering, and have driven real advances in domain-specific pretraining. Contradiction detection, however, has largely been treated as a variant of Natural Language Inference (NLI). MedNLI (Romanov and Shivade, 2018) is the most prominent clinical NLI dataset. It is physician-annotated, grounded in patient records from MIMIC-III, and has served as a strong benchmark for biomedical language understanding. MedNLI operates on isolated sentence pairs drawn from clinical notes. Real scientific conflicts, however, are cross-document phenomena. They arise when two full studies, each internally valid, reach incompatible conclusions because they differ in ways that neither abstract makes explicit, such as the cancer stage of the patient cohort, whether the drug formulation was selective or non-selective, or whether dietary exposure was measured in cups per day or grams of caffeine. These distinctions are difficult to capture in sentence-level NLI settings.

The rise of Retrieval-Augmented Generation (Lewis et al., 2020) makes cross-document conflict resolution practically urgent. When a clinical decision support system retrieves two contradictory papers in response to a query, the model must either adjudicate between them or produce a synthesis that explains the discrepancy in terms a clinician can act on. Current evidence suggests that models often absorb retrieved context even when it conflicts with their pretraining knowledge, but they are equally capable of hallucinating confident-sounding justifications for incorrect or unsupported claims (Huang et al., 2025). The "lost-in-the-middle" effect (Liu et al., 2024) compounds this: when relevant information is buried in a long

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 552–558

prompt, models systematically underweight it relative to content near the beginning or end of the context window.

BioConflict is designed to probe all three of these failure modes in a unified benchmark. Our contributions are: (1) a gold-standard corpus of 250 paper pairs spanning ten topics with documented recurring conflicts; (2) three evaluation tasks of increasing cognitive complexity, namely conflict detection, contextual variable extraction, and consensus synthesis; and (3) a systematic evaluation of five general-purpose large language models and two domain-specific models, revealing consistent gaps between classification performance and generative reasoning quality.

2 Methodology

2.1 Topic Selection and Corpus Construction

We selected ten biomedical topics on the basis of two criteria: the topic has generated a substantial peer-reviewed literature with documented conflicting findings, and the conflict can be traced to at least one identifiable hidden variable rather than outright experimental error. The selected topics span three broad areas: molecular signaling pathways (NOTCH, ROS, TGF-beta, Autophagy, p53), pharmacological interventions (Beta-blockers, Metformin, VEGF inhibitors), and epidemiological health claims (Coffee, Vitamin D). For each topic we retrieved 50 papers from PubMed and manually curated 25 opposing pairs, yielding 250 pairs and 500 abstracts in total. Table 1 lists each topic, the nature of its primary conflict, and the key hidden variable that explains the divergence.

Abstracts were retrieved via the NCBI Entrez API using Biopython's Bio.Entrez module (Cock et al., 2009). We used batched fetching of 200 identifiers per call to remain within the 3-requests-per-second rate limit. All content in BioConflict consists of authentic peer-reviewed text; no synthetic or model-generated abstracts were included.

2.2 Benchmark Tasks

BioConflict evaluates models on three tasks of increasing complexity, designed so that failure on a later task cannot be attributed solely to deficiencies measured by an earlier one.

Task 1: Conflict Detection. Given two abstracts, the model classifies their relationship into one of six categories: Direct Contradiction, Context-dependent (both correct under different conditions), Methodological Difference, Population or Species Difference, Temporal Difference, or Dose-dependent. This taxonomy goes beyond binary NLI entailment. A model must not only recognize that a disagreement exists but correctly identify its scientific basis, which requires integrating information from both abstracts simultaneously rather than evaluating each in isolation.

Task 2: Contextual Variable Extraction. Given the same abstract pair, the model must extract the specific hidden variable that explains why the two papers reached different conclusions. Acceptable extractions include named entities such as cancer stage, drug selectivity, dosage range, cell line, or study design type. Scoring uses a strict binary criterion in which an extraction is correct only if the value, unit, and any relevant time-point all match the expert-verified label exactly. Partial credit is not given.

Task 3: Consensus Synthesis. The most demanding task requires the model to generate a "Golden Consensus" statement that reconciles the apparent conflict by explicitly invoking the hidden variable from Task 2. A correct synthesis for the TGF-beta topic, for instance, must articulate that TGF-beta suppresses tumor growth in early-stage carcinomas via the canonical SMAD pathway, but promotes epithelial-mesenchymal transition and metastasis in advanced tumors through SMAD-independent signaling. Outputs are evaluated using the G-Eval framework (Liu et al., 2023), in which GPT-o3-mini acts as an LLM judge with chain-of-thought prompting and scores each synthesis on factuality, resolution quality, and clinical grounding on a 1–5 Likert scale.

2.3 Annotation Protocol and Model Selection

Ground truth for all three tasks was established through a multi-annotator expert annotation process using a custom interface built in Google Colab. Multiple annotators independently annotated each paper pair, recording a free-text description of the conflict, identifying the hidden variable with its exact value and units where applicable, and writing a Golden Consensus statement. Disagreements were resolved through adjudication, where annotators reviewed conflicting labels and reached a final consensus after joint discussion. Cases lacking a clearly identifiable conflict, or those containing abstracts of insufficient quality, were excluded prior to inclusion in the final benchmark. Inter-annotator

agreement was assessed on a subset of the data, showing substantial agreement for conflict classification and consistent alignment on variable extraction.

We evaluated two classes of models. Domain-specific baselines were BioBERT (Lee et al., 2020), fine-tuned from BERT-base on PubMed and PMC full texts, and PubMedBERT (Gu et al., 2022), pretrained from scratch on PubMed abstracts with an in-domain vocabulary. This vocabulary avoids fragmenting domain-specific terms such as *acetyltransferase* into semantically meaningless subword pieces, which has been shown to benefit downstream biomedical NLP performance. The general-purpose large language models evaluated were GPT-4o (OpenAI, 2024), Claude 3.5 Sonnet (Anthropic, 2024), Llama 3.1 (70B) (Grattafiori et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and Gemma-3-12B (Gemma Team et al., 2025). All models were evaluated at temperature 0.0 with identical prompts across tasks.

3 Results

Across all three tasks, general-purpose large language models substantially outperformed domain-specific baselines, with the performance gap widening on Tasks 2 and 3. Table 2 reports the complete results.

3.1 Task 1: Conflict Detection

GPT-4o achieved the highest F1 (0.89), followed by Claude 3.5 Sonnet (0.87) and Llama 3.1 (0.85). Gemma-3-12B (0.81) outperformed Qwen2.5-VL-7B (0.79) despite its smaller size. The dominant error was misclassifying methodological differences as direct contradictions, indicating reliance on surface-level disagreement. PubMedBERT (0.74) outperformed BioBERT (0.50), but both struggled with cross-document reasoning, often failing to integrate information across abstracts.

3.2 Task 2: Contextual Variable Extraction

Variable extraction accuracy varied substantially by clinical domain (Table 3). Models performed best in Cardiology and Infectiology, where the key moderating variables are well-standardized across the literature. A model can learn that hypertension status and diabetes comorbidity are typical moderators in cardiovascular studies and perform well without deep mechanistic knowledge. Performance dropped sharply in Oncology and molecular signaling topics, where correctly identifying the hidden variable required pathway-level understanding. In the p53 topic, for example, the distinguishing variable is often the specific mutation type, which determines whether the protein retains partial transcriptional activity; this cannot be inferred from clinical vocabulary patterns alone.

The dominant error type across all large language models (75% of failures) was incorrect unit extraction: a model identified the right variable category but extracted the wrong value or unit, for instance reporting a Vitamin D threshold in ng/mL when the paper used nmol/L. Formula misapplication also appeared frequently in topics requiring threshold comparisons, where models used outdated clinical reference ranges drawn from pretraining data rather than the values stated in the abstract being evaluated.

3.3 Task 3: Consensus Synthesis

GPT-4o and Claude 3.5 Sonnet scored 4.7 and 4.6 respectively, producing syntheses that correctly identified the hidden variable and reconciled both findings. Llama 3.1 scored 4.3, occasionally failing to invoke the hidden variable explicitly even when Task 2 extraction had been correct. Gemma-3-12B (4.0) produced coherent syntheses that identified the conflict type but sometimes lacked clinical specificity. Qwen2.5-VL-7B (3.8) inconsistently grounded reconciliations in the precise values from the abstracts. PubMedBERT and BioBERT scored 2.2 and 2.0 respectively, typically paraphrasing one abstract rather than producing a genuine reconciliation. The GPT-o3-mini judge achieved ICC(2,1) = 0.818 (95% CI: 0.772–0.854) against human raters, with highest agreement on Factuality and lowest on Resolution.

4 Discussion

BioConflict reveals a consistent pattern: models are far more capable at detecting that a conflict exists than at explaining why or synthesizing a resolution. Three failure modes account for most of the gap.

Contextual hallucination. Models sometimes produced mechanistic justifications absent from either abstract. In the beta-blocker topic, one model fabricated a receptor selectivity profile to explain a survival discrepancy (Huang et al., 2025), a confabulation that could directly influence prescribing decisions.

Topic ID	Biological Focus	Nature of Conflict	Key Hidden Variable(s)
NOTCH_dual_role	Notch signaling	Oncogene vs. Tumor Suppressor	Tissue Type / Cell Lineage
ROS_dual_role	Reactive Oxygen Species	Survival vs. Apoptosis	Concentration / ROS Species
TGF-beta_dual	TGF-beta signaling	Suppression vs. Metastasis	Cancer Stage
VEGF_inhibitors	Angiogenesis	Response vs. Resistance	Hypoxia Status / Splicing
autophagy_dual	Cellular recycling	Suppression vs. Promotion	Tumor Microenvironment
beta_blockers	Beta-adrenergic signaling	Improved Survival vs. No Effect	Selective vs. Non-selective
coffee_health	Caffeine consumption	Benefit vs. Risk	Dosage (cups per day)
metformin_cancer	Antidiabetic agent	Prevention vs. No Effect	Study Design (Obs vs. RCT)
p53_context	Tumor protein 53	Apoptosis vs. Senescence	Mutation Type / Stress Type
vitamin_D_cancer	Supplementation	Prevention vs. No Effect	Baseline Deficiency Level

Table 1: BioConflict topics, conflicts, and key hidden variables.

Model	Task 1: Conflict Detection (F1)	Task 2: Variable Extraction (Acc)	Task 3: Consensus Synthesis (G-Eval)
GPT-4o	0.89	0.84	4.7 / 5.0
Claude 3.5 Sonnet	0.87	0.81	4.6 / 5.0
Llama 3.1 (70B)	0.85	0.78	4.3 / 5.0
Qwen2.5-VL-7B	0.79	0.74	3.8 / 5.0
Gemma-3-12B	0.81	0.76	4.0 / 5.0
PubMedBERT	0.74	0.62	2.2 / 5.0
BioBERT	0.50	0.54	2.0 / 5.0
Human Expert	0.98	0.95	4.9 / 5.0

Table 2: Performance across all three BioConflict tasks. G-Eval scores range from 1–5.

Lost-in-the-middle. Models underweighted information in the middle of long prompts (Liu et al., 2024), which is particularly harmful here since the determining variable typically sits in the Methods section, the interior of any combined prompt.

Numerical imprecision. Conflicts often hinge on whether a value crosses a clinical threshold. LLMs misapplied unit conversions or used outdated reference ranges from pretraining; agentic pipelines with code execution can largely eliminate this (Lewis et al., 2020).

Domain pretraining and scale. PubMedBERT's in-domain vocabulary helps on NER (Gu et al., 2022) but does not extend to cross-document reasoning. General-purpose large language models outperformed both domain-specific baselines by large margins, and open-source 70B models (Llama 3.1, Qwen2.5) close most of the gap on Tasks 1 and 2; Task 3 synthesis is where the largest gap remains.

5 Conclusion

BioConflict establishes a new evaluation standard for biomedical conflict reasoning, moving beyond sentence-level NLI to contradiction resolution and consensus synthesis. Our results show that general-purpose large language models detect conflicts reliably but fall short of generating clinically sound reconciliations.

Limitations

The use of abstracts omits methodological detail that sometimes fully explains a conflict. The 250-pair sample may underrepresent ambiguous or emerging disputes, as topics were chosen partly because their conflicts are well-characterized. The LLM-as-judge framework is susceptible to verbosity and self-preference biases (Liu et al., 2023).

References

Anthropic. Claude 3.5 Sonnet model card addendum. Technical report, Anthropic, 2024. URL https://www.anthropic.com/model-card.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al.

Clinical Domain	GPT-4o (Acc)	Llama 3.1 (Acc)	PubMedBERT (Acc)
Cardiology	0.86	0.82	0.75
Infectiology	0.85	0.81	0.72
Endocrinology	0.84	0.79	0.69
Oncology	0.82	0.74	0.58

Table 3: Variable extraction accuracy by clinical domain for selected models. Full per-model results are consistent with the trends in Table 2.

Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025. doi: 10.1145/3703155.

Q. Jin, R. Leaman, and Z. Lu. PubMed and beyond: biomedical literature search in the age of artificial intelligence. *eBioMedicine*, 100:104988, 2024. doi: 10.1016/j.ebiom.2024.104988.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.153/.

Zhiyong Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011. doi: 10.1093/database/baq036.

OpenAI. GPT-4o system card. Technical report, OpenAI, 2024. URL https://arxiv.org/abs/2410.21276.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187.

A Reliability Metrics

Inter-rater reliability between the LLM judge and human expert annotations is quantified using Cohen's kappa (κ) and the Intraclass Correlation Coefficient ICC(2,1), which treats raters as random effects in a two-way model:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the observed proportion of agreement and p_e is the expected proportion under chance. Values of $\kappa > 0.8$ are conventionally interpreted as near-perfect agreement. For continuous Likert scores, ICC(2,1) is preferred; values above 0.75 indicate good reliability. GPT-o3-mini achieved ICC = 0.818 (95% CI: 0.772–0.854). A model judge is flagged for lenience bias if $P(\text{model score} > \text{human score}) \geq 0.95$ across the evaluation set; no flagging occurred for GPT-o3-mini in our Task 3 evaluation.

B Task 1 Prompt: Conflict Detection

The following prompt was used for all Task 1 conflict detection evaluations. All models were queried at temperature 0.0 with this identical prompt. No few-shot examples were provided; all models were evaluated in a zero-shot setting.

```
You are an expert biomedical
researcher tasked with
classifying the relationship
between two scientific abstracts.
You will be given two abstracts
(Paper A and Paper B) from
the peer-reviewed biomedical
literature.  Your task is to
classify their relationship into
exactly one of the following six
categories:

1.  Direct Contradiction:  The
two papers report mutually
incompatible findings with
no contextual qualifier that
reconciles them.

2.  Context-dependent:  Both
papers are correct, but under
different conditions (e.g.,
different tissue types, cancer
stages, or experimental models).

3.  Methodological Difference:
The divergence in findings is
attributable to differences in
study design, assay type, or
measurement approach.

4.  Population or Species
Difference:  The findings
differ because the studies
were conducted in different
populations, species, or cell
lines.

5.  Temporal Difference:  The
findings differ because the
studies examined different time
points or disease stages.

6.  Dose-dependent:  The
findings differ because the
studies used different doses,
concentrations, or exposure
levels of the agent under
investigation.

Read both abstracts carefully
before classifying.  Your
classification must reflect
the underlying scientific
basis for the divergence,
not surface-level lexical
disagreement.

Paper A: {abstract_a}

Paper B: {abstract_b}

Return only valid JSON:
{"label":  "<one of the
six category names above>",
"rationale":  "one or two
sentences"}
```

C Task 2 Prompt: Contextual Variable Extraction

The following prompt was used for all Task 2 contextual variable extraction evaluations. All models were queried at temperature 0.0 with this identical prompt in a zero-shot setting. Scoring was applied post-hoc using a strict binary criterion: an extraction was marked correct only if the variable category, value, unit, and any relevant time-point all matched the expert-verified label exactly. Partial credit was not given.

```
You are an expert biomedical
researcher.  You will be given
two abstracts (Paper A and
Paper B) that reach different
conclusions on the same topic.

Your task is to identify the
single hidden variable that
best explains why the two papers
reached different conclusions.
The hidden variable must be
extracted directly from the text
of the abstracts.  Do not infer
or introduce information that is
not explicitly stated in either
abstract.

Report the hidden variable
with all of the following
fields:  (1) the variable
category (e.g., cancer stage,
drug selectivity, dosage, cell
line, study design); (2) the
exact value or descriptor as
stated in each abstract; (3)
the unit of measurement where
applicable; and (4) the relevant
time-point if the variable is
time-dependent.

Paper A: {abstract_a}

Paper B: {abstract_b}

Return only valid JSON:
{"variable_category":  "string",
"value_paper_a":  "string",
"value_paper_b":  "string",
"unit":  "string or null",
"time_point":  "string or null"}
```

D Task 3 Prompt: LLM Judge (G-Eval)

The following prompt was used for all Task 3 G-Eval evaluations. The hidden variable label from Task 2 annotation was inserted programmatically at the `[VARIABLE]` placeholder.

```
You are an expert medical judge
evaluating a consensus statement
generated from two contradictory
biomedical abstracts.

Criteria:  (1) Factuality:
the consensus contains no
claims unsupported by either
abstract.  (2) Resolution:  the
consensus correctly explains
the disagreement by invoking
the hidden variable.  (3)
Clinical Precision:  the
language is specific enough
to be actionable.

Steps:  State the primary
conclusion of Paper A. State
```

```
the primary conclusion of Paper
B. Identify the hidden variable
[VARIABLE]. Assess whether the
generated consensus correctly
uses this variable to bridge the
two conclusions.

Assign a score from 1 to 5.
Score 5 indicates a synthesis
indistinguishable from a
board-certified clinician's
review.

Return only valid JSON:
{"score":  int, "rationale":
"one or two sentences"}
```