Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MedConclusion: A Benchmark for Biomedical Conclusion Generation from Structured Abstracts

Mengyu Wang

mengyu_wang@meei.harvard.edu

Harvard Ophthalmology AI lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School

Weiyue Li

Harvard University

Ruizhi Qian

University of Southern California

Yi Li

Carnegie Mellon University

Yongce Li

Stanford University

Yunfan Long

Carnegie Mellon University

Jiahui Cai

Harvard University

Yan Luo

Harvard University https://orcid.org/0000-0001-5135-0316

Article

Keywords: biomedical NLP, scientific reasoning, conclusion generation, structured abstracts, large language models

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10531914/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

MedConclusion: A Benchmark for Biomedical Conclusion Generation from Structured Abstracts

Weiyue Li[1†], Ruizhi Qian[1,2†], Yi Li[1,3†], Yongce Li[4], Yunfan Long[3], Jiahui Cai[1], Yan Luo[1*], Mengyu Wang[1,5*]

[1]Harvard AI and Robotics Lab, Harvard Medical School, Boston, MA, USA.
[2]University of Southern California, Los Angeles, CA, USA.
[3]Carnegie Mellon University, Pittsburgh, PA, USA.
[4]Stanford University, Stanford, CA, USA.
[5]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): yluo16@meei.harvard.edu; mengyu_wang@meei.harvard.edu;
Contributing authors: weiyueli@fas.harvard.edu; ruizhiqi@usc.edu; yili6@andrew.cmu.edu; yongceli@stanford.edu; justinlo@andrew.cmu.edu; jiahuicai@fas.harvard.edu;
†These authors contributed equally to this work.

Abstract

Large language models (LLMs) are widely explored for reasoning-intensive research tasks, yet resources for testing whether they can infer scientific conclusions from structured biomedical evidence remain limited. We introduce **MedConclusion**, a large-scale dataset of **5.7M** PubMed structured abstracts for biomedical conclusion generation. Each instance pairs the non-conclusion sections of an abstract with the original author-written conclusion, providing naturally occurring supervision for evidence-to-conclusion reasoning. MedConclusion also includes journal-level metadata such as biomedical category and SJR, enabling subgroup analysis across biomedical domains. As an initial study, we evaluate diverse LLMs under conclusion and summary prompting settings and score outputs with both reference-based metrics and LLM-as-a-judge. We find that conclusion writing is behaviorally distinct from summary writing, strong models remain closely clustered under current automatic metrics, and judge identity can substantially shift absolute scores. MedConclusion provides a reusable data resource for studying

scientific evidence-to-conclusion reasoning. Our code and data are available at: https://github.com/Harvard-AI-and-Robotics-Lab/MedConclusion.

Keywords: biomedical NLP, scientific reasoning, conclusion generation, structured abstracts, large language models

1 Introduction

Large language models (LLMs) have shown strong reasoning capability across a wide range of demanding settings, including mathematical thinking, long-form creative writing, and scientific discovery assistance [1–6]. As these capabilities improve, there is growing interest in using LLMs to support research workflows, not only to retrieve or summarize papers, but also to infer scientific conclusions from evidence. Structured abstracts provide a particularly convenient setting for studying this capability as we can frame this evidence-to-conclusion reasoning problem as: given **Background**, **Methods**, and **Results**, the model should infer a **Conclusion** without injecting unprovided context. However, finding large-scale data sources from diverse scientific domains is challenging, which motivates us to shift the focus to biomedicine, where structured abstracts are widely used.

Existing work has explored this direction, but current resources remain limited in two ways. First, existing *datasets* are often narrow in scope, focusing on specific study types or specialized report formats, such as randomized controlled trial abstracts or echocardiography notes, rather than broad biomedical literature [7, 8]. Other work uses conclusion reconstruction mainly as a proxy for premise–conclusion alignment or as a training objective, rather than as a reusable data resource [9, 10]. These resources also typically do not emphasize journal-level metadata such as biomedical category and SJR, which limits analysis of how difficulty varies across subfields or venue strata. Second, existing *benchmarking designs* do not fully isolate the reasoning problem of conclusion generation. Some adjacent biomedical resources focus on question answering, medical exam reasoning, treatment-effect inference, or claim verification rather than deriving the author-written conclusion itself [11–18]. Moreover, open-ended conclusion generation is difficult to evaluate reliably because reference-based metrics are incomplete, and LLM judges can vary substantially in calibration [19–23].

To address these limitations, we present **MedConclusion**, a large-scale **5.7M** dataset for biomedical conclusion generation from PubMed structured abstracts. Each instance pairs the non-conclusion sections of an abstract with its original author-written conclusion, yielding naturally occurring supervision for evidence-to-conclusion reasoning. In addition, MedConclusion includes journal-level metadata, including *biomedical category* labels and *SJR* records. This combination of large-scale author-written supervision, broad biomedical coverage, and journal metadata enables analyses that are difficult to conduct in prior conclusion-generation settings. We further provide an initial empirical study using diverse LLMs, contrasting conclusion prompting against summary prompting, evaluating with a hybrid rule-based reference metrics and LLM judges, and examining robustness across judge backbones.

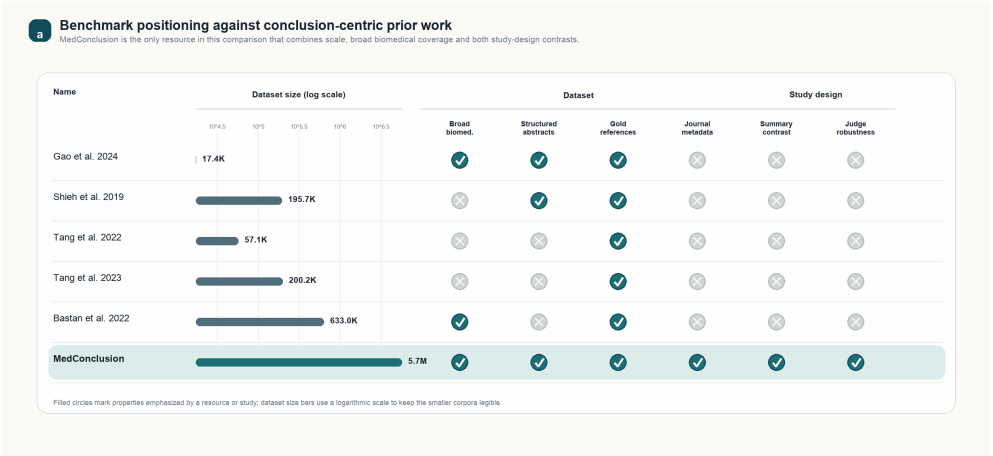


Fig. 1: Comparison of MedConclusion with conclusion-centric prior work. Dataset size is shown on a logarithmic scale, and filled circles denote properties that are central and explicitly emphasized by each resource or study. For multi-part resources, total dataset size sums all released components used in the paper.

In summary, our work makes three contributions. First, we curate **MedConclusion**, a **5.7M**-example dataset of PubMed structured abstracts for biomedical conclusion generation. Second, we augment the dataset with journal-level metadata, enabling aggregate and subgroup analysis across biomedical domains and venue strata. Third, we provide a first empirical study of the dataset by evaluating diverse LLMs, contrasting conclusion versus summary prompting, and studying the sensitivity of automatic evaluation to judge identity. We hope MedConclusion serves as a reusable data resource for future study of scientific evidence-to-conclusion reasoning.

2 Related work

Adjacent biomedical reasoning resources.

A large body of biomedical NLP work studies reasoning over scientific and clinical text, but not specifically the task of inferring an abstract conclusion from preceding evidence. PubMedQA and BioASQ evaluate biomedical question answering [11, 12]; MedQA and MedMCQA focus on exam-style medical reasoning [13, 14]; EBM-NLP and Evidence Inference study treatment-effect reasoning from structured evidence [15–17]; and SciFact evaluates scientific claim verification [18]. More recently, EvidenceBench studies sentence-level evidence extraction for biomedical hypotheses from full papers rather than conclusion generation from structured abstracts [24]. Structured abstracts and discourse-aware scientific summarization have also been widely studied [25–29]. MedConclusion differs from these resources by centering benchmarking the evidence-to-conclusion reasoning step itself.

Conclusion-centric generation and reconstruction.

The closest prior work to MedConclusion studies conclusion reconstruction directly. Gao et al. [9] reconstruct conclusions of structured scientific abstracts to evaluate premise–conclusion alignment, treating conclusion generation mainly as an evaluation proxy. Shieh et al. [7] study conclusion generation for randomized controlled trial abstracts, Tang et al. [8] focus on echocardiography notes, Tang et al. [30] study factual consistency in conclusion-oriented clinical-study summarization, and Bastan et al. [10] use large-scale PubMed conclusion generation as a training objective. In contrast, MedConclusion contributes a broader **5.7M**-example PubMed resource with author-written targets and journal-level metadata, and uses it to study conclusion generation as a reasoning task rather than only as a proxy objective. Figure 1 shows a detailed comparison.

Evaluation of open-ended scientific reasoning.

Evaluating conclusion generation is challenging because open-ended outputs can differ in wording, scope, and detail while remaining partially valid. Prior work in summarization has shown that lexical overlap and fluency metrics are incomplete proxies for factual correctness [19, 31–38]. LLM-as-a-judge provides a scalable alternative [20, 21], but recent work documents sensitivity to judge identity and grading scale, verbosity bias, position bias, and broader generalization concerns [23, 39–43]. These findings motivate the evaluation protocol used in this paper, but they are secondary to our main goal of introducing MedConclusion as a large-scale dataset for biomedical conclusion generation.

3 Methodology

3.1 Data curation

MedConclusion is constructed from PubMed articles with structured abstracts published between 2000 and 2025. We identify candidate papers using PubMed's constraint `hasstructuredabstract` and collect the corresponding records for downstream processing.

3.1.1 Data collection pipeline

Our data collection pipeline is implemented with Entrez Direct (EDirect) [1] and a custom XML parser. We first query PubMed for all UIDs satisfying `hasstructuredabstract` within the target time span, and deduplicate the retrieved identifiers. We then batch-download the corresponding PubMed XML records via the `epost` and `efetch` commands provided by EDirect, and parse each record into a JSONL representation containing article metadata, keywords, and structured abstract segments represented as (`label`, `nlm_category`, `text`) tuples. Records with missing abstract labels are filtered out during parsing.

[1] Kans J. Entrez® Direct: E-utilities on the Unix Command Line. 2013 Apr 23 [Updated 2025 Mar 25]. In: Entrez® Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK179288/

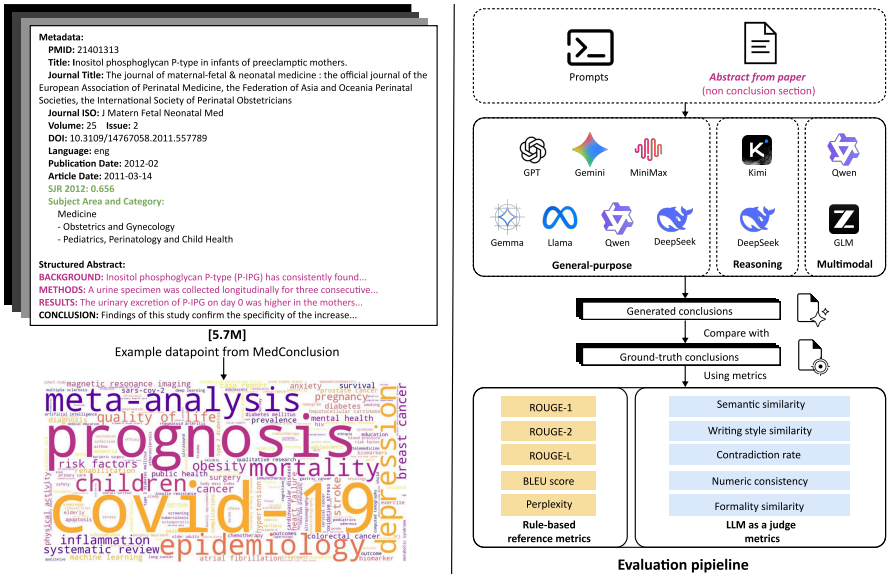


Fig. 2: Overview of MedConclusion and the evaluation pipeline. Left: an example MedConclusion instance, including article metadata, subject categories, and a structured abstract, where the non-conclusion sections are used as model input and the author-written conclusion serves as the gold reference. Right: the non-conclusion abstract is paired with prompts and given to diverse LLM families to generate conclusions, which are then compared against the ground-truth conclusion using both rule-based reference metrics and multi-dimensional LLM-as-a-judge metrics.

After parsing, we perform record-level deduplication using PMID, DOI, and normalized title. We then apply a rule-based cleaning procedure that keeps only English-language records with non-empty core bibliographic fields, normalizes date fields and missing metadata, and retains only articles with at least three abstract segments and at least one conclusion section. Conclusion sections are identified by matching normalized labels against a curated set of conclusion variants (Appendix G). We further remove records with malformed labels and clean keywords by dropping empty, overlong, or non-ASCII entries. The resulting cleaned corpus contains 5,692,839 structured abstract records with at least one conclusion section.

3.1.2 Construction of MedConclusion and dataset statistics

The structured abstract records in MedConclusion span a total of 3,772 unique journals. For each journal, we retrieved its subject category assignments and annual SJR scores from the SCImago Journal & Country Rank (SJR) database,[2] a publicly available bibliometric resource derived from Scopus. Across the full corpus, the 3,772 journals are distributed across 141 subject categories. For SJR scores, we collected annual values

[2] https://www.scimagojr.com

from each journal's first indexed year in the SJR database through 2024. Dataset statistics and a formal definition of the SJR score are provided in Appendix F.

3.2 Task, evaluation, and experimental setup

Given a structured abstract with its **Conclusion** removed, let x be the concatenation of all remaining sections and let $y^\star$ be the original conclusion. A model generates $\hat{y} = f_\theta(x)$.

We evaluate four prompting modes. Ⓐ (default) asks the model to write a formal academic *conclusion*, with no explicit length or style constraints. Ⓑ asks the model to write a formal academic *summary*, with no explicit length or style constraints. Ⓒ asks the model to write a formal academic *conclusion*, with explicit sentence- and word-count targets and instructions to match the abstract's writing style. Ⓓ asks the model to write a formal academic *summary*, with the same sentence- and word-count targets and the same instruction to match the abstract's writing style. Exact prompts are given in Appendix B.

We score outputs with two classes of automatic metrics. First, we use multi-dimensional LLM-as-a-judge scoring. Given $(y^\star, \hat{y})$, the judge outputs five scores in $[0, 100]$: **semantic similarity**, **writing style similarity**, **non-contradiction**, **numeric consistency**, and **formality similarity**. Second, we report lightweight diagnostics and reference-based metrics: word-count ratio, sentence-count ratio, embedding cosine similarity [33], ROUGE-1/2/L [31], BLEU [32], and perplexity on the original and generated conclusions under a fixed external language model (`GPT-2`). Details for reference-based metrics could be found in Appendix D.

We evaluate a diverse set of LLMs spanning closed-source frontier models, open-source instruction-tuned models, multimodal models, reasoning-oriented models, and small models. For each model and prompting mode, we generate one output per instance using the corresponding prompt template. All modes enforce a *no-new-claims* instruction, and we record length-ratio diagnostics to quantify format compliance. Due to cost constraints, we evaluate a randomly sampled 30K subset. Detailed model configurations are in Appendix H.

To study evaluation sensitivity, we use `GPT-5.4-mini` [44] as the primary judge and `Gemini 3 Flash` [45] as a secondary judge (prompts in Appendix C). Our two main comparisons are **conclusion prompting versus summary prompting** and **judge-backbone robustness**. The first tests whether models treat conclusion writing as a distinct discourse function [25, 26]; the second measures how much absolute scores and model rankings depend on judge identity [20, 22, 23, 39, 41, 42].

Figure 2 shows our overall pipeline, and Appendix A shows an example data point.

4 Results

4.1 Overall performance under conclusion generation

The left panel of Figure 3 shows that `GPT-5.4` is the strongest model under the primary judge, leading all five judge dimensions. At the same time, models such as `Gemini 3.1 Pro`, `Gemini 3 Flash`, `DeepSeek-V3.2`, `Gemma-3-27B`, and `GLM-4.6V` lie within

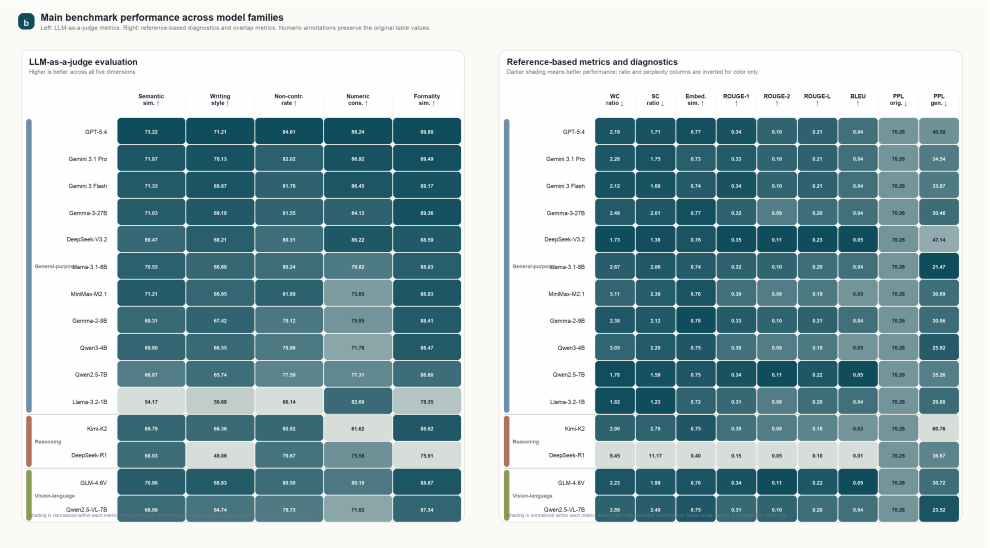


Fig. 3: Nature-style summary of the main benchmark results. Left: LLM-as-Judge evaluation scores for conclusion generation, with models grouped by primary capability. Right: rule-based evaluation scores and diagnostic measures for conclusion generation in the same order. Cell shading is normalized within each metric; for down-arrow metrics, darker cells indicate lower and therefore better values.

only a few points of the top model on most judge dimensions. This score compression suggests that current reference-comparison evaluation separates strong models only weakly, even though the task is not trivial.

The right panel of Figure 3 tells a partly different story. `DeepSeek-V3.2` attains the best ROUGE-1/2/L and ties for the best BLEU, while `GPT-5.4` remains the strongest model under judge-based semantic and non-contradiction scores. Likewise, `Gemma-2-9B` has the highest embedding similarity despite clearly lower judge scores than the best closed models. Perplexity-based fluency diagnostics also decouple from task quality: several mid-sized open or multimodal models have substantially lower generated-text perplexity than `GPT-5.4`, yet they do not approach its semantic or contradiction scores. These mismatches indicate that lexical overlap, embedding similarity, fluency, and judge agreement capture different aspects of biomedical conclusion generation, and indicate our hybrid evaluation approach provides a more comprehensive evaluation than traditional reference-based metrics.

4.2 Conclusion generation is not summary writing

The left two panels of Figure 4 show a clear discourse-function effect. Relative to **A**, **C** slightly reduces semantic and style similarity for both `GPT-5.4` and `Gemini 3 Flash`, but it improves numeric consistency to about 91 for both models. This suggests that explicit style and length control helps models better match the reference conclusion's level of numeric selectivity, even when it slightly constrains broader content choice.

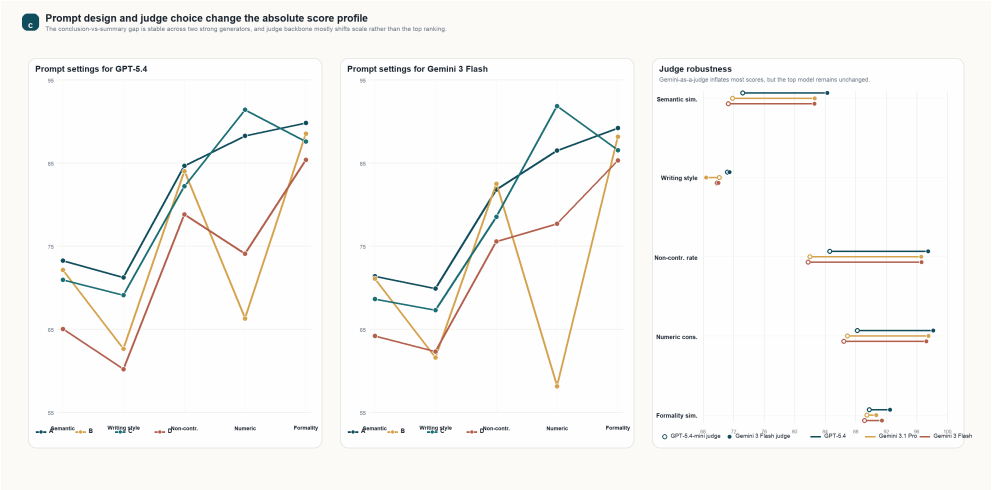


Fig. 4: LLM-as-Judge evaluation scores across prompt settings, generation modes and judge backbones. Left and center: `GPT-5.4` and `Gemini 3 Flash` across prompt settings Ⓐ–Ⓓ, where Ⓐ and Ⓒ generate conclusions and Ⓑ and Ⓓ generate summaries. Right: score shifts when the judge backbone changes from `GPT-5.4-mini` to `Gemini 3 Flash`. Higher is better for all metrics.

When the target is changed from a conclusion to a summary, performance shifts more substantially. Under Ⓓ, semantic similarity drops by about 7–8 points for both models relative to Ⓐ, with even larger drops in writing style similarity and numeric consistency. Ⓑ reveals an even more interesting pattern: semantic similarity rebounds to nearly the Ⓐ level (within 1.11 points for `GPT-5.4` and 0.29 points for `Gemini 3 Flash`), but writing style similarity remains more than 8 points lower and numeric consistency collapses by 22.00 and 28.36 points, respectively. This suggests that unconstrained summaries often preserve the broad meaning of the abstract while selecting different details, especially numbers, scope qualifiers, and level of detail, than the published conclusion.

We therefore interpret these results as evidence that conclusion writing is behaviorally distinct from summary writing in this benchmark setting. Because the judge only compares outputs to the gold conclusion, the effect should be read as a difference in *reference agreement and discourse targeting*, not as proof that summary-mode outputs are unsupported by the input. Some of the additional detail produced in summary modes may still be compatible with the source abstract even when it lowers the agreement with the reference conclusion. We further examine whether this distinction holds across biomedical subfields in Appendix E.11.

4.3 Judge robustness: score scale shifts across judges

The right panel of Figure 4 shows large absolute calibration shifts when the judge backbone is changed. Switching from `GPT-5.4-mini` to `Gemini 3 Flash` as judge raises semantic similarity, non-contradiction, and numeric consistency across the same three

generation models. By contrast, writing style similarity changes only modestly. Thus, the absolute score scale is highly judge-dependent. However, ranking is relatively stable. `GPT-5.4` remains the top generator under both judges across all five dimensions, yet the middle ordering can flip occasionally.

5 Analysis

We further analyze benchmark behavior across journal-level subgroups to understand where conclusion generation is relatively easier or harder. In particular, we study variation along two axes: **journal prestige**, measured by SJR score, and **biomedical category**. These analyses are descriptive rather than part of the benchmark definition itself, and are intended to reveal whether performance differences are associated with venue prestige or with the broader structure of biomedical subfields. All results in this section use `GPT-5.4` under setting A as the representative condition unless otherwise noted.

5.1 SJR score

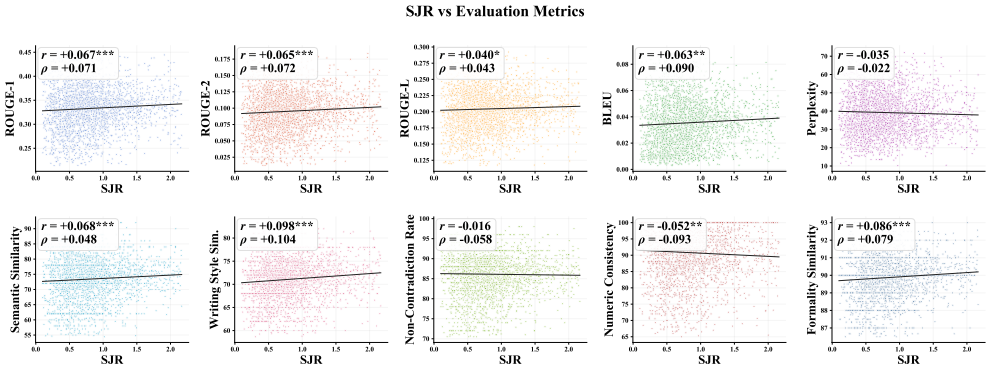
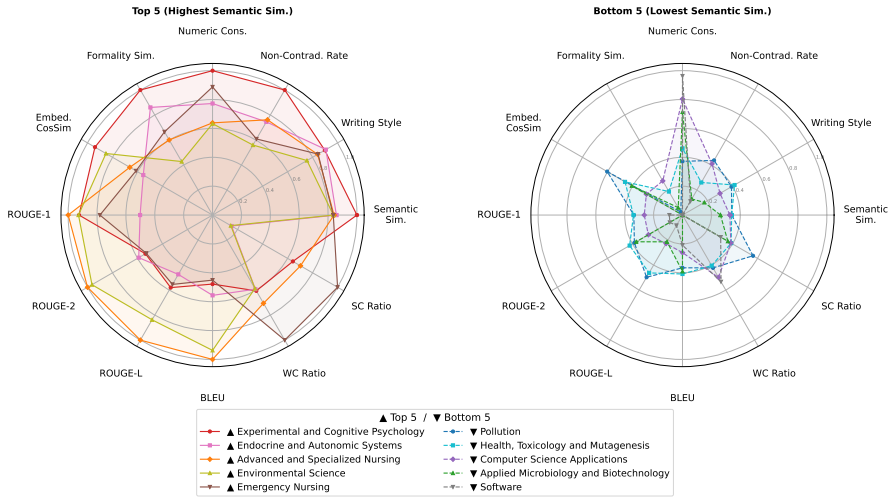


Fig. 5: Scatter plots of journal-level SJR scores versus evaluation metrics for `GPT-5.4` under the A setting (outliers removed). Each point represents one journal, aggregated by mean score. The top row shows reference-based metrics and the bottom row shows LLM-judge dimensions. Pearson (r) and Spearman (ρ) correlations are annotated in each panel ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

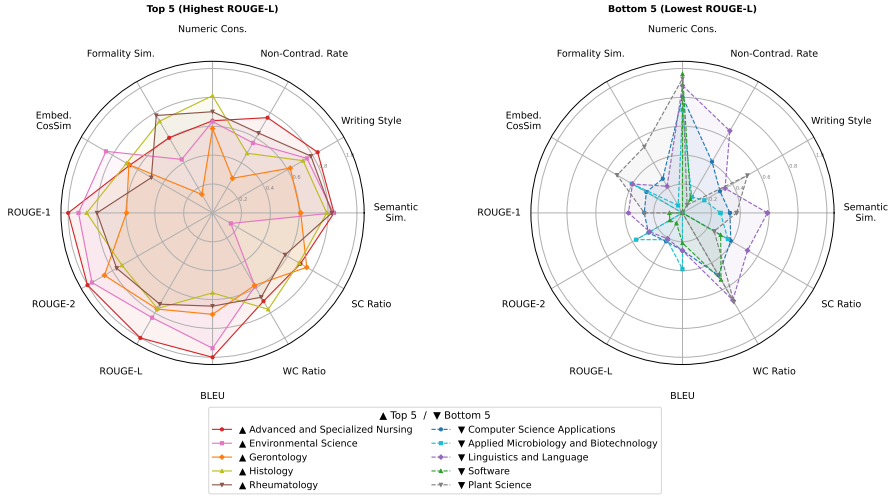
Figure 5 plots journal-level SJR scores against each evaluation metric for `GPT-5.4` under setting A. Most reference-based and judge-based metrics show small but statistically significant positive associations with SJR: ROUGE-1, ROUGE-2, Semantic Similarity, Writing Style Similarity, and Formality Similarity are all significant at $p < 0.001$. By contrast, Perplexity and Non-Contradiction Rate show no significant trend, while Numeric Consistency exhibits a small but significant negative correlation. These results suggest that journals with higher SJR scores tend to produce abstracts whose conclusions are slightly easier to match in terms of lexical overlap and writing

style, yet this advantage does not extend to factual consistency dimensions. The overall effect sizes remain modest, indicating that venue prestige is a weak rather than dominant predictor of conclusion-generation difficulty.

5.2 Category



(a) Top 5 and bottom 5 biomedical categories ranked by mean Semantic Similarity.



(b) Top 5 and bottom 5 biomedical categories ranked by mean ROUGE-L.

Fig. 6: Radar chart comparison of the top 5 and bottom 5 biomedical categories ranked by both mean Semantic Similarity and mean ROUGE-L for `GPT-5.4` under setting Ⓐ. Each axis is normalized to $[0, 1]$.

Figure 6 compares the top-5 and bottom-5 biomedical categories under two ranking criteria: mean semantic similarity (panel a) and mean ROUGE-L (panel b), both for `GPT-5.4` under setting Ⓐ, with all twelve metrics min-max normalized to $[0, 1]$.

When categories are ranked by semantic similarity (panel a), the top-5 categories form large, uniformly filled polygons: high semantic similarity co-occurs with high writing style similarity, numeric consistency, non-contradiction rate, and lexical overlap metrics alike. In contrast, when categories are ranked by ROUGE-L (panel b), the top-5 polygons are visibly lopsided. ROUGE-1/2/L axes are high by construction, but judge-based dimensions such as writing style and numeric consistency do not follow. For example, Gerontology ranks among the top-5 by ROUGE-L yet falls well below the semantic-similarity top-5 on writing style and numeric consistency.

This asymmetry reveals that **lexical overlap with the reference conclusion is not a reliable proxy for overall conclusion quality**. Categories whose generated conclusions share many surface n-grams with the gold reference do not necessarily match its rhetorical style, numeric selectivity, or broader discourse structure. By contrast, high semantic similarity appears to act as a more holistic quality indicator that correlates with strong performance across both reference-based and judge-based dimensions. This observation reinforces the motivation for our hybrid evaluation protocol (Section 4.2): relying on ROUGE or BLEU alone would mask meaningful quality differences that only LLM-as-a-judge scoring can detect.

The bottom-5 categories under both rankings share substantial overlap: Software, Computer Science Applications, and Applied Microbiology and Biotechnology appear in both lists, confirming that these interdisciplinary, non-clinical fields are consistently the hardest for conclusion generation regardless of the evaluation axis. Their radar profiles are highly irregular. For instance, Software has the lowest semantic similarity across all 112 categories (61.0) yet among the highest numeric consistency (96.4), illustrating that no single metric suffices to characterize conclusion-generation difficulty. Appendix E provides representative generation examples from both high- and low-performing categories.

6 Conclusion

We introduce MedConclusion, a large-scale benchmark of 5.7M structured PubMed abstracts for biomedical conclusion generation, paired with author-written conclusions and journal-level metadata. Our experiments show that conclusion generation is behaviorally distinct from summary writing, that strong LLMs remain closely clustered under current automatic metrics, and that absolute LLM-judge scores are sensitive to judge identity. Analyses across journal prestige and biomedical categories further show that task difficulty is heterogeneous and that lexical-overlap metrics alone do not adequately capture conclusion quality.

Supplementary information. Supplementary sections in this source provide the example datapoint, generation prompts, judge prompts, reference-metric definitions, additional category examples, dataset statistics, conclusion-label variants, model configurations, and LLM-usage details.

Acknowledgements. Not applicable.

Declarations

Funding. This work was supported by NIH R01 EY036222, R01 EY038687, and NIH P30 EY003790.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. This study uses publicly available PubMed structured abstracts and public journal metadata from SCImago Journal & Country Rank. No human-participant enrollment or intervention was performed for this study.

Consent for publication. Not applicable.

Data availability. MedConclusion is derived from PubMed structured abstracts and public journal metadata from SCImago Journal & Country Rank. The manuscript, released benchmark assets and code are available at https://github.com/Harvard-AI-and-Robotics-Lab/MedConclusion.

Materials availability. Benchmark metadata, prompts and generated figure assets described in this study are available from the same repository.

Code availability. Code for data processing, evaluation and figure generation is available at https://github.com/Harvard-AI-and-Robotics-Lab/MedConclusion.

Author contributions. Weiyue Li, Ruizhi Qian and Yi Li contributed equally to this work. Yan Luo and Mengyu Wang jointly supervised the study. All authors contributed to manuscript revision and approved the final version.

Appendix A Example data

Example Datapoint from MedConclusion

Metadata:

PMID: 21401313

Title: Inositol phosphoglycan P-type in infants of preeclamptic mothers.

Journal Title: The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians

Journal ISO: *J Matern Fetal Neonatal Med*

Volume: 25 **Issue:** 2

DOI: 10.3109/14767058.2011.557789

Language: eng

Publication Date: 2012-02

Article Date: 2011-03-14

SJR 2012: 0.656

Subject Area and Category:

Medicine

- Obstetrics and Gynecology
- Pediatrics, Perinatology and Child Health

Structured Abstract:

BACKGROUND: Inositol phosphoglycan P-type (P-IPG) has consistently found to be elevated during active preeclampsia, although the biosynthetic source has to be identified yet. This multicenter prospective cross-sectional case-control study evaluated the fetus/newborn as the source of P-IPG.

METHODS: A urine specimen was collected longitudinally for three consecutive days after delivery from 90 newborns and their mothers, and ordered according to clinical diagnosis of preeclampsia, gestational hypertension, or healthy pregnancy.

RESULTS: The urinary excretion of P-IPG on day 0 was higher in the mothers in all groups ($p < 0.05$) with higher levels in preeclamptic women ($p < 0.01$) in the mothers compared to their newborns in the preeclamptic group ($p < 0.01$). The difference persisted at least two days post partum.

CONCLUSION: Findings of this study confirm the specificity of the increase in urinary excretion of P-IPG in preeclamptic mothers at day of birth compared to healthy pregnancy and GH, but does not extend to their newborns.

Appendix B Prompts for conclusion/summary generation

B.1 Prompts for conclusion generation (A and C)

Conclusion Generation (Ⓐ vs. Ⓒ)

System: You are a senior scientist in generating conclusions for scientific papers.

User:
You will be provided with a structured abstract of a scientific paper.
The abstract contains sections such as Background, Objective, Methods, Results, etc., but the corresponding Conclusion section is missing.

Your task is to infer and write the most plausible **CONCLUSION** section that would appear in this abstract.

Here are the requirements:
- Output ONLY the text for the conclusion itself. Do NOT include any section headers or explanations.
- Only use the information provided by the abstract to derive your conclusion.
- Do NOT introduce new experiments, datasets, numerical values, or claims that are not supported by the abstract.

Please generate the conclusion in the same writing style as the given abstract sections. The conclusion should be a concise paragraph with `{sen_num}` sentences totalling `{word_num}` words.

Use formal academic writing style.

Structured Abstract:
```
<Abstract>
{abstract_text}
</Abstract>
```

B.2 Prompts for summary generation (Ⓑ and Ⓓ)

Summary Generation (Ⓑ vs. Ⓓ)

System: You are a senior scientist in generating summaries for scientific papers.

User:
You will be provided with a structured abstract of a scientific paper.
The abstract contains sections such as Background, Objective, Methods, Results, etc.

Your task is to summarize the core information in the given abstract sections.

Here are the requirements:
- Output ONLY the text for the summary itself. Do NOT include any section headers or explanations.
- Only use the information provided by the abstract to derive your summary.
- Do NOT introduce new experiments, datasets, numerical values, or claims that are not supported by the abstract.

Please generate the summary in the same writing style as the given abstract sections. The summary should be a concise paragraph with `{sen_num}` sentences totalling `{word_num}` words.

Use formal academic writing style.

Structured Abstract:
```
<Abstract>
{abstract_text}
</Abstract>
```

Appendix C Prompts for LLM judges

LLM Judge Prompt

System: You are an expert evaluator of scientific writing.

User:
Your task is to compare the Generated Conclusion against the Original (Reference) Conclusion and score multiple dimensions from 0 to 100 (decimals allowed). Use ONLY the two conclusions provided. Do NOT provide any explanations.

Scoring dimensions:
- semantic similarity: How similar the meaning and core claims are.
- writing style similarity: How similar the tone, phrasing, structure, and rhetorical style are.
- contradiction rate: Degree of contradiction between the Generated Conclusion and the Original Conclusion.
  - 100 = no contradiction
  - 0 = severe contradiction
- numeric consistency: Consistency of all numerical information, quantities, directions, and magnitudes.
  - 100 = fully consistent or no numeric content in either text
  - 0 = major numeric inconsistency
- formality similarity: Similarity in academic/formal writing level and register.

INPUT:
- Original Conclusion (reference): `{original_conclusion}`
- Generated Conclusion: `{generated_conclusion}`

OUTPUT FORMAT (STRICT):
```
{"semantic similarity": <0-100>, "writing style similarity": <0-100>,
"contradiction rate": <0-100>, "numeric consistency": <0-100>, "formality
similarity": <0-100>}
```

Appendix D Reference-based metrics

In addition to LLM-as-a-judge scoring, we report a set of lightweight diagnostics and reference-based metrics that compare the generated conclusion $\hat{y}$ against the author-written reference conclusion $y^\star$. These metrics are inexpensive to compute, easy to reproduce, and provide complementary signals about lexical overlap, semantic proximity, length control, and fluency. We do not treat any single metric as a complete measure of conclusion quality; rather, we use them as a bundle of auxiliary indicators.

Let $|y|_{\text{word}}$ denote the number of words in a text y, and let $|y|_{\text{sent}}$ denote its number of sentences.

Word-count ratio.

To measure length matching at the tokenized word level, we compute

$$\text{WCR}(\hat{y}, y^\star) = \frac{|\hat{y}|_{\text{word}}}{|y^\star|_{\text{word}}}. \tag{D1}$$

A value close to 1 indicates that the generated conclusion has similar length to the reference. Values below 1 indicate shorter generations, while values above 1 indicate longer generations.

Sentence-count ratio.

To assess structural length control at the sentence level, we compute

$$\mathrm{SCR}(\hat{y}, y^\star) = \frac{|\hat{y}|_{\mathrm{sent}}}{|y^\star|_{\mathrm{sent}}}. \tag{D2}$$

Embedding cosine similarity.

To capture semantic similarity beyond surface lexical overlap, we encode $\hat{y}$ and $y^\star$ using an off-the-shelf ALL-MPNET-BASE-V2[3] sentence embedding model and compute the cosine similarity between the two vector representations [33]:

$$\mathrm{CosSim}(\hat{y}, y^\star) = \frac{\phi(\hat{y})^\top \phi(y^\star)}{\|\phi(\hat{y})\| \, \|\phi(y^\star)\|}, \tag{D3}$$

where $\phi(\cdot)$ denotes the embedding function. Higher values indicate greater semantic proximity.

ROUGE.

We report ROUGE-1, ROUGE-2, and ROUGE-L [31]. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L measures longest-common-subsequence overlap. These metrics quantify the extent to which the generated conclusion reuses words and short phrases appearing in the reference. Because multiple valid conclusions may use different wording, ROUGE should be interpreted as a lexical-overlap signal rather than a direct measure of scientific correctness.

BLEU.

We also report BLEU [32], which measures n-gram precision of the generated text with a brevity penalty. As with ROUGE, BLEU is sensitive to phrasing and therefore is best viewed as an approximate indicator of closeness to the reference wording, not as a standalone measure of conclusion quality.

Perplexity under an external language model.

To estimate fluency and distributional typicality, we compute perplexity using a fixed external language model, `GPT-2`, on both the reference conclusion $y^\star$ and the generated conclusion $\hat{y}$. For a sequence of tokens $y = (w_1, \ldots, w_T)$, perplexity under language model p is

$$\mathrm{PPL}(y) = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \log p(w_t \mid w_{<t}) \right). \tag{D4}$$

Lower perplexity indicates that the text is more probable under the external language model. Reporting perplexity for both $y^\star$ and $\hat{y}$ helps contextualize whether model-generated conclusions are comparably fluent to author-written ones under the same scoring model.

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Implementation notes.

All reference-based metrics are computed between each generated conclusion $\hat{y}$ and its paired author-written conclusion $y^{\star}$. We aggregate scores over the evaluation set using the arithmetic mean unless otherwise noted. Since these metrics emphasize different aspects of generation quality, we recommend interpreting them jointly with the LLM-as-a-judge results in the main paper.

Appendix E Additional category analysis

E.1 Example 1

Example: Experimental and Cognitive Psychology (HIGH)

Metadata:
PMID: 32422422
Journal Title: Cortex
Publication Year: 2020
SJR 2020: 1.786
Subject Area and Category:
- Experimental and Cognitive Psychology

Structured Abstract:

BACKGROUND: Numerous studies have shown visuoperceptual/visuospatial deficits in dementia with Lewy bodies (DLB) and Alzheimer's disease (AD). Visual texture recognition is also impaired in patients with DLB and AD. Although patients with DLB often exhibit visual misidentifications of objects, there are few studies on the relationships between visual texture recognition and viewpoints for object recognition.

OBJECTIVES: The aim of this study was to clarify how viewpoints, textures, and visual cognitive functions affect object recognition and result in visual misidentifications in patients with DLB or AD.

METHODS: A total of 37 patients with probable DLB and 58 with probable AD and 32 age-matched healthy controls underwent neuropsychological and visuoperceptual assessments, and performed object identification tasks under four conditions (non-canonical view + blurry texture, non-canonical view + clear texture, canonical view + blurry texture, canonical view + clear texture). The relationship between object identification and other visuoperceptual functions was analyzed.

RESULTS: Patients with DLB and AD exhibited significantly impaired object recognition under non-canonical viewing with blurry texture conditions, with the DLB patients exhibiting a significantly worse performance than the AD patients. Patients with DLB and AD exhibited visual misidentifications during object identification tasks under non-canonical viewing. In patients with DLB, the number of visual misidentifications was significantly correlated with the scores of visual texture recognition.

CONCLUSION: The present study showed that significantly impaired object recognition in patients with DLB under the influences by both viewpoint and visual texture and in those with AD under the influence by viewpoint. Visual misidentification in object recognition could be associated with impaired visual texture recognition in DLB.

GENERATED: Object recognition is particularly vulnerable in DLB and, to a lesser extent, in AD when objects are presented from non-canonical viewpoints with degraded texture information. The greater impairment in DLB and the association between visual misidentifications and visual texture recognition in this group suggest that deficits in processing texture cues, especially under atypical viewing conditions, contribute to visual misidentification. These findings indicate that altered viewpoint and texture processing are important mechanisms underlying visual recognition disturbances in DLB.

E.2 Example 2

Example: Endocrine and Autonomic Systems (HIGH)

Metadata:
PMID: 37580720
Journal Title: Thyroid research
Publication Year: 2023
SJR 2023: 0.492
Subject Area and Category:
- Endocrine and Autonomic Systems

Structured Abstract:

PURPOSE: Lacrimal gland enlargement can be a feature of thyroid eye disease (TED). Unilateral or asymmetric lacrimal gland enlargement is poorly described and may impede diagnosis. We present the histological and clinical findings of four patients with asymmetric lacrimal gland enlargement.

METHODS: A retrospective case note review was performed for patients over two tertiary orbital clinics (Royal Adelaide Hospital, South Australia and the Sussex Eye Hospital, Brighton, United Kingdom) presenting with an asymmetrical lacrimal gland enlargement with a background of TED that underwent biopsy to exclude alternate diagnoses. Baseline data was collected for each patient and histopathological images and reports were reviewed.

RESULTS: All four patients were hyperthyroid at time of lacrimal gland biopsy. Biopsy demonstrated nonspecific, lymphoid aggregates, typically of B cell type, with no diagnostic findings to support lymphocyte clonality or IgG4-related disease. One biopsy specimen demonstrated evidence of some fibrosis.

CONCLUSION: Asymmetrical lacrimal gland enlargement can occur as part of the TED spectrum but may require biopsy to exclude alternate pathology. Histology demonstrates a non-specific lymphocytic infiltrate.

GENERATED: In patients with thyroid eye disease, asymmetric lacrimal gland enlargement may occur and can mimic alternative orbital pathology. In this series, biopsy showed only nonspecific benign lymphoid infiltration without evidence of lymphoma or IgG4-related disease, supporting asymmetric lacrimal gland involvement as a possible manifestation of TED. These findings may aid clinical recognition, although biopsy remains important when the presentation is atypical and alternative diagnoses must be excluded.

E.3 Example 3

Example: Advanced and Specialized Nursing (HIGH)

Metadata:
PMID: 32951753
Journal Title: Complementary therapies in medicine
Publication Year: 2020
SJR 2020: 0.58
Subject Area and Category:
- Advanced and Specialized Nursing

Structured Abstract:

BACKGROUND AND OBJECTIVE: Walnut intake is considered a healthy dietary approach worldwide, particularly as a nutritional tool for the management of obesity and cardiometabolic disorders. Among these lines, leptin and adiponectin, as well as glycemic biomarkers, deserve further attention. We aimed to examine the impact of walnut intake on circulation levels of leptin and adiponectin through a systematic review and meta-analysis of randomized clinical trials (RCTs); secondarily, assessing the glycemic profile as well.

METHODS: The literature search was implemented in four following databases: Web of Science, Scopus, PubMed/Medline, and Google Scholar, thus, determining studies that measured the effects of walnut consumption on adiponectin, leptin, and glycemic biomarkers levels from 2004 up to December 2019.

RESULTS: Fourteen trials were include in the meta-analysis, with an intervention period ranging from 5 weeks to 12 months.Walnut intake increased leptin (weighted mean difference (WMD): 2.502 ng/mL; 95 % CI: 2.147-2.856, p 0.001) and adiponectin (WMD: 0.440 ng/mL; 95 % CI: 0.323 to 0.557, p 0.001) levels. Pertaining to glycemic biomarkers, neither overall analyses nor sub-analyses corroborated with changes in fasting blood glucose (WMD: 0.500 mg/dL, 95 % CI: -0.596, 1.596, p = 0.371), insulin (WMD: -0.21 mg/dL, 95 % CI: -0.67, 0.24, p = 0.367), and glycated hemoglobin (WMD: 0.004 mg/dL, 95 % CI: -0.041, 0.049, p = 0.870) concentrations.

CONCLUSION: Walnut intake may increase leptin and adiponectin levels but does not improve glycemic biomarkers.

GENERATED: Walnut consumption appears to favorably modulate adipokine status by increasing circulating adiponectin and leptin levels, while showing no significant effect on fasting blood glucose, insulin, or glycated hemoglobin. These findings suggest that walnuts may contribute to obesity and cardiometabolic management primarily through effects on adipokine regulation rather than glycemic control. Further well-designed randomized clinical trials are warranted to confirm these effects and clarify their clinical relevance.

E.4 Example 4

Example: Environmental Science (HIGH)

Metadata:
PMID: 31054526
Journal Title: Environmental research
Publication Year: 2019
SJR 2019: 1.52
Subject Area and Category:
Environmental Science
Structured Abstract:

BACKGROUND: Hypertension and air pollution are two important risk factors for cardiovascular morbidity and mortality. Although several studies suggest that air pollution has a significant impact on blood pressure, studies on long-term effects are sparse and still controversial.

OBJECTIVE: To evaluate the effects of exposure of outdoor workers to different levels of traffic-generated PM2.5 on blood pressure.

DESIGN: This is an observational panel study.

PARTICIPANTS: 88 non-smoking workers exposed to different concentrations of air pollution were evaluated weekly along four successive weeks.

MEASUREMENTS: In each week, personal monitoring of 24-h PM2.5 concentration and 24-h ambulatory blood pressure were measured. The association between blood pressure variables and PM2.5, adjusted for age, body mass index, time in job, daily work hours, diabetes, hypertension and cholesterol was assessed by means of multiple linear regression models fitted by least squares.

RESULTS: Exposure to PM2.5 (ranging from 8.5 to 89.7 μg/m[3]) is significantly and consistently associated with an increase in average blood pressure. An elevation of 10 μg/m[3] in the concentration of PM2.5 is associated with increments of 3.9 mm Hg (CI 95% = [1.5; 6.3]) in average systolic 24-h blood pressure for hypertensive and/or diabetic workers.

CONCLUSION: Exposure to fine particles, predominantly from vehicular traffic, is associated with elevated blood pressure in hypertensive and/or diabetic workers.

GENERATED: Personal exposure to traffic-generated PM2.5 was positively associated with higher 24-hour blood pressure in this cohort of outdoor workers, with particularly pronounced effects among workers with hypertension and/or diabetes. These findings support the view that long-term exposure to fine particulate air pollution contributes to elevated blood pressure and may increase cardiovascular risk in susceptible populations.

E.5 Example 5

Example: Emergency Nursing (HIGH)

Metadata:
PMID: 21458134
Journal Title: Resuscitation
Publication Year: 2011
SJR 2011: 1.736
Subject Area and Category:
- Emergency Nursing

Structured Abstract:

AIM: Body mass index (BMI) may influence the quality of cardiopulmonary resuscitation and may influence prognosis after cardiac arrest. To review the direct effect of obesity on outcome after cardiac arrest, the following cohort study was conducted.

METHODS: This study based on a cardiac arrest registry comprising all adult patients with cardiac arrest of non-traumatic origin and restoration of spontaneous circulation (ROSC) admitted to the department of emergency medicine of a tertiary-care facility. Data were collected between January 1992 and December 2007 according to the Utstein criteria. We assessed the association between BMI according to the WHO classification (underweight, BMI18.5; normal weight, 18.5-24.9; overweight, 25.0-29.9; obese $\geq$ 30), six-month survival and neurological recovery.

RESULTS: Analysis was carried out on a total of 1915 adult patients (32% female). Patients had a median age of 59 years (interquartile range [IQR] 49-70) and a median BMI of 26.0 (IQR 23.9-29.1). Survival to six months was 50%. There was no significant difference in survival between the BMI groups (underweight 46%, normal weight 47%, overweight 52%, obese 51%). In a multivariate analysis neurological outcome was better in overweight patients as compared to subjects with normal BMI (odds ratio 1.35; 95% confidence interval 1.02-1.79).

CONCLUSION: Body mass index may have no direct influence on six-month survival after cardiac arrest, but patients with moderately elevated BMI may have a better neurological prognosis.

GENERATED: In this cohort of adult patients with non-traumatic cardiac arrest and ROSC, BMI was not significantly associated with six-month survival. However, overweight patients showed a modestly better neurological outcome than patients with normal BMI after adjustment for confounders, suggesting that excess body weight does not adversely affect prognosis after cardiac arrest and may be associated with improved neurological recovery.

E.6 Example 6

Example: Pollution (LOW)

Metadata:
PMID: 36497716
Journal Title: International journal of environmental research and public health
Publication Year: 2022
SJR 2022: 0.828
Subject Area and Category:
- Pollution

Structured Abstract:

BACKGROUND: Urinary incontinence (UI) and poor sleep negatively affect health-related quality of life (HRQoL). This study explored the UI-related factors and the relationships between UI, sleep quality, and HRQoL.

METHODS: This cross-sectional study collected data from 237 women with type 2 diabetes. Multivariate logistic regression was conducted to identify the factors associated with UI. One-way analysis of variance was used to compare the mean sleep quality and HRQoL scores of women without UI and those who experienced UI of varying severities. Correlation coefficients were estimated, and multivariate linear regression was conducted to examine the relationships between UI severity, sleep quality, and HRQoL.

RESULTS: Of the 237 women, 115 (48.52%) experienced UI and 139 (58.65%) were poor sleepers. The three factors associated with UI were advanced age, a higher body mass index, and a history of vaginal delivery. Significant associations between UI severity and sleep quality and between sleep quality and HRQoL were revealed. UI severity and night-time voiding frequency were both associated with sleep quality.

CONCLUSION: One factor associated with UI (body mass index) is modifiable. UI severity is associated with sleep quality as the possible influence of night-time voiding frequency on sleep quality has been considered.

GENERATED: Urinary incontinence is common among women with type 2 diabetes and is associated with older age, higher body mass index, and a history of vaginal delivery. Greater UI severity, particularly with more frequent night-time voiding, is related to poorer sleep quality, and poor sleep is in turn associated with worse health-related quality of life. These findings highlight the importance of routine assessment and management of UI and sleep problems in this population to improve overall well-being.

E.7 Example 7

Example: Health, Toxicology and Mutagenesis (LOW)

Metadata:
PMID: 28934092
Journal Title: Environmental health perspectives
Publication Year: 2017
SJR 2017: 3.41
Subject Area and Category:
- Health, Toxicology and Mutagenesis

Structured Abstract:

BACKGROUND: Some epidemiologic and laboratory studies suggest that insecticides are related to increased breast cancer risk, but the evidence is inconsistent. Women engaged in agricultural work or who reside in agricultural areas may experience appreciable exposures to a wide range of insecticides.

OBJECTIVE: We examined associations between insecticide use and breast cancer incidence among wives of pesticide applicators (farmers) in the prospective Agricultural Health Study.

METHODS: Farmers and their wives provided information on insecticide use, demographics, and reproductive history at enrollment in 1993-1997 and in 5-y follow-up interviews. Cancer incidence was determined via cancer registries. Among 30,594 wives with no history of breast cancer before enrollment, we examined breast cancer risk in relation to the women's and their husbands' insecticide use using Cox proportional hazards regression to estimate adjusted hazard ratios (HRs) and 95% confidence intervals (CIs).

RESULTS: During an average 14.7-y follow-up, 39% of the women reported ever using insecticides, and 1,081 were diagnosed with breast cancer ... Among women who did not apply pesticides, the husband's use of fonofos was associated with elevated risk, although no exposure-response trend was observed.

CONCLUSION: Use of several organophosphate insecticides was associated with elevated breast cancer risk. However, associations for the women's and husbands' use of these insecticides showed limited concordance. Ongoing cohort follow-up may help clarify the relationship, if any, between individual insecticide exposures and breast cancer risk. https://doi.org/10.1289/EHP1295.

GENERATED: Overall insecticide use was not associated with breast cancer incidence among wives of pesticide applicators; however, use of specific insecticides, particularly the organophosphates chlorpyrifos and terbufos, was associated with modestly increased risk. These associations appeared to be driven primarily by premenopausal breast cancer, with little evidence of variation by estrogen receptor status. The observed elevation in risk with husbands' fonofos use among nonapplying wives, in the absence of an exposure-response trend, warrants cautious interpretation. These findings suggest that certain insecticides may contribute to breast cancer risk in this agricultural population and merit further investigation.

E.8 Example 8

Example: Computer Science Applications (LOW)

Metadata:
PMID: 39014177
Journal Title: International journal of computer assisted radiology and surgery
Publication Year: 2025
SJR 2025: 0.658
Subject Area and Category:
- Computer Science Applications

Structured Abstract:

PURPOSE: Augmented reality guidance in laparoscopic liver resection requires the registration of a preoperative 3D model to the intraoperative 2D image. However, 3D-2D liver registration poses challenges owing to the liver's flexibility, particularly in the limited visibility conditions of laparoscopy. Although promising, the current registration methods are computationally expensive and often necessitate manual initialisation.

METHODS: The first neural model predicting the registration (NM) is proposed, represented as 3D model deformation coefficients, from image landmarks. The strategy consists in training a patient-specific model based on synthetic data generated automatically from the patient's preoperative model. A liver shape modelling technique, which further reduces time complexity, is also proposed.

RESULTS: The NM method was evaluated using the target registration error measure, showing an accuracy on par with existing methods, all based on numerical optimisation. Notably, NM runs much faster, offering the possibility of achieving real-time inference, a significant step ahead in this field.

CONCLUSION: The proposed method represents the first neural method for 3D-2D liver registration. Preliminary experimental findings show comparable performance to existing methods, with superior computational efficiency. These results suggest a potential to deeply impact liver registration techniques.

GENERATED: A patient-specific neural approach can accurately estimate 3D-2D liver registration from image landmarks while substantially reducing computational cost relative to optimisation-based methods. By leveraging automatically generated synthetic data and an efficient liver shape modelling strategy, the proposed framework removes the need for costly numerical optimisation and supports the prospect of real-time augmented reality guidance in laparoscopic liver resection.

E.9 Example 9

Example: Applied Microbiology and Biotechnology (LOW)

Metadata:
PMID: 26497155
Journal Title: Journal of applied microbiology
Publication Year: 2016
SJR 2016: 0.84
Subject Area and Category:
Biochemistry, Genetics and Molecular Biology
Immunology and Microbiology
Medicine
- Applied Microbiology and Biotechnology
- Biotechnology
- Medicine

Structured Abstract:

AIMS: The effect of ohmic heating (OH) in a pilot plant system which had a zig-zag shaped (elbow-type) ohmic heater with five sequential voltage electrodes was investigated on Bacillus cereus spores in a commercial tsuyu sauce.

METHODS AND RESULTS: The electrical field was fixed at 26·7 V cm(-1) with an alternating current frequency of 25 kHz. Raw tsuyu sauce (50 l) inoculated with B. cereus spores was submitted in a 4 × 3 factorial design to the OH system and heated at 95, 105, 115, and 125°C each for 30, 60, and 90 s. Survival of B. cereus spores and colour change in the commercial tsuyu sauce were both measured before and after treatment. As the treatment temperature and time increased, the number of surviving B. cereus spores decreased. The OH treatment in a bath-type process at 105°C for more than 30 s resulted in the total inactivation of the inoculated B. cereus spores (average 5·4 log reductions to undetectable levels after treatment). The OH protocol of heating at 105°C for 60 s which ensure complete eradication of the inoculated spores without compromising product quality was chosen and investigated for its suitability for commercial application on bulk quantities of samples (500 l). Reliable and reproducible reductions in B. cereus spore counts of 4·7-5·5 log CFU ml(-1) (mean ± standard deviation = 5·1 ± 0·3 CFU ml(-1)) were achieved by the selected protocol of the continuous OH treatment (105°C for 60 s).

SIGNIFICANCE AND IMPACT OF THE STUDY: This procedure will enhance the microbiological quality of liquid foods while minimizing quality deterioration.

CONCLUSION: This study suggests that OH treatment with five sequential elbow-type electrodes has great potential as an industrial sterilizing method for liquid food contaminated with B. cereus spores.

GENERATED: Ohmic heating using the pilot-scale zig-zag electrode system effectively inactivated Bacillus cereus spores in commercial tsuyu sauce in a temperature- and time-dependent manner. Heating at 105°C for 60 s achieved complete spore eradication in the bath-type process and produced reliable, reproducible reductions during continuous treatment of bulk samples, while maintaining product colour quality. These findings indicate that the developed ohmic heating protocol is a practical approach for commercial processing of liquid foods to improve microbiological safety with minimal quality deterioration.

E.10 Example 10

Example: Software (LOW)

Metadata:
PMID: 29157445
Journal Title: Computer methods and programs in biomedicine
Publication Year: 2018
SJR 2018: 0.753
Subject Area and Category:
- Software

Structured Abstract:

BACKGROUND AND OBJECTIVES: Diabetic retinopathy (DR) is one of the leading causes of preventable blindness in the world. Its earliest sign are red lesions, a general term that groups both microaneurysms (MAs) and hemorrhages (HEs). In daily clinical practice, these lesions are manually detected by physicians using fundus photographs. However, this task is tedious and time consuming, and requires an intensive effort due to the small size of the lesions and their lack of contrast. Computer-assisted diagnosis of DR based on red lesion detection is being actively explored due to its improvement effects both in clinicians consistency and accuracy. Moreover, it provides comprehensive feedback that is easy to assess by the physicians. Several methods for detecting red lesions have been proposed in the literature, most of them based on characterizing lesion candidates using hand crafted features, and classifying them into true or false positive detections. Deep learning based approaches, by contrast, are scarce in this domain due to the high expense of annotating the lesions manually.

METHODS: In this paper we propose a novel method for red lesion detection based on combining both deep learned and domain knowledge. Features learned by a convolutional neural network (CNN) are augmented by incorporating hand crafted features. Such ensemble vector of descriptors is used afterwards to identify true lesion candidates using a Random Forest classifier.

RESULTS: We empirically observed that combining both sources of information significantly improve results with respect to using each approach separately. Furthermore, our method reported the highest performance on a per-lesion basis on DIARETDB1 and e-ophtha, and for screening and need for referral on MESSIDOR compared to a second human expert.

CONCLUSION: Results highlight the fact that integrating manually engineered approaches with deep learned features is relevant to improve results when the networks are trained from lesion-level annotated data. An open source implementation of our system is publicly available at https://github.com/ignaciorlando/red-lesion-detection.

GENERATED: The proposed red lesion detection framework, which integrates CNN-derived features with hand-crafted descriptors and classifies lesion candidates using a Random Forest, provides superior performance to either deep learned or hand-crafted features alone. Its strong results across DIARETDB1, e-ophtha, and MESSIDOR indicate that combining data-driven representation learning with domain knowledge is an effective strategy for accurate red lesion detection and for supporting diabetic retinopathy screening and referral decisions in clinical practice.

E.11 The conclusion–summary distinction holds across categories

Section 4.2 shows that summary-mode outputs recover most of the semantic similarity of conclusion-mode outputs while diverging sharply in writing style and numeric consistency. We now test whether this pattern is universal or driven by a subset of categories, by computing the per-category gap (Mode Ⓐ−Ⓑ) across all five judge dimensions for `GPT-5.4`.

Across all 112 categories, the gap is *always* positive for writing style similarity (min +3.8, mean +8.3, max +13.7) and numeric consistency (min +11.3, mean +21.6, max +41.3).

Table E15 shows the ten categories with the largest |ΔSemantic Similarity|. Even within this group, the gap magnitudes vary substantially across dimensions: numeric consistency ranges from +18.6 to +33.0 and writing style from +4.5 to +13.7, indicating that the specific dimensions along which the two modes diverge are category-dependent. Biotechnology is a notable outlier: it is the only entry where summary-mode outputs are semantically *closer* to the reference ($\Delta = -4.0$), yet numeric consistency still drops by +27.2 points.

Table E16 isolates the ten categories where the semantic gap is smallest ($|\Delta| < 0.3$). Even here, writing style still differs by +4.8 to +8.4 points and numeric consistency by +14.3 to +23.1 points. Semantic convergence does not imply behavioral equivalence: summaries differ from conclusions in phrasing, structure, and numeric detail even when their meaning is indistinguishable. This confirms that the conclusion–summary distinction is a structural property of the two discourse functions, not an artifact of category-level heterogeneity.

Category	Δ Semantic Sim.	Δ Writing Style	Δ Non-Contrad. Rate	Δ Numeric Cons.	Δ Formality Sim.
Immunology & Microbiology	+6.0	+13.0	+4.8	+28.0	+2.3
Pharm., Toxicology & Pharmaceutics	+5.5	+13.5	+4.9	+33.0	+2.4
Exp. & Cognitive Psychology	+5.3	+7.9	+3.4	+27.5	+2.1
Arts & Humanities	+5.1	+12.3	+6.5	+32.4	+2.8
Virology	+4.8	+13.0	+7.2	+23.2	+3.1
Nursing	+4.4	+13.7	+3.8	+30.3	+1.8
Social Psychology	+4.1	+11.4	+2.9	+18.6	+2.1
Biotechnology	−4.0	+4.5	−4.6	+27.2	+0.6
Education	+3.9	+12.3	+3.6	+27.6	+1.7
Health Informatics	+3.5	+12.7	+4.0	+31.5	+3.1

Table E15: Top 10 categories ranked by the absolute semantic-similarity gap between modes A and B for `GPT-5.4`. Gap magnitudes vary substantially across dimensions within this group, indicating that the specific dimensions along which conclusion and summary modes diverge are category-dependent.

Category	Δ Semantic Sim.	Δ Writing Style	Δ Non-Contrad. Rate	Δ Numeric Cons.	Δ Formality Sim.
Cancer Research	+0.2	+7.6	−0.5	+16.7	+0.7
Organic Chemistry	−0.2	+7.1	−1.2	+18.1	+1.2
Endocrinology, Diabetes & Metab.	+0.1	+7.3	−0.2	+14.3	+0.9
Cellular & Mol. Neuroscience	−0.1	+6.8	+0.4	+16.5	+1.1
Hepatology	−0.1	+4.8	+0.7	+15.1	+1.1
Biochemistry	−0.1	+5.4	+0.6	+16.4	+0.2
Applied Psychology	+0.0	+5.9	−1.2	+23.1	+2.5
Dentistry	−0.0	+8.4	−0.5	+19.6	+1.1
Critical Care & ICM	+0.0	+5.6	+0.6	+18.6	+0.4
Pharmaceutical Science	−0.0	+6.1	−0.2	+17.7	+1.2

Table E16: Bottom 10 categories ranked by the absolute semantic-similarity gap between modes A and B for `GPT-5.4`. Despite near-zero semantic gaps, writing style (+4.8 to +8.4) and numeric consistency (+14.3 to +23.1) remain substantially positive.

Appendix F MedConclusion dataset statistics

SJR.

The SJR score quantifies a journal's scientific influence by accounting for both the volume of citations received and the prestige of the citing sources. It is computed via an iterative algorithm analogous to PageRank: a citation from a highly-ranked journal contributes more to a journal's SJR score than one from a lower-ranked journal, and self-citations are down-weighted to mitigate inflation [46]. This design renders SJR size-independent and more robust to citation manipulation than raw impact factors. Because SJR scores are published annually, our dataset captures the longitudinal prestige trajectory of each journal from its earliest available record through 2024.

F.1 Abstracts' year distribution

F.2 Journal category & SJR score distribution

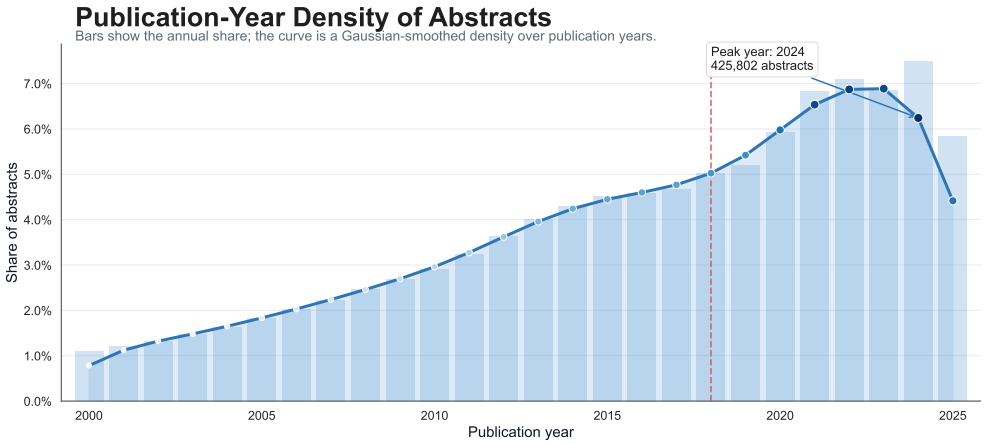


Fig. F1: Publication-year density of abstracts in MedConclusion from 2000–2025. The median publication year is 2018.

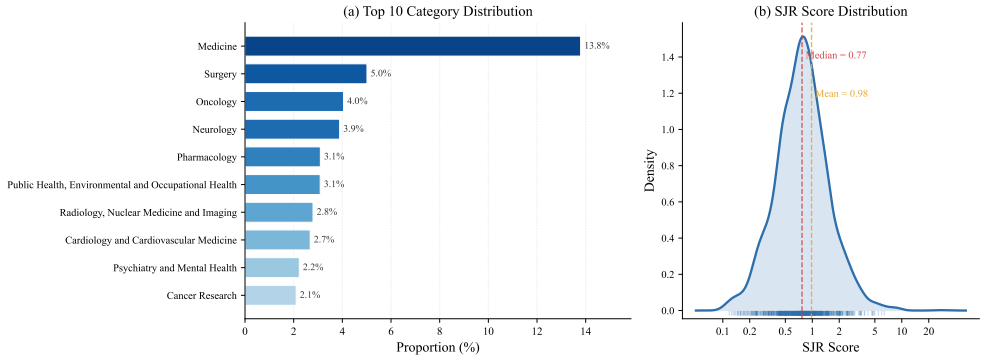


Fig. F2: Distribution statistics of MedConclusion. (a) Top 10 subject categories by proportion of abstracts in the dataset, with Medicine being the dominant category at 13.8%. (b) Distribution of SJR scores across the 3,772 journals (median = 0.77, mean = 0.98), showing a right-skewed distribution concentrated in the low-to-moderate prestige range.

Appendix G Conclusion Label Variants

We use the conclusion label variants shown below to identify conclusion-type sections in structured abstracts.

Conclusion Label Variants

- CONCLUSION
- CONCLUSIONS
- CONCLUSION(S)
- CONCLUSIONS AND RELEVANCE
- CONCLUSION AND RELEVANCE
- CONCLUSIONS AND IMPLICATIONS
- CONCLUSION AND IMPLICATIONS
- CONCLUSIONS AND IMPORTANCE
- CONCLUSION AND IMPORTANCE
- CONCLUSION AND SIGNIFICANCE
- CONCLUSIONS AND SIGNIFICANCE
- CONCLUSION AND INTERPRETATION
- CONCLUSIONS AND INTERPRETATION
- CONCLUSIONS AND CLINICAL RELEVANCE
- CONCLUSION AND CLINICAL RELEVANCE
- CONCLUSIONS AND CLINICAL IMPORTANCE
- CONCLUSION AND CLINICAL IMPORTANCE
- AUTHORS' CONCLUSIONS
- AUTHORS' CONCLUSION
- MAIN CONCLUSIONS
- MAIN CONCLUSION

Appendix H Model configurations

Short Name	Full Name	Access	Capability	Scale	Max New Tokens	Temp.
General-purpose Models						
GPT-5.4	gpt-5.4 [44]	Proprietary	General-purpose	Large	1024	0
Gemini 3.1 Pro	gemini-3.1-pro [47]	Proprietary	General-purpose	Large	1024	0
Gemini 3 Flash	gemini-3-flash [45]	Proprietary	General-purpose	Medium	1024	0
DeepSeek-V3.2	DeepSeek-V3.2 [48]	Proprietary	General-purpose	Large	1024	0
MiniMax-M2.1	MiniMax-M2.1 [49]	Proprietary	General-purpose	Large	1024	0
Gemma-3-27B	gemma-3-27b-it [50]	Open-weight	General-purpose	Medium	1024	0
Llama-3.1-8B	Llama-3.1-8B-Instruct [51]	Open-weight	General-purpose	Small	1024	0
Gemma-2-9B	gemma-2-9b-it [52]	Open-weight	General-purpose	Small	1024	0
Qwen2.5-7B	Qwen2.5-7B-Instruct [53]	Open-weight	General-purpose	Small	1024	0
Qwen3-4B	Qwen3-4B-Instruct-2507 [54]	Open-weight	General-purpose	Small	1024	0
Llama-3.2-1B	Llama-3.2-1B-Instruct [55]	Open-weight	General-purpose	Small	1024	0
Reasoning Models						
Kimi-K2	Kimi-K2-Thinking [56]	Open-weight	Reasoning	Large	1024	0
DeepSeek-R1	DeepSeek-R1 [57]	Open-weight	Reasoning	Large	1024	0
Vision-Language Models						
GLM-4.6V	GLM-4.6V [58]	Proprietary	Vision-language	Large	1024	0
Qwen2.5-VL-7B	Qwen2.5-VL-7B-Instruct [59]	Open-weight	Vision-language	Small	1024	0

Table H18: Model metadata and run-configuration summary for all evaluated models. For general-purpose models that expose a `thinking` mode, `thinking=none` was used.

Appendix I The use of Large Language Models (LLMs)

LLM is used only to aid writing quality (proofreading and polishing grammar). No ideas, claims, methods, results, or references are generated by LLMs. All content decisions and revisions are made by the authors.

References

[1] Luo, Z., Yang, Z., Xu, Z., Yang, W. & Du, X. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306* (2025).

[2] Zheng, T. *et al. From automation to autonomy: A survey on large language models in scientific discovery*, 17744–17761 (2025).

[3] Li, W. *et al.* Llm review: Enhancing creative writing via blind peer review feedback. *arXiv preprint arXiv:2601.08003* (2026).

[4] Zhang, J., Petrui, C., Nikolić, K. & Tramèr, F. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. *arXiv preprint arXiv:2505.12575* (2025).

[5] Yu, Z. *et al.* Formalmath: Benchmarking formal mathematical reasoning of large language models. *arXiv preprint arXiv:2505.02735* (2025).

[6] Fein, D. *et al. Litbench: A benchmark and dataset for reliable evaluation of creative writing*, 7740–7755 (2026).

[7] Shieh, A. T.-W., Chuang, Y.-S., Su, S.-Y. & Chen, Y.-N. *Towards understanding of medical randomized controlled trials by conclusion generation*, 108–117 (2019).

[8] Tang, L. *et al. Echogen: generating conclusions from echocardiogram notes*, 359–368 (2022).

[9] Gao, Y., Gu, N., Lam, J., Henderson, J. & Hahnloser, R. *Evaluating unsupervised argument aligners via generation of conclusions of structured scientific abstracts*, 151–160 (2024).

[10] Bastan, M., Shankar, N., Surdeanu, M. & Balasubramanian, N. Calzolari, N. *et al.* (eds) *SuMe: A dataset towards summarizing biomedical mechanisms.* (eds Calzolari, N. *et al.*) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6922–6931 (European Language Resources Association, Marseille, France, 2022). URL https://aclanthology.org/2022.lrec-1.748/.

[11] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. *Pubmedqa: A dataset for biomedical research question answering*, 2567–2577 (2019).

[12] Tsatsaronis, G. *et al.* An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics* **16**, 138 (2015).

[13] Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).

[14] Pal, A., Umapathi, L. K. & Sankarasubbu, M. Flores, G., Chen, G. H., Pollard, T., Ho, J. C. & Naumann, T. (eds) *Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering.* (eds Flores, G., Chen, G. H., Pollard, T., Ho, J. C. & Naumann, T.) *Proceedings of the Conference on Health, Inference, and Learning*, Vol. 174 of *Proceedings of Machine Learning Research*, 248–260 (PMLR, 2022). URL https://proceedings.mlr.press/v174/pal22a.html.

[15] Nye, B. *et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature*, 197–207 (2018).

[16] Lehman, E., DeYoung, J., Barzilay, R. & Wallace, B. C. *Inferring which medical treatments work from reports of clinical trials*, 3705–3717 (2019).

[17] DeYoung, J., Lehman, E., Nye, B., Marshall, I. & Wallace, B. C. *Evidence inference 2.0: More data, better models*, 123–132 (2020).

[18] Wadden, D. *et al. Fact or fiction: Verifying scientific claims*, 7534–7550 (2020).

[19] Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. *On faithfulness and factuality in abstractive summarization*, 1906–1919 (2020).

[20] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023).

[21] Liu, Y. *et al. G-eval: Nlg evaluation using gpt-4 with better human alignment*, 2511–2522 (2023).

[22] Shi, L. *et al. Judging the judges: A systematic study of position bias in llm-as-a-judge*, 292–314 (2025).

[23] Huang, H. *et al. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4*, 5880–5895 (2025).

[24] Wang, J. *et al.* Evidencebench: A benchmark for extracting evidence from biomedical papers. *arXiv preprint arXiv:2504.18736* (2025).

[25] Teufel, S. & Moens, M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* **28**, 409–445 (2002).

[26] Cohan, A. *et al. A discourse-aware attention model for abstractive summarization of long documents*, 615–621 (2018).

[27] Dernoncourt, F. & Lee, J.-Y. *Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts*, 308–313 (2017).

[28] Cachola, I., Lo, K., Cohan, A. & Weld, D. S. *Tldr: Extreme summarization of scientific documents*, 4766–4777 (2020).

[29] Yasunaga, M. *et al.* *Scisummnet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks*, AAAI'19/IAAI'19/EAAI'19 (AAAI Press, 2019). URL https://doi.org/10.1609/aaai.v33i01.33017386.

[30] Tang, X., Cohan, A. & Gerstein, M. Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K. & Rumshisky, A. (eds) *Aligning factual consistency for clinical studies summarization through reinforcement learning.* (eds Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K. & Rumshisky, A.) *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 48–58 (Association for Computational Linguistics, Toronto, Canada, 2023). URL https://aclanthology.org/2023.clinicalnlp-1.7/.

[31] Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*, 74–81 (2004).

[32] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. *Bleu: a method for automatic evaluation of machine translation*, 311–318 (2002).

[33] Reimers, N. & Gurevych, I. *Sentence-bert: Sentence embeddings using siamese bert-networks*, 3982–3992 (2019).

[34] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[35] Kryściński, W., McCann, B., Xiong, C. & Socher, R. *Evaluating the factual consistency of abstractive text summarization*, 9332–9346 (2020).

[36] Durmus, E., He, H. & Diab, M. *Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization*, 5055–5070 (2020).

[37] Scialom, T. *et al.* *Questeval: Summarization asks for fact-based evaluation*, 6594–6604 (2021).

[38] Laban, P., Schnabel, T., Bennett, P. N. & Hearst, M. A. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* **10**, 163–177 (2022).

[39] Dubois, Y., Galambosi, B., Liang, P. & Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475* (2024).

[40] Li, W. *et al.* Grading scale impact on llm-as-a-judge: Human-llm alignment is highest on 0-5 grading scale. *arXiv preprint arXiv:2601.03444* (2026).

[41] Ye, J. *et al.* Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736* (2024).

[42] Gu, J. *et al.* A survey on llm-as-a-judge. *The Innovation* (2024).

[43] Zhu, L., Wang, X. & Wang, X. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631* (2023).

[44] OpenAI. Introducing gpt-5.4 (2026). URL https://openai.com/index/introducing-gpt-5-4/. OpenAI product announcement, accessed 2026-03-31.

[45] Google DeepMind. Gemini 3 flash: Model card (2025). URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Published December 2025; accessed 2026-03-31.

[46] González-Pereira, B., Guerrero-Bote, V. P. & Moya-Anegón, F. A new approach to the metric of journals' scientific prestige: The sjr indicator. *Journal of informetrics* **4**, 379–391 (2010).

[47] Google DeepMind. Gemini 3.1 pro: Model card (2026). URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. Published February 2026; accessed 2026-03-31.

[48] DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models (2025).

[49] MiniMax. Minimax m2.1: Significantly enhanced multi-language programming, built for real-world complex tasks (2025). URL https://www.minimax.io/news/minimax-m21. MiniMax news post, published 2025-12-23; accessed 2026-03-31.

[50] Gemma Team. Gemma 3 (2025). URL https://goo.gle/Gemma3Report.

[51] Grattafiori, A. *et al.* The llama 3 herd of models (2024). URL https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[52] Gemma Team. Gemma (2024). URL https://www.kaggle.com/m/3301.

[53] Qwen Team. Qwen2.5: A party of foundation models (2024). URL https://qwenlm.github.io/blog/qwen2.5/.

[54] Qwen Team. Qwen3 technical report (2025). URL https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[55] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models (2024). URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Meta AI blog post, published 2024-09-25; accessed 2026-03-31.

[56] Kimi Team *et al.* Kimi k2: Open agentic intelligence (2026). URL https://arxiv.org/abs/2507.20534. arXiv:2507.20534.

[57] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025). URL https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[58] Team, V. *et al.* Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025). URL https://arxiv.org/abs/2507.01006. arXiv:2507.01006.

[59] Qwen Team. Qwen2.5-vl (2025). URL https://qwen.ai/blog?id=qwen2.5-vl.