

MedRAGChecker: Claim-Level Verification for Biomedical Retrieval-Augmented Generation

Yuelyu Ji Min Gu Kwak Hang Zhang
Xizhi Wu Chenyu Li Yanshan Wang*
University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Biomedical retrieval-augmented generation (RAG) produces long-form answers that mix supported, under-evidenced, and contradicted claims, demanding fine-grained verification. We propose MEDRAGCHECKER, a claim-level diagnostic framework that combines textual natural language inference (NLI) with a biomedical knowledge-graph (KG) compatibility signal. Its asymmetric selective-gating rule preserves the textual prediction for most claims and applies KG adjustment only to confident textual entailments assigned low plausibility by the KG back-end. We distill GPT-4.1 claim decompositions and NLI pseudo-labels into compact student models and evaluate three biomedical generators on PubMedQA, MedQuAD, LiveQA, and MedRedQA. The student-ensemble results show that selective BioPortal gating increases macro-averaged faithfulness from 60.8% to 64.4% and reduces the contradiction rate from 19.9% to 15.0%. On 172 annotations of 86 claims whose decisions change after fusion, agreement with human evidence judgments increases from 63.4% to 69.8%. A comparison with DRKG shows that fusion behavior varies substantially with KG coverage, entity linking, score construction, and calibration; the comparison therefore characterizes back-end effects rather than isolating a causal effect of coverage.

1 Introduction

Large language models (LLMs) now pass standardized medical exams, yet their ungrounded generations still produce clinically harmful errors (Yu et al., 2025; Singhal et al., 2025; Vladika et al., 2024). Retrieval-augmented generation (RAG)

*Corresponding author: yanshan.wang@pitt.edu
Code, and diagnostics are available at: <https://github.com/JoyDajunSpaceCraft/MedicalRagChecker>

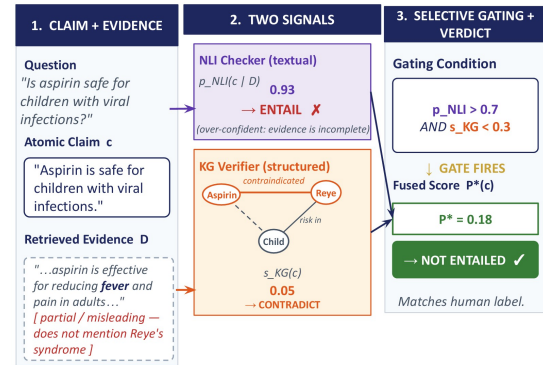


Figure 1: **Selective gating in MEDRAGCHECKER.** This schematic example illustrates the decision rule rather than a reported evaluation instance. For a claim assigned high textual entailment but low KG plausibility, selective gating lowers the fused support score and flags the claim as NOT ENTAILED. A low KG score is a back-end-specific compatibility signal, not proof that the claim is factually false. A real decision-flip example from the human evaluation is provided in Appendix G.

partly addresses this by conditioning answers on up-to-date medical literature (Achiam et al., 2023; Tu et al., 2024; Lewis et al., 2020), but noisy retrieval can still cause the model to anchor on misleading snippets (Ning et al., 2025). Long-form biomedical answers contain many distinct factual statements, and whole-answer scoring hides isolated but clinically important mistakes. This motivates *claim-level* evaluation: decomposing each answer into atomic factual claims (e.g., a single drug–effect or drug–contraindication assertion) and verifying each one independently (Min et al., 2023; Ru et al., 2024).

Standard claim-level verification frames each claim as a natural language inference (NLI) hypothesis against retrieved evidence. In biomedicine, however, textual entailment alone is insufficient: a claim can be linguistically supported by retrieved passages yet violate known drug–disease contraindications. Recent work therefore aug-

ments NLI with biomedical knowledge graphs (KGs) (Vladika et al., 2024; Ning et al., 2025), under the implicit assumption that adding a KG can only help. *This assumption is unsafe.* Figure 1 illustrates this failure mode using an aspirin safety claim. The textual checker assigns a high entailment probability because the retrieved evidence omits the association between aspirin use in children with viral infections and Reye’s syndrome. The KG verifier instead returns low support, activating the selective gating rule. This example motivates examining how KG back-end characteristics—including coverage, entity linking, score construction, and calibration—influence biomedical claim verification.

We propose MEDRAGCHECKER (Figure 2), a claim-level diagnostic framework that combines textual NLI with a KG compatibility signal under selective gating. We instantiate the KG module with DRKG and BioPortal, which differ in scope, entity linking, score construction, and calibration. Because these back-ends differ along multiple dimensions, their comparison does not isolate a causal effect of coverage. Instead, it demonstrates that KG back-end selection can substantially change fusion behavior and motivates a conservative rule that limits when the KG can alter a textual entailment decision.

Contributions. (i) We identify a failure mode in biomedical claim verification: indiscriminately combining textual entailment with back-end-specific KG scores can introduce spurious decision changes because KG coverage, entity linking, and calibration vary across back-ends. (ii) We introduce an asymmetric selective-gating rule that preserves the textual NLI prediction for most claims and applies KG adjustment only to high-confidence textual entailments with low KG plausibility. Claim extraction and student distillation are supporting components rather than methodological contributions. (iii) Across four datasets, three generators, and two KG back-ends, we characterize these back-end effects and evaluate KG-induced decision flips against blinded human evidence judgments.

2 Related Work

Biomedical QA. Despite strong performance on standardized exams (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019), long-form biomedical answers remain prone to hallucinations (Asgari et al., 2025), making claim-level validation necessary.

RAG Evaluation and Fact-Checking. Prior work formulates RAG factuality checking as an NLI-style claim-evidence problem, splitting generations into statements and verifying each against retrieved passages (Lewis et al., 2020; Fu et al., 2024). RAGAS (Es et al., 2024) reports answer-level evaluation metrics, while RAGChecker (Ru et al., 2024) decomposes responses into claims and aggregates claim-level errors into system-level diagnostics. MEDRAGCHECKER specializes this claim-level evaluation setting to biomedicine by adding a structured KG compatibility signal and safety-oriented relation diagnostics, then aggregating verified claims into answer-level summaries. HealthFC (Vladika et al., 2024) offers a strong clinical baseline but targets short, well-formed statements, while we target multi-sentence RAG answers.

Knowledge Graphs for QA. Biomedical KGs support QA reasoning (Himmelstein et al., 2017), but their incompleteness limits standalone use (Chen et al., 2024). We use KGs as external verifiers complementing text-based NLI, fusing both signals to improve robustness on safety-critical claims.

3 MedRAGChecker

Figure 2 overviews the pipeline. MEDRAGCHECKER combines a textual NLI signal from teacher-distilled student checkers with a KG compatibility signal instantiated with DRKG or BioPortal. It reports the component scores and a selectively fused support score for downstream diagnostics.

3.1 Task Definition and Outputs

We cast biomedical RAG evaluation as claim-centric verification. Given a RAG instance (q, D, a) with question q , generated answer a , and top- k retrieved passages $D = \{d_j\}_{j=1}^k$, MEDRAGCHECKER decomposes a into atomic claims $\mathcal{C} = \{c_1, \dots, c_n\}$ and assigns each claim (1) a discrete verdict $\hat{y}_i \in \{\text{ENTAIL}, \text{NEUTRAL}, \text{CONTRADICT}\}$ and (2) a support score $s(c_i) \in [0, 1]$. Two signals produce the verdict: a textual NLI probability $p_{\text{NLI}}(c_i)$ against retrieved evidence, and a KG plausibility score $s_{\text{KG}}(c_i)$ against a biomedical knowledge base. We study how KG back-end choice—including coverage, entity linking, score

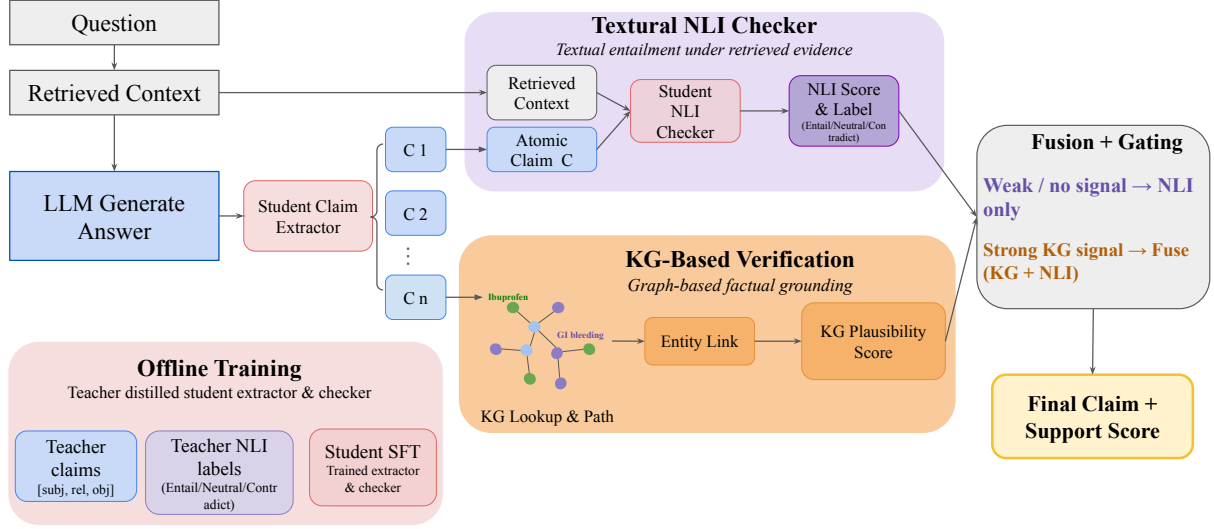


Figure 2: **Overview of MEDRAGCHECKER.** For each atomic claim extracted from the answer, a textual NLI checker (purple) and a KG-based compatibility module (orange) produce complementary signals. These are combined by a selective gating module: when textual entailment is high but KG plausibility is low, the KG score is fused with the NLI score; otherwise the NLI score is used alone. The student NLI checker and claim extractor are distilled from GPT-4.1 pseudo-labels offline (dashed arrows); at inference no proprietary LLM calls are made. We instantiate the KG module with DRKG and BioPortal, which differ in coverage, entity linking, and score construction (Section 4.5).

construction, and calibration—is associated with the quality of the final verdict.

In our implementation, the textual checker provides $p_{\text{NLI}}(c_i) \triangleq P(\text{ENTAIL} \mid c_i, D)$ from the ensemble distribution. When KG evidence is available, we fuse this textual signal with KG support to obtain the final support score (Section 3.5).

3.2 Claim Extraction

We use GPT-4.1 as a teacher to decompose each answer a into atomic claims, represented as subject–predicate–object (SPO) triples (e.g., ["Ibuprofen", "may cause", "GI bleeding"]), in the spirit of FActScore-style atomic decomposition (Min et al., 2023). The triple form aligns with downstream KG entity linking (§3.4). To avoid teacher calls at inference time, we distill a biomedical LLM (e.g., Med42-Llama3-8B) via supervised fine-tuning on $(q, a) \rightarrow c_1, \dots, c_n$ with teacher claim lists as targets. Because claim extraction is not our primary methodological contribution, we treat it as a front-end component and evaluate its errors separately. Its outputs remain teacher-dependent because the student extractor is trained on GPT-4.1 decompositions. Extraction quality (BERTScore vs. teacher claims, human Likert ratings) is validated in Section 4.4; training details are in Appendix P; qualitative error analysis

is in Appendix I.

3.3 Textual NLI Verification

For each extracted claim c_i , a checker predicts $y_i \in \{\text{ENTAIL}, \text{NEUTRAL}, \text{CONTRADICT}\}$ with evidence D as premise and the claim as hypothesis, producing $p_{\text{NLI}}(c_i) \triangleq P(\text{ENTAIL} \mid c_i, D)$. We distill GPT-4.1 verdicts into Meditron3-8B, MedQwen2-7B, and Med42-Llama3-8B via supervised fine-tuning (SFT), then combine them in an F1-weighted ensemble so minority-class specialists contribute more to CONTRADICT decisions. Unless otherwise stated, all downstream diagnostics use this ensemble; details are in Appendix P.

3.4 KG-Based Verification

In biomedical settings, a claim may receive a high textual entailment score even when retrieval omits clinically important context. We therefore augment the textual NLI signal with structured biomedical knowledge, while retaining the component scores as separate diagnostics.

Our KG verification module is designed as a generic interface with three stages: (1) *entity linking* from the extracted SPO triple to KG entities; (2) *plausibility scoring* over linked entities; and (3) *signal fusion* with the textual NLI score (Section 3.5). Given an extracted SPO claim, the KG

operations require no proprietary LLM calls. This provides a non-LLM external signal, but does not make the end-to-end pipeline teacher-independent because the extractor and NLI checker are distilled from GPT-4.1 (cf. Limitations). Different KG back-ends can be plugged into this interface; we instantiate the module with two biomedical KGs in our experiments (Section 4.2). For both back-ends, KG plausibility is computed only over the linked subject–object pair in the extracted SPO claim. We do not construct or reason over multi-hop KG subgraphs. Additional nodes shown in Figure 1, such as Reye’s syndrome, provide background context only and do not directly enter the KG score.

3.4.1 Entity Linking

Unlike general-purpose KGs (e.g., Wikidata (Vrandečić and Krötzsch, 2014), ConceptNet (Liu and Singh, 2004)) that ship with high-quality entity embeddings, the biomedical KGs we use either do not provide pretrained entity embeddings aligned with our claim vocabulary (BioPortal) or provide only relation-specific TransE embeddings without natural-language surface forms (DRKG). We therefore link entities via text-based matching enhanced with biomedical entity representations (SapBERT) where applicable.

For each claim c , the claim extractor returns an SPO triple (s, r, o) . During training these targets originate from GPT-4.1; during inference they are produced by the distilled extractor. We then link s and o to the target KG.

For **DRKG**, we perform string-based matching against DRKG entity names and map canonical claim relations (e.g., TREATS, CAUSES) to DRKG edge types, computing a normalized string similarity score $s_{\text{text}}(c) \in [0, 1]$.

For **BioPortal**, we query the BioPortal API across entity-type- appropriate ontologies, scoring each candidate entity by

$$s_{\text{link}} = 0.3 \cdot s_{\text{text}} + 0.5 \cdot s_{\text{SapBERT}} + 0.2 \cdot s_{\text{syn}}, \quad (1)$$

where s_{text} is Jaccard string similarity, s_{SapBERT} is the cosine similarity between SapBERT (Liu et al., 2021) biomedical entity embeddings, and s_{syn} is synonym-set match. We retain the top-scoring entity per mention with $s_{\text{link}} \geq 0.5$; otherwise the claim is treated as KG-uncovered and falls back to NLI-only.

3.4.2 KG Plausibility Scoring

Given linked entities, we compute a claim-level plausibility score $s_{\text{KG}}(c) \in [0, 1]$.

The score is a back-end-specific compatibility signal rather than a calibrated probability that the claim is true. In particular, a low score does not by itself establish a factual contradiction under the open-world assumption of biomedical KGs.

DRKG (TransE). For each aligned triple (h, r, t) , we compute $p_{\text{KGE}}(h, r, t) = \sigma(-\|\mathbf{e}_h + \mathbf{r}_r - \mathbf{e}_t\|_2)$ using pre-trained translation-based KG embeddings (TransE) (Bordes et al., 2013), taking the maximum over candidate triples. We combine this with text alignment:

$$s_{\text{KG}}^{\text{DRKG}}(c) = (1-\alpha) \cdot p_{\text{KGE}}(c) + \alpha \cdot s_{\text{text}}(c), \quad (2)$$

with α tuned on dev.

BioPortal (ontology-based). For each linked entity pair, we compute the maximum of three signals: (i) *direct property lookup* (score 0.9, e.g., indication, contraindication); (ii) *hierarchical proximity* via shared ontology ancestors (0.2–0.7); and (iii) *cross-ontology mapping* through LOOM links (0.8):

$$s_{\text{KG}}^{\text{Bio}}(c) = \max(s_{\text{direct}}, s_{\text{hier}}, s_{\text{loom}}). \quad (3)$$

Scoring details are in Appendix F.

3.5 Signal Fusion with Selective Gating

3.5.1 Logit-Space Fusion

For claims linked to the KG, we combine NLI and KG signals in logit space:

$$P^*(c_i) = \sigma\left(\beta \cdot \text{logit}(p_{\text{NLI}}(c_i)) + (1-\beta) \cdot \text{logit}(s_{\text{KG}}(c_i))\right), \quad (4)$$

where β controls the relative contribution of textual and KG evidence. Because the two scores are produced by different mechanisms, their fusion remains sensitive to calibration. We therefore estimate KG calibration statistics on the development split, fix them before test evaluation, and report calibrated and uncalibrated sensitivity analyses in Appendix L. For claims that cannot be linked to the KG, we fall back to $P^*(c_i) = p_{\text{NLI}}(c_i)$.

3.5.2 Selective Gating

A symmetric application of Eq. 4 can allow a poorly calibrated KG score to alter every linked claim. We

instead restrict KG adjustment to a narrow disagreement regime; with shorthand $\ell_{\text{NLI}} \triangleq \text{logit}(p_{\text{NLI}})$, $\ell_{\text{KG}} \triangleq \text{logit}(s_{\text{KG}})$:

$$P^*(c_i) = \begin{cases} \sigma(\beta \ell_{\text{NLI}} + (1-\beta) \ell_{\text{KG}}), \\ \text{if } s_{\text{KG}} < 0.3 \text{ and } p_{\text{NLI}} > 0.7 \\ p_{\text{NLI}}(c_i), & \text{otherwise.} \end{cases} \quad (5)$$

Gating fires only when NLI is confident of entailment but the KG back-end assigns low plausibility. The rule is intentionally asymmetric. A low KG score may reduce an over-confident textual entailment, but a high KG score is not allowed to overturn a low textual entailment probability. In the symmetric low- p_{NLI} , high- s_{KG} case, the method retains the NLI score. This conservative design prevents high KG plausibility from overriding textual contradiction or negation cues in safety-critical claims. We tune (α, β, τ) on the development split (Appendix L).

3.6 Metrics

From $P^*(c_i)$ and the predicted labels, we derive two core answer-level metrics: *Faithfulness* $\text{Faith}(a) = \frac{1}{n} \sum_i \mathbb{I}[\hat{y}_i = \text{ENTAIL}]$ and *Hallucination rate* $\text{Halluc}(a) = \frac{1}{n} \sum_i \mathbb{I}[\hat{y}_i = \text{CONTRADICT}]$. We additionally compute (i) claim extraction quality (span-F1 and BERTScore (Zhang et al., 2020)), (ii) retrieval diagnostics (ClaimRec, Ctx-Prec, answer-level ClaimF1), (iii) a parametric SelfKnow score (NLI-only, empty context), and (iv) a SafetyErr rate—the fraction of KG-aligned safety-critical claims (drug–disease, drug–adverse-event) labeled CONTRADICT, isolating clinically dangerous errors from benign inaccuracies. Formal definitions of all metrics are in Appendix A. A *decision-flip* claim is one whose verdict (Entail vs. not-Entail) differs between NLI-only and the Fused setting; flips are the loci where KG augmentation actually changes evaluation outcomes.

4 Experiments

4.1 Experimental Setup

Data and RAG setup. We evaluated on four biomedical QA benchmarks: PubMedQA (Jin et al., 2019), MedQuAD (Ben Abacha and Demner-Fushman, 2019), LiveQA (Yang et al., 2017), and MedRedQA (Nguyen et al., 2023). We use a PubMed-based RAG pipeline (Appendix O) with fixed retrieval settings across runs and compare

three biomedical generators: Meditron3-8B (Sallinen et al., 2025), Med-Qwen2-7B (Bai et al., 2023), and Med42-Llama3-8B (Christophe et al., 2024). PMC-LLaMA-13B (Wu et al., 2024) is excluded from the main comparison due to degenerate extraction behavior (only 1.19 claims per answer vs. 3–5 for the other generators, and BERT-F1=45.7 against teacher claims vs. 78–84), and is reported separately in Appendix C.

Checker training and evaluation protocol. We use GPT-4.1 to provide claims and NLI pseudo-labels on train/dev, then distill (i) a student claim extractor and (ii) student NLI checkers. Unless stated otherwise, we use the F1-weighted ensemble (Section 3.3) for downstream diagnostics. We report teacher-checker, single-student, and student-ensemble results. We split the teacher-labeled data into train/dev/test (80/10/10) by question before constructing claim-level instances, preventing claims derived from the same question from crossing splits, and fine-tuned open-source students with low-rank adaptation (LoRA) (Hu et al., 2022); training details are in Appendix P. For KG-enhanced verification, we computed a KG support score from each backend (DRKG or BioPortal, Section 3.4) and fused it with textual NLI via Eq. 4 or Eq. 5; (α, β, τ) were tuned on dev and then fixed for all test results. Unless noted, all test results use student models only, with no GPT-4.1 calls at inference.

Marginal inference cost. The distilled extractor and student ensemble process approximately 180 claims/min on a single A100-80GB in our batched implementation. This measurement covers local model inference; retrieval and uncached BioPortal API latency are excluded and reported separately in the released evaluation logs. BioPortal does not charge a per-query fee, but requires an API key, so we cache ontology responses. These figures describe marginal inference and exclude the one-time costs of GPT-4.1 pseudo-label generation and student training. We therefore do not claim a fixed monetary cost ratio, which depends on token counts, hardware pricing, batching, and cache hit rate.

4.2 KG Back-end Instantiations

To study how KG choice affects verification, we instantiate the KG module with two back-ends that expose a common input–output interface but use different entity-linking and plausibility-scoring

Table 1: Distilled claim extractor and checker fidelity to the GPT-4.1 teacher. **Panel A:** BERTScore F1 between student-extracted and teacher-extracted claims, and average number of claims per answer. **Panel B:** Checker accuracy and per-class F1 on the 3-way NLI task. F1-E, F1-N, and F1-C denote per-class F1 for the ENTAIL, NEUTRAL, and CONTRADICT labels respectively. **Bold** marks the column-wise maximum (including the ensemble row where reported). **Green** highlights the best value *among single students only*; **red** highlights the worst. The ensemble row is excluded from the green/red comparison.

Generator	Panel A: Extractor		Panel B: Checker (SFT)				
	BERT-F1 $\uparrow$	#Claims	Acc (%)	Mac-F1 (%)	F1-E (%)	F1-N (%)	F1-C (%)
Med-Qwen2-7B	78.3	3.25	83.7	55.0	40.1	94.9	42.4
Med42-Llama3-8B	83.6	4.70	85.6	67.3	50.3	94.5	27.3
Meditron3-8B	82.1	5.28	81.2	49.7	46.3	92.9	23.4
Ensemble (F1-wt)			87.4	60.5	52.8	95.0	41.3

Table 2: **Verification behavior across KG back-ends.** Results are macro-averaged across four datasets. DRKG and BioPortal differ in coverage, entity linking, score construction, and calibration; therefore, differences between their rows should not be attributed to coverage alone. “vanilla” applies Eq. 4 to all KG-linked claims, whereas “gated” uses Eq. 5. Green and red indicate improvement and degradation relative to NLI-only, respectively. Bold marks the best value within each generator block and checker setting.

Generator	KG Setting	Teacher-checker (GPT-4.1)			Student-checker (SFT ensemble)		
		Faith. $\uparrow$	Halluc. $\downarrow$	SafetyErr $\downarrow$	Faith. $\uparrow$	Halluc. $\downarrow$	SafetyErr $\downarrow$
Med-Qwen2-7B	NLI-only	68.1	21.8	29.9	62.3	20.7	19.6
	+ DRKG	75.6	18.6	24.4	68.6	10.4	12.6
	+ BioPortal (vanilla)	72.5	23.4	25.7	68.8	20.8	18.3
	+ BioPortal (gated)	72.1	22.3	25.4	69.6	9.1	16.8
Med42-Llama3-8B	NLI-only	70.2	18.1	29.8	67.3	24.3	28.9
	+ DRKG	77.0	23.0	17.5	58.4	25.7	19.6
	+ BioPortal (vanilla)	73.3	16.7	16.8	67.8	24.8	24.4
	+ BioPortal (gated)	73.2	14.3	16.4	69.1	22.4	23.7
Meditron3-8B	NLI-only	58.2	20.6	41.8	52.7	14.6	29.0
	+ DRKG	66.8	33.2	23.7	59.6	27.3	27.2
	+ BioPortal (vanilla)	67.9	23.6	37.9	53.4	15.7	26.7
	+ BioPortal (gated)	69.3	23.5	36.7	54.6	13.5	25.7

mechanisms. They also differ in scope and edge-attribute richness:

DRKG (Ioannidis et al., 2020) is a drug-repurposing-focused biomedical KG containing $\approx 97K$ entities and $\approx 5.8M$ edges, primarily covering drug–gene–disease relations. Its edges carry minimal attribute information beyond the relation type, and we score plausibility using pre-trained TransE embeddings. On our benchmarks, DRKG links both entities for 62.7% of extracted claim pairs (averaged over the three main generators; PMC-LLaMA-13B excluded for consistency with §4.1).

BioPortal (Salvadores et al., 2013) is a multi-ontology hub integrating biomedical ontologies (SNOMED-CT, MeSH, NCI, DOID, RxNorm)

with cross-ontology links, accessed via the public API in early 2026. Its edges carry richer ontology-level attributes (direct property relations such as indication/contraindication, hierarchical ancestors, LOOM mappings), which support ontology-based scoring without requiring trained embeddings. BioPortal links both entities for 94.2% of extracted claim pairs (a 31.5pp increase over DRKG). We use this entity-pair definition for all main-text coverage results; node-level coverage is reported separately in Appendix J.

The two back-ends differ both in *coverage* over our claim distribution and in the *type of edge attributes* available for scoring. Section 4.5 analyzes how these differences propagate to downstream verification quality.

4.3 Research Questions

We organize experiments around three questions: how KG back-end characteristics are associated with verification behavior (Sec. 4.5), whether selective gating mitigates back-end-specific fusion failures (Sec. 4.6), and whether KG-induced decision flips better align with blinded human evidence judgments (Sec. 4.7). Claim extraction and checker distillation are supporting components evaluated separately.

4.4 Distillation Quality

Before presenting our main results, we briefly validate that the distilled student checkers and extractors are faithful enough to serve as the front-end of our pipeline. A full analysis is in Appendix C; here we summarize the key findings.

Extractor. Student extractors achieve BERTScore F1 of 0.78–0.84 against teacher claims (Med42-Llama3-8B: 83.6; Meditron3-8B: 82.1; Med-Qwen2-7B: 78.3), producing 3–5 atomic claims per answer that closely match the teacher’s decomposition. Qualitative analysis (Appendix I) further shows that hallucinated claims are rare (6%) and over-splitting is the dominant but benign failure mode.

Checker ensemble. After SFT, the F1-weighted student ensemble achieves the highest accuracy (87.4%) and the highest ENTAIL and NEUTRAL F1 against GPT-4.1 pseudo-labels (Table 1). Med42-Llama3-8B achieves the highest macro-F1 (67.3%), while Med-Qwen2-7B achieves the highest CONTRADICT F1 (42.4%). We therefore report ensemble and single-backbone results without claiming uniform ensemble superiority.

4.5 RQ1: How KG Back-end Characteristics Affect Verification

We compare NLI-only, NLI+DRKG, BioPortal vanilla fusion, and BioPortal selective gating under identical student checkers. Table 2 reports macro-averages across four datasets for each generator. Unless otherwise stated, summary values below are simple macro-averages over the three student-ensemble generator rows.

Main finding: KG back-end choice changes verification behavior. For the student ensemble, macro-averaged faithfulness is 60.8% with NLI-only, 62.2% with DRKG, and 64.4% with selectively gated BioPortal. The corresponding contra-

diction rates are 19.9%, 21.1%, and 15.0%. DRKG and BioPortal differ in coverage, entity linking, score construction, and calibration. Consequently, these differences should be interpreted as effects associated with the complete KG back-end configuration rather than as a causal effect of coverage alone.

Per-generator consistency. Table 2 shows heterogeneous effects across the three generators under the student-checker setting. For Meditron3-8B, DRKG increases the contradiction rate from 14.6% to 27.3%, while selective BioPortal gating reduces it to 13.5%. SafetyErr decreases under both back-ends in the reported student rows. Because completely uncovered claims fall back to NLI-only, observed fusion failures arise from linked claims with noisy mappings or poorly calibrated compatibility scores, not from uncovered claims.

Per-dataset coverage effect. DRKG node-level and entity-pair coverage varies across datasets and generators (Appendix J, Table 11). This pattern is consistent with coverage contributing to back-end differences, but it does not separate coverage from linking and scoring effects. We therefore report the aggregate BioPortal–DRKG entity-pair comparison in the main text and avoid a causal per-dataset interpretation. Per-generator coverage and KG-subset behavior are in Table 11.

Human validation on decision-flip. On 172 annotator–claim judgments from 86 decision-flip claims, agreement is 69.8% for BioPortal-fused predictions and 63.4% for NLI-only predictions. Macro-F1 increases from 59.2 to 64.7. This targeted analysis provides descriptive evidence on the sampled flip subset; it does not establish a general improvement over all claims. Full protocol is in Appendix N.

4.6 RQ2: Selective Gating Mitigates Back-end-Specific Fusion Failures

The main result in RQ1 uses selective gating (Eq. 5) in the BioPortal configuration. Here we ablate gating to isolate its contribution.

Vanilla fusion baseline. To separate the effect of the gating rule from the effect of KG coverage, we compare selective gating (Eq. 5) against a *vanilla fusion* baseline that applies Eq. 4 to all KG-linked claims without the gating condition. Both variants share the same KG and fusion settings (β, τ), so any difference in Table 2 between Bio-

Portal (vanilla) and BioPortal (gated) rows isolates the gating rule itself.

Gating trigger rate. The gating condition ($s_{KG} < 0.3$ and $p_{NLI} > 0.7$) fires on 12.3% of KG-linked claims on average across our four datasets in the reported evaluation. For the remaining 87.7% of KG-linked claims, at least one gating condition is not satisfied, so the method retains p_{NLI} .

Gating ablation on the same KG. Comparing BioPortal (vanilla) and BioPortal (gated) in Table 2 isolates gating’s effect under fixed coverage: across all three generators, gating reduces the contradiction rate in all six generator–checker comparisons, with effect sizes ranging from 0.1 to 11.7 percentage points, while faithfulness changes are comparatively small. This same-back-end comparison isolates the gating rule from coverage. We restrict this ablation to BioPortal; applying the same fixed thresholds to differently scaled DRKG scores would not constitute a controlled calibration comparison.

Relation-type analysis. Appendix G illustrates that decision flips occur in both safety-critical drug relations and broader lifestyle claims. Because the relation strata are small, we treat this analysis as descriptive and do not infer that a flip is a correction unless it is supported by a human evidence judgment.

4.7 RQ3: Human Alignment on KG-Induced Decision Flips

Our RQ3 reports two targeted human studies: a claim-extraction sanity check on outputs sampled from 100 questions, and a blinded evaluation of 86 claims whose decisions change after BioPortal fusion.

Human ratings of extracted claim quality. Med-Qwen2-7B: 4.97 ± 0.18 ; Med42-Llama3-8B: 4.73 ± 0.99 ; Meditron3-8B: 4.62 ± 0.18 (1–5 Likert; two annotators, 100 questions; quadratic-weighted $\kappa = 0.67$ for claim-extraction quality). Near-ceiling scores confirm the distilled extractor produces faithful atomic decompositions.

Decision-flip agreement. Across 172 annotator–claim judgments, BioPortal+gating raises agreement from 63.4% to 69.8% (+6.4pp). Given the targeted sample and modest size, we interpret the difference descriptively.

Table 3: Comparison with general-purpose RAG evaluators. All numbers use Med42-Llama3-8B as a representative generator to enable controlled comparison on the same outputs. – indicates the metric is not supported by that tool. The scores are outputs of different evaluator definitions and should not be read as directly interchangeable accuracy estimates. MEDRAGCHECKER produces scores in a comparable range while additionally providing claim-level diagnostics, biomedical KG verification, and safety-critical error detection.

Dataset	Metric	MedRAG-Checker	RAGAS	RAG-Checker
PubMedQA	Faith. $\uparrow$	88.9	89.5	87.2
	Halluc. $\downarrow$	4.7	–	5.1
LiveQA	Faith. $\uparrow$	82.5	74.9	70.8
	Halluc. $\downarrow$	8.6	–	19.5
MedQuAD	Faith. $\uparrow$	87.6	84.7	83.9
	Halluc. $\downarrow$	5.5	–	6.8
MedRedQA	Faith. $\uparrow$	82.4	82.1	81.5
	Halluc. $\downarrow$	6.4	–	8.2

4.8 Comparison with Existing RAG Evaluators

To position MEDRAGCHECKER against general-purpose RAG evaluation tools, we compare with RAGAS (Es et al., 2024) and RAGChecker (Ru et al., 2024) on answer-level faithfulness using the same retrieval pipeline and generator outputs (Table 3). We use Med42-Llama3-8B as a representative generator (best macro-F1 as a checker; comparable claim quality across generators, see §4.7).

All three frameworks produce faithfulness scores in a similar range, although their decompositions and evaluator definitions differ. RAGChecker already operates at claim level; MEDRAGCHECKER instead differs through its biomedical KG compatibility signal, safety-oriented relation diagnostics, and inference without proprietary LLM calls. The BioPortal configuration still requires API access or cached ontology responses.

5 Conclusion

MEDRAGCHECKER provides claim-level diagnostics for biomedical RAG by combining distilled textual NLI checkers with structured KG signals under selective gating. It achieves faithfulness comparable to general-purpose evaluators while additionally identifying safety-critical errors. Our comparison of DRKG and BioPortal shows that KG back-end choice is associated with substantial differences in fusion behavior. Because the back-ends

differ in coverage, linking, and score construction, the results do not isolate coverage as the sole cause. On the targeted decision-flip subset, selective BioPortal fusion has higher descriptive agreement with blinded human evidence judgments than NLI-only verification.

Limitations

Teacher supervision as pseudo-ground truth. Our distillation and evaluation rely on teacher labels for claim decomposition and NLI judgments, which may contain biases or systematic errors, especially for rare biomedical conditions or ambiguous questions. Our teacher-sensitivity analysis (Appendix E, Table 9) confirms that generator rankings are stable under both GPT-4.1 and GPT-4o teachers across research-style datasets, though absolute values on consumer-health data (MedRedQA, $\kappa = 0.13$ between teachers) remain teacher-conditional.

Checker calibration and class imbalance. Contradiction is typically a minority class. Even with ensembling, performance varies across datasets, and subtle contradictions can be under-detected. MedRedQA exhibits low inter-teacher agreement (Appendix Table 9); we therefore interpret absolute Faith/Halluc values on MedRedQA with caution and focus on within-teacher relative comparisons. **Limited KG back-end comparison.** We evaluate only DRKG and BioPortal. These back-ends differ not only in coverage but also in entity linking, relation representation, score construction, and calibration. Consequently, our experiments characterize differences associated with back-end choice but do not isolate a causal effect of coverage alone. Both back-ends may also miss rare entities, complex paraphrases, and relations that cannot be represented by a single subject-object pair. Evaluating additional resources such as UMLS and Hetionet under a standardized scoring interface is important future work.

Retrieval-conditional evaluation. All verifier variants use the same retrieved passages, which controls retrieval across method comparisons. However, absolute diagnostic values remain conditional on retrieval quality. Under our three-way labeling scheme, retrieval failures and missing evidence are more likely to increase NEUTRAL than CONTRADICT. We do not evaluate robustness across alternative retrievers or retrieval depths.

Circularity of teacher supervision. Although

the KG operations do not call GPT-4.1, claim extraction and the primary NLI signal remain derived from GPT-4.1 pseudo-labels, creating a partially circular evaluation loop. A fully teacher-free pipeline would require human-annotated claims at scale; we partially mitigate this through human validation (Section 4.7) and teacher-robustness analysis (Appendix E).

Domain generalization. Our methodology (claim extraction, NLI + KG verification, fusion) transfers to other technical domains such as legal and financial QA, but requires a domain-specific KG and re-distillation of student models. Cross-domain evaluation is left for future work.

Scale of human evaluation. Our claim-extraction study samples outputs from 100 questions and three generators. Inter-annotator agreement is substantial for claim-extraction quality (quadratic-weighted $\kappa = 0.67$). The decision-flip study contains 86 claims and 172 independent judgments from two graduate annotators; it should therefore be interpreted as descriptive evidence. Larger-scale validation with clinician annotators remains future work.

Ethical Considerations

MEDRAGCHECKER is designed to *detect* unsupported or contradictory claims in biomedical RAG outputs and to surface safety-critical error patterns. Nevertheless, automatic verification can be wrong; false negatives may miss harmful claims and false positives may over-warn. We recommend using MEDRAGCHECKER as a screening and debugging aid, with human oversight for high-stakes settings.

Our experiments use publicly available datasets and retrieved biomedical literature.

For our human evaluation, annotators were informed of the study purpose, compensated appropriately, and instructed not to provide medical advice. We do not process private patient data. We will release prompts, model checkpoints (where licenses allow), and evaluation code to support transparency and reproducibility.

Acknowledgments

Use of Generative AI. GPT-4.1 is used as a methodological component of this work, serving as the teacher model for claim extraction (Section 3.2) and NLI verification (Section 3.3) in our distillation pipeline, and as the basis for the teacher-sensitivity comparison against GPT-4o (Appendix E). Given

extracted claims, the KG verification module operates without GPT-4.1 calls. At inference time, we use only distilled student models and the selected KG back-end; the full mapping of models to pipeline stages is documented in Section 3.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. [A question-entailment approach to question answering](#). *BMC Bioinform.*, 20(1):511:1–511:23.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Weijian Chen, Yixin Cao, Fuli Feng, Xiangnan He, and Yongdong Zhang. 2024. HoGRN: Explainable sparse knowledge graph completion via high-order graph reasoning network. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8462–8475.
- Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. [Med42-v2: A suite of clinical llms](#).
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.
- Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xi-angxiang Zeng, and George Karypis. 2020. Drkg - drug repurposing knowledge graph for covid-19. <https://github.com/gnn4dr/DRKG/>.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4228–4238.

- Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. [MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.
- Yingpeng Ning, Yuanyuan Sun, Ling Luo, Yanhua Wang, Yuchen Pan, and Hongfei Lin. 2025. Medtrustrag: Evidence verification and trust alignment for biomedical question answering. *arXiv preprint arXiv:2510.14400*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, and 1 others. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.
- Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, and 1 others. 2025. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.
- Manuel Salvadores, Paul R Alexander, Mark A Musen, and Natalya F Noy. 2013. Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic web*, 4(3):277–284.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):A10a2300138.
- Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.
- Yuan Yang, Jingcheng Yu, Ye Hu, Xiaoyao Xu, and Eric Nyberg. 2017. Cmu livemedqa at trec 2017 liveqa: A consumer health question answering system. *arXiv preprint arXiv:1711.05789*.
- Erlan Yu, Xuehong Chu, Wanwan Zhang, Xiangbin Meng, Yaodong Yang, Xunming Ji, and Chuanjie Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

A Full Metric Definitions

Notation. For an example (q, D, a) , the extractor outputs claims $\mathcal{C} = \{c_i\}_{i=1}^n$. The checker predicts $\hat{y}_i \in \{\text{ENTAIL}, \text{NEUTRAL}, \text{CONTRADICT}\}$ and a fused support score $P^*(c_i) \in [0, 1]$ (Eq. 4). In the fused setting, we replace the textual NLI ENTAIL logit with the fused logit from $P^*(c_i)$, keep NEUTRAL/CONTRADICT logits unchanged, renormalize via softmax, and predict $\hat{y}_i = \arg \max_y p(y | c_i, D)$. Binary supported/unsupported uses threshold τ (tuned on dev): $\mathbb{I}_{\text{sup}}(c_i) = \mathbb{I}[P^*(c_i) \geq \tau]$.

Claim extraction quality. Against teacher claims (GPT-4.1), we compute soft span-level P/R/F1 via token-overlap matching (Appendix H) and report average claims per answer.

Retrieval diagnostics.

$$\text{ClaimRec} = \frac{1}{|\mathcal{C}^{\text{ref}}|} \sum_{c \in \mathcal{C}^{\text{ref}}} \mathbb{I}_{\text{sup}}(c), \quad (6)$$

where $\mathcal{C}^{\text{ref}}$ are reference claims (reported only when references are available).

$$\text{CtxPrec} = \frac{1}{k} \sum_{j=1}^k \mathbb{I} \left[\exists i : d_j \text{ cited for } c_i \wedge P^*(c_i) \geq \tau \right]. \quad (7)$$

ClaimF1. We match generated claims $\mathcal{C}^{\text{gen}}$ to reference claims $\mathcal{C}^{\text{ref}}$ by token-overlap, count a match as correct if supported ($P^* \geq \tau$), and report the harmonic mean of Precision and Recall.

Self-knowledge. We disable KG fusion and re-run NLI with empty context $\emptyset$: $\text{SelfKnow} = \frac{1}{n} \sum_i \mathbb{I}[p_{\text{NLI}}(c_i | \emptyset) \geq \tau_{\text{NLI}}]$.

Safety-critical error rate. $\text{SafetyErr} = \frac{1}{|\mathcal{C}_{\text{safety}}|} \sum_{c \in \mathcal{C}_{\text{safety}}} \mathbb{I}[\hat{y}(c) = \text{CONTRADICT}]$, where $\mathcal{C}_{\text{safety}}$ are claims mapped to safety-critical KG relations.

B Dataset

This section summarizes the four biomedical QA benchmarks used in our experiments. Table 4 reports length and retrieval-context statistics, and Table 5 provides a qualitative overview of question style and evidence source.

Table 4: Dataset statistics for MedRAGChecker RAG inputs. $|q|$ and $|a|$ denote tokenized question and (ground-truth) answer lengths (whitespace tokens). $|d|$ denotes retrieved document length; median/max refer to median and maximum values across all retrieved contexts.

Dataset	#Q	Median $ q $	Max $ q $	Median $ a $	Max $ a $	Median #Doc	Max #Doc	Median/Max $ d $
MedRedQA	799	167	3732	56	1187	8	8	123 / 187
TREC LiveQA Medical	104	12	66	88	486	8	8	205 / 972
MedQuAD	1000	9	18	86	731	10	10	123 / 203
PubMedQA	1000	15	43	35	263	8	8	239 / 735

Table 5: Qualitative overview of the biomedical QA datasets used with MedRAGChecker, including domain, answer type, and typical question/evidence style.

Dataset	Domain	Answer type	Question style / evidence
MedRedQA	Consumer health (Reddit r/AskDocs)	Free-text answers by verified clinicians	Medium-long layperson questions covering diverse symptoms and concerns; thread content and PubMed-style references.
TREC LiveQA Medical	Consumer health questions to NLM	Free-text answers with quality judgments	Long, noisy real-world user questions; reference answers plus retrieved passages with human relevance labels.
MedQuAD	NIH consumer health websites	Free-text factoid / definition QA	Short-medium consumer questions about diseases, drugs, and procedures; long snippets from NIH pages with source URLs.
PubMedQA	Biomedical research (PubMed abstracts)	Yes/No/Maybe + long-answer rationale	Short research questions derived from article titles; abstract body used as context, conclusion paragraph as gold answer.

C Full Diagnostic Tables

Per-generator end-to-end diagnostics are in Table 6; teacher-checker results across both GPT-4.1 and GPT-4o are in Table 7; per-claim NLI label distributions are in Table 8.

Table 6: Full end-to-end MEDRAGCHECKER diagnostics across all datasets and generators (teacher = GPT-4.1). **Bold** marks the best value per dataset and column.

Dataset	Generator	Faith. $\uparrow$	Halluc. $\downarrow$	CtxPrec $\uparrow$	SelfKnow. $\uparrow$	SafetyErr $\downarrow$
PubMedQA	Med-Qwen2-7B	83.6	7.4	41.8	28.2	7.0
	Med42-Llama3-8B	88.9	4.7	41.4	32.7	6.5
	Meditron3-8B	69.2	3.3	53.2	30.4	5.9
	PMC-LLaMA-13B	57.9	11.7	35.6	21.6	12.4
MedQuAD	Med-Qwen2-7B	85.3	6.0	42.4	32.4	8.5
	Med42-Llama3-8B	87.6	5.5	46.3	31.2	7.8
	Meditron3-8B	73.1	7.6	49.4	24.0	8.2
	PMC-LLaMA-13B	63.2	8.7	44.6	20.0	9.7
LiveQA	Med-Qwen2-7B	72.4	11.3	43.2	23.1	7.6
	Med42-Llama3-8B	82.5	8.6	42.7	28.6	6.2
	Meditron3-8B	72.3	11.3	48.6	31.4	9.5
	PMC-LLaMA-13B	55.7	12.8	38.1	22.6	10.4
MedRedQA	Med-Qwen2-7B	84.2	7.3	41.7	26.4	7.6
	Med42-Llama3-8B	82.4	6.4	42.7	27.2	6.7
	Meditron3-8B	71.2	8.3	44.3	28.0	9.3
	PMC-LLaMA-13B	63.4	9.5	41.3	22.5	12.7

Table 7: Teacher checker results (teacher = GPT-4.1 vs GPT-4o). **Bold** marks the best value per dataset, teacher, and column (excluding degenerate rows where Faith. or ClaimF1 collapses to 0).

Dataset	Generator	Teacher = GPT-4.1					Teacher = GPT-4o				
		ClaimF1	ClaimRec	CtxPrec	Faith.	Halluc.	ClaimF1	ClaimRec	CtxPrec	Faith.	Halluc.
PubMedQA	Med-Qwen2-7B	26.7	97.6	49.9	88.6	7.4	23.2	96.9	41.8	83.6	7.4
	Med42-Llama3-8B	27.6	97.9	49.8	92.4	5.2	21.1	99.0	41.4	88.9	4.7
	Meditron3-8B	23.5	97.2	32.9	69.2	3.3	21.0	97.5	41.9	63.7	1.0
	PMC-LLaMA-13B	14.6	98.3	50.0	58.0	29.6	14.7	97.4	50.7	60.5	29.5
MedQuAD	Med-Qwen2-7B	39.3	65.7	51.4	88.1	9.4	30.1	62.3	47.0	82.9	6.2
	Med42-Llama3-8B	43.4	61.5	51.4	86.1	6.4	39.5	62.8	46.0	80.2	8.5
	Meditron3-8B	31.8	63.3	52.8	83.1	4.9	30.1	63.9	49.4	73.1	7.6
	PMC-LLaMA-13B	0.0	63.5	51.9	0.0	5.2	0.0	61.3	47.2	4.0	6.0
LiveQA	Med-Qwen2-7B	6.3	26.8	43.1	72.4	18.3	5.3	23.8	37.6	65.9	22.4
	Med42-Llama3-8B	10.8	27.5	42.2	70.3	12.6	8.8	24.8	38.1	63.7	7.5
	Meditron3-8B	5.7	26.4	41.1	53.9	22.9	4.6	26.3	41.9	29.6	13.5
	PMC-LLaMA-13B	0.0	30.1	45.6	27.3	45.2	0.0	15.6	27.5	0.0	0.0
MedRedQA	Med-Qwen2-7B	10.7	16.2	12.3	31.4	48.0	1.5	13.9	9.5	12.7	5.6
	Med42-Llama3-8B	11.5	14.6	11.2	32.0	48.0	2.0	13.2	9.8	14.0	4.5
	Meditron3-8B	7.4	14.5	11.3	26.7	51.4	1.2	13.5	10.1	9.7	6.0
	PMC-LLaMA-13B	9.8	14.4	10.5	7.4	75.0	0.0	9.2	7.5	3.7	5.3

Table 8: Text-level claim verification summary (ENTAIL/NEUTRAL/CONTRADICT histograms) across runs. Ent/Neu/Con denote the percentage of generated claims assigned to each NLI label.

Dataset	Generator	Avg. Resp. Claims	Avg. GT Claims	Ent (%)	Neu (%)	Con (%)
PubMedQA	Med-Qwen2-7B	2.10	4.90	88.6	4.0	7.4
	Med42-Llama3-8B	2.00	4.90	92.3	2.4	5.2
	Meditron3-8B	1.95	4.90	29.1	69.0	1.8
	PMC-LLaMA-13B	6.80	4.90	58.0	12.4	29.6
MedQuAD	Med-Qwen2-7B	4.80	10.20	88.1	2.4	9.4
	Med42-Llama3-8B	5.10	10.20	86.1	7.5	6.4
	Meditron3-8B	4.70	10.20	83.1	12.0	4.9
	PMC-LLaMA-13B	12.50	10.20	4.0	90.0	6.0
LiveQA	Med-Qwen2-7B	5.60	14.00	72.4	9.3	18.3
	Med42-Llama3-8B	6.81	13.06	12.3	86.1	1.7
	Meditron3-8B	5.20	14.12	10.6	87.9	1.6
	PMC-LLaMA-13B	17.50	14.00	27.3	27.5	45.2
MedRedQA	Med-Qwen2-7B	3.92	5.81	15.5	82.1	2.3
	Med42-Llama3-8B	3.82	5.80	19.3	78.2	2.5
	Meditron3-8B	7.98	7.71	54.6	45.1	0.3
	PMC-LLaMA-13B	21.01	9.23	28.1	71.1	0.8

D Prompt Templates

D.1 Teacher claim extraction prompt

You are a claim extraction assistant.

Task: Given an answer, extract a list of ATOMIC factual claims as SPO triples.

Output format (STRICT): Return ONLY a valid JSON array of triples: [["subject", "relation", "object"], ...] - No prose, no markdown, no preface, no trailing commas. - Use double quotes for all strings. - Each triple must contain exactly 3 strings.

Atomicity & faithfulness constraints: - Each triple must express a single, check-

able fact stated in the answer. - Do NOT paraphrase the whole sentence; split conjunctions into separate triples. - Do NOT introduce any new facts that are not explicitly stated in the answer. - Keep entity names as they appear in the answer when possible.

Negation / uncertainty / condition handling: - If the answer negates a fact, reflect negation in the RELATION (preferred) or OBJECT. Example: ["Drug A", "is not recommended for", "Condition B"] - If the answer is uncertain/probabilistic (may/can/likely), reflect it in RELATION. Example: ["Drug A", "may cause", "Side effect X"] - If the claim is conditional (if/when), include the condition in RELATION. Example: ["Drug A", "reduces X when", "taken with food"]

Now extract triples from the answer below.

ANSWER: answer

D.2 Teacher NLI verification prompt

You are a strict NLI verifier for biomedical QA.

Goal: Decide whether the CLAIM is supported by the provided PASSAGES.

Input: - CLAIM: a single atomic claim (hypothesis). - PASSAGES: retrieved text snippets (premises). Each passage includes a doc_id.

Decision labels: - ENTAIL: at least one passage explicitly supports the claim. - CONTRADICT: at least one passage explicitly states the opposite of the claim. - NEUTRAL: the passages do not provide sufficient information to entail or contradict. NEUTRAL includes: (a) insufficient evidence (relevant but incomplete) (b) irrelevant evidence (not about the claim)

Constraints: - Use ONLY the information in PASSAGES. Do NOT use external knowledge. - If evidence is missing, choose NEUTRAL (insufficient). - Prefer Entail/Contradict only with explicit textual support.

Output format (STRICT JSON): Return ONLY: "label": "Entail" | "Neu-

Table 9: Teacher sensitivity: GPT-4o vs GPT-4.1 on the same claim–evidence pairs.

Dataset	Label agreement $\uparrow$	κ $\uparrow$
PubMedQA	96.2	0.84
MedQuAD	91.9	0.74
LiveQA	68.9	0.53
MedRedQA	32.1	0.13

tral" | "Contradict", "prob": "Entail":
<float>, "Neutral": <float>, "Contradict":
<float>, "neutral_type": "insufficient" |
"irrelevant" | null, "rationale": <string>,
"spans": ["doc_id": <string>, "quote":
<string>] - prob values must sum to
1. - quote should be a short support-
ing/contradicting span (≤ 25 words). -
If label is Neutral, spans can be [].

Now verify.

CLAIM: claim

PASSAGES:

topk_passages_with_doc_id

E Teacher Sensitivity

We compare GPT-4.1 and GPT-4o teachers on the same claim–evidence pairs to assess teacher robustness; results are in Table 9. Generator rankings are stable on research-style datasets (PubMedQA, MedQuAD), but absolute values on consumer-health data (MedRedQA, $\kappa = 0.13$) are teacher-conditional.

F BioPortal Scoring Details

For linked entity pairs in BioPortal, we compute plausibility using three complementary methods and take the maximum:

1. *Direct Property Lookup* (score: 0.9): query the source entity’s ontology properties (e.g., indication, contraindication) and check if the target entity appears in the property values.
2. *Hierarchical Proximity* (score: 0.2–0.7): traverse the ontology hierarchy to find common ancestors; score increases with the number of shared ancestors.
3. *Cross-Ontology Mapping via LOOM* (score: 0.8): check whether the entity pair is connected through LOOM (Lexical Ontology Mapping) links across ontologies.

The direct-property signal is relation-aware: a property contributes only when its mapped predicate and direction are compatible with the claim predicate. Hierarchical proximity and LOOM mappings provide entity-pair compatibility rather than direct evidence for the claim relation. Accordingly, the combined value is treated as a heuristic compatibility score, not a probability of biomedical truth.

The final BioPortal KG score is:

$$s_{KG}^{\text{Bio}}(c) = \max(s_{\text{direct}}, s_{\text{hier}}, s_{\text{loom}}). \quad (8)$$

If no entity can be linked, the claim is treated as KG-uncovered and falls back to NLI-only verification.

G KG Case Study

When does KG fusion help vs. hurt? We categorize the 86 human-evaluated decision-flip claims by relation family. Both beneficial and harmful flips occur: drug-related claims provide clinically relevant opportunities for external structure to change an over-confident textual decision, whereas broader lifestyle claims are more susceptible to noisy hierarchical matches. Because the strata are small, we report them as qualitative error categories and do not describe a flip as a correction without a corresponding human evidence judgment.

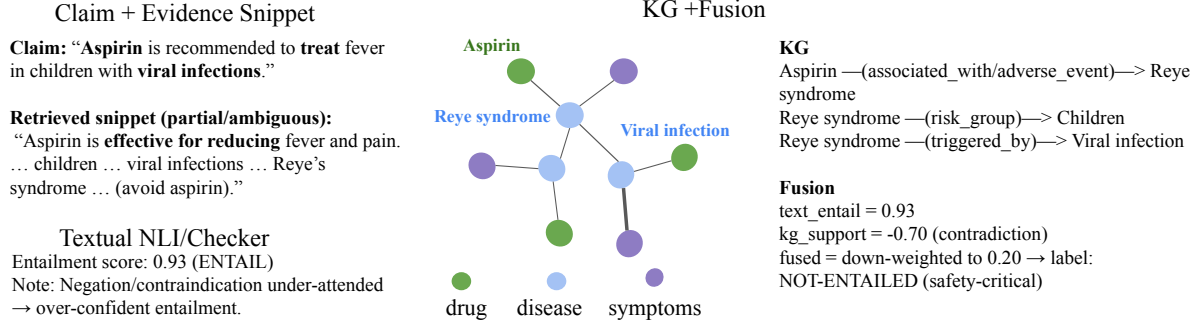


Figure 3: A decision-flip example from the human-evaluated subset. The figure reports the claim, retrieved evidence, linked entities, textual score, KG compatibility score, fused score, and blinded human evidence judgment. A low KG score is not interpreted as standalone proof that the claim is false.

H Span-Level Extraction Metrics

Table 10 reports conservative token-overlap P/R/F1 for claim extraction. These numbers are lower than the BERTScore-based evaluation in Table 1 due to surface-form differences in span granularity.

Table 10: Span-level token-overlap P/R/F1 for claim extraction. Values are conservative due to granularity mismatch; see Table 1 for semantic evaluation. **Bold** marks the best value in each column.

Generator	P	R	F1
Med-Qwen2-7B	11.9	12.5	12.2
Med42-Llama3-8B	21.4	26.7	23.7
Meditron3-8B	20.5	25.1	22.6
PMC-LLaMA-13B	0.3	0.2	0.2

I Claim Extraction Error Analysis

To complement the quantitative extraction metrics in Table 1, we provide a qualitative analysis of common extraction failure modes. We manually inspected 50 student-extracted claim sets (across all four generators) and categorized errors into three types:

Over-splitting (12/50). The student decomposes a single factual unit into two or more redundant claims. *Example:* For the answer “Metformin is the first-line treatment for type 2 diabetes and works by reducing hepatic glucose production,” the teacher extracts one claim [“Metformin”, “works by reducing”, “hepatic glucose production”], while the student produces two: [“Metformin”, “works by”, “reducing glucose”] and [“Metformin”, “reduces”, “hepatic glucose production”]. Over-splitting inflates claim counts but rarely introduces factual errors; answer-level metrics are unaffected because both sub-claims remain verifiable.

Under-splitting (8/50). The student merges multiple atomic facts into a single compound claim. *Example:* For the answer “Ibuprofen is an NSAID; it can cause GI bleeding and is contraindicated in renal failure,” the teacher produces three

atomic triples, while the student produces one compound: [“Ibuprofen”, “is an NSAID that can cause”, “GI bleeding and is contraindicated in renal failure”]. Under-splitting reduces granularity but does not introduce hallucinations; the NLI checker can still verify the merged claim if all sub-facts are supported.

Hallucinated claims (3/50). Rarely, the student introduces a claim not present in the original answer. *Example:* Given an answer about hypertension management that mentions “ACE inhibitors,” the student adds [“ACE inhibitors”, “are preferred over”, “ARBs for diabetic patients”], a preference not stated in the answer. These errors are the most harmful but are infrequent (6%) and are flagged by our NLI checker as NEUTRAL (unsupported by retrieved evidence).

Summary. Over-splitting is the most common failure (24%) but is benign for downstream diagnostics. Hallucinated extracted claims are rare in this sample (6%). These findings are consistent with the high claim quality scores (4.62–4.97/5) reported in §4.7.

J Per-Dataset KG Coverage and Consistency

Table 11: DRKG coverage and KG-subset behavior (pair-hit definition). We report node/pair coverage and KG-subset faithfulness/hallucination metrics on the DRKG-aligned subset. **Bold** marks the best value per dataset and column; degenerate rows (PMC-LLaMA-13B on MedQuAD with Faith=4.0) are excluded from the bolding for Faith/Hall columns where the underlying generator output collapses.

Dataset	Generator	KG-Cov _{node} (%)	KG-Cov _{pair} (%)	Faith	Faith _{KG}	Hall _{KG}
LiveQA	Med-Qwen2-7B	84.2	51.3	72.4	86.3	8.8
	Med42-Llama3-8B	68.5	56.4	63.7	87.8	9.4
	Meditron3-8B	70.0	52.6	53.9	65.7	26.1
	PMC-LLaMA-13B	65.0	30.0	27.3	25.6	57.7
PubMedQA	Med-Qwen2-7B	83.0	49.0	88.6	96.1	3.9
	Med42-Llama3-8B	95.0	83.0	92.4	94.5	5.0
	Meditron3-8B	88.3	62.7	69.2	85.8	6.2
	PMC-LLaMA-13B	72.0	35.0	60.5	57.1	42.9
MedQuAD	Med-Qwen2-7B	93.0	69.0	88.1	87.2	11.2
	Med42-Llama3-8B	94.0	88.0	86.1	89.0	7.2
	Meditron3-8B	84.0	66.0	83.1	89.5	7.4
	PMC-LLaMA-13B	8.0	4.0	4.0	0.0	92.3
MedRedQA	Med-Qwen2-7B	90.0	53.4	37.8	32.8	50.3
	Med42-Llama3-8B	92.0	74.3	31.3	36.7	48.5
	Meditron3-8B	90.2	46.3	26.7	26.0	55.1
	PMC-LLaMA-13B	92.7	86.6	7.4	6.1	76.2

K Annotation Rating Scales and Guidelines

This appendix summarizes the annotation protocol. Each row in the annotation sheet corresponds to a single question, containing (i) the original question, (ii) model-generated answers from each main generator (Meditron3-8B, Med42-Llama3-8B, and Med-Qwen2-7B), and (iii) the automatically extracted claims for each model. Annotators independently rate how faithfully each claim set represents its corresponding answer using the 1–5 Likert scale defined in Appendix M (Table 12).

L Calibration Details

Selected operating point. The main experiments use $\beta = 0.7$, support threshold $\tau = 0.5$, BioPortal entity-linking threshold 0.5, and selective-gating thresholds $s_{KG} < 0.3$ and $p_{NLI} > 0.7$. DRKG’s interpolation weight α is selected on the development split and is used only for DRKG scoring; all selected values are fixed before test evaluation.

L.1 Fusion-weight sensitivity (β -sweep)

We study how the fused support score P^* and downstream supported/not-supported decisions change as we vary the mixing weight β in Eq. 4. Unless otherwise specified, we fix α to the dev-tuned value and sweep $\beta \in \{0, 0.1, \dots, 1.0\}$.

Metrics. We report (i) the fused supported rate (fraction of claims with $P^* \geq \tau$) and (ii) the deci-

sion flip rate relative to a near-NLI reference setting (e.g., $\beta = 0.9$).

Supported rate and flip rate. For a given mixing weight β , we compute fused scores $P_\beta^*(c)$ via Eq. 4 and binarize them with the dev-tuned threshold τ :

$$\mathbb{I}_{\text{sup},\beta}(c) \triangleq \mathbb{I}[P_\beta^*(c) \geq \tau].$$

We summarize fusion behavior on a claim set $\mathcal{C}$ using

$$\text{fused_E}(\beta) = 100 \cdot \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}_{\text{sup},\beta}(c),$$

and the decision flip rate relative to a near-text-only reference β_0 (we use $\beta_0=0.9$):

$$\text{flip_rate}(\beta) = \frac{100}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}[\mathbb{I}_{\text{sup},\beta}(c) \neq \mathbb{I}_{\text{sup},\beta_0}(c)]. \quad (9)$$

A lower flip_rate means fusion decisions are less dominated by the KG score scale and thus more stable across β .

Min–max calibration of KG scores. Raw KG plausibility scores (e.g., TransE-based p_{KGE} and the derived s_{KG}) can be poorly scaled and concentrated in a narrow range. To reduce scale mismatch in the logit-mixture fusion (Eq. 4), we apply min–max calibration to the KG score using development-split statistics:

$$\tilde{s}_{KG}(c) = \text{clip}\left(\frac{s_{KG}(c) - s_{\min}}{s_{\max} - s_{\min}}, \epsilon, 1 - \epsilon\right), \quad (10)$$

where $s_{\min}$ and $s_{\max}$ are the minimum and maximum KG scores observed on the dev set (restricted to KG-aligned claims), and ϵ is a small constant to avoid infinite logits. We then replace $s_{\text{KG}}(c)$ with $\tilde{s}_{\text{KG}}(c)$ when computing Eq. 4.

Why calibrating KG scores matters. Raw KG plausibility scores can be poorly calibrated and concentrated in a narrow range, which makes the logit term $\text{logit}(s_{\text{KG}})$ disproportionately large (in magnitude) and causes fusion decisions to change sharply with β . Min-max calibration aligns the KG score scale to the NLI probability range while preserving ranking, substantially reducing flip rates and yielding flatter supported-rate curves in Fig. 4.

L.2 Threshold robustness (τ -sweep)

We additionally sweep the support threshold τ around the dev-tuned value and observe stable trends in the supported rate and SafetyErr within a reasonable range.

For completeness, we report teacher checker results for all dataset/generator configurations in Table 7, complementing the representative results in the main text.

M Human Evaluation: Protocol

We conduct human evaluation to assess whether extracted claim sets faithfully represent the corresponding model answers.

M.1 Setup

Sampling. We sample 100 questions from MedQuAD, LiveQA, MedRedQA, and Pub-MedQA. For each question, we evaluate answers from the three main generators (Meditron3-8B, Med-Qwen2-7B, and Med42-Llama3-8B) and their extracted claim sets.

Annotators. Annotators are graduate students with training in biomedical or health sciences. Each model output is rated independently; annotators do not rank models.

Unit of annotation. For each (*question, model*) pair, annotators compare the model answer with its extracted claim set.

M.2 Dimensions and scoring

Annotators assign a 1–5 Likert score for **claim-extraction quality**: whether the claims faithfully reflect the answer without omitting central factual content or introducing new content. Scores 5/3/1

Table 12: Anchor rubric for claim-extraction quality (2 and 4 are intermediate).

Score	Claim-extraction anchor
5	Faithfully covers the answer’s key factual content without adding unsupported claims.
3	Contains noticeable omissions, redundancy, or distortions but remains partly usable.
1	Fails to represent the answer or introduces clear claims not present in it.

correspond to *excellent / mixed / poor*, with 2 and 4 as intermediate levels.

When to comment. Annotators provide a brief comment whenever claim quality is ≤ 3 .

M.3 Agreement and reporting

We report $\text{mean} \pm \text{std}$ over samples after averaging the two ratings for each item and quadratic-weighted Cohen’s κ for inter-annotator agreement.

Full guidelines. Detailed examples and interface schemas are provided in our public repository.

N Human Evaluation on KG-Induced Decision-Flips

Sampling. We sample claims from the KG-aligned subset where the final verification decision differs between NLI-only and Fused (NLI+KG) settings. Claims are stratified by relation families used in the safety analysis (e.g., drug–disease, drug–adverse-event, gene–disease).

Sample size. The decision-flip subset consists of $n = 86$ claims drawn from outputs for 100 questions and three generators, stratified across drug–disease, drug–adverse-event, gene–disease, and lifestyle relations. Each claim is independently annotated by both annotators. We evaluate predictions against all 172 annotator–claim judgments rather than constructing a majority label from two annotators. Given the modest sample size, we interpret the +6.4pp difference descriptively.

Annotation unit. Each annotation instance consists of a single atomic claim and the retrieved passages used for verification. Annotators are blinded to the NLI prediction, fused prediction, KG score, and aligned KG triple, and judge the claim only against the retrieved passages.

Labels. Annotators assign one of {ENTAIL, NEUTRAL, CONTRADICT} to each claim. We compare

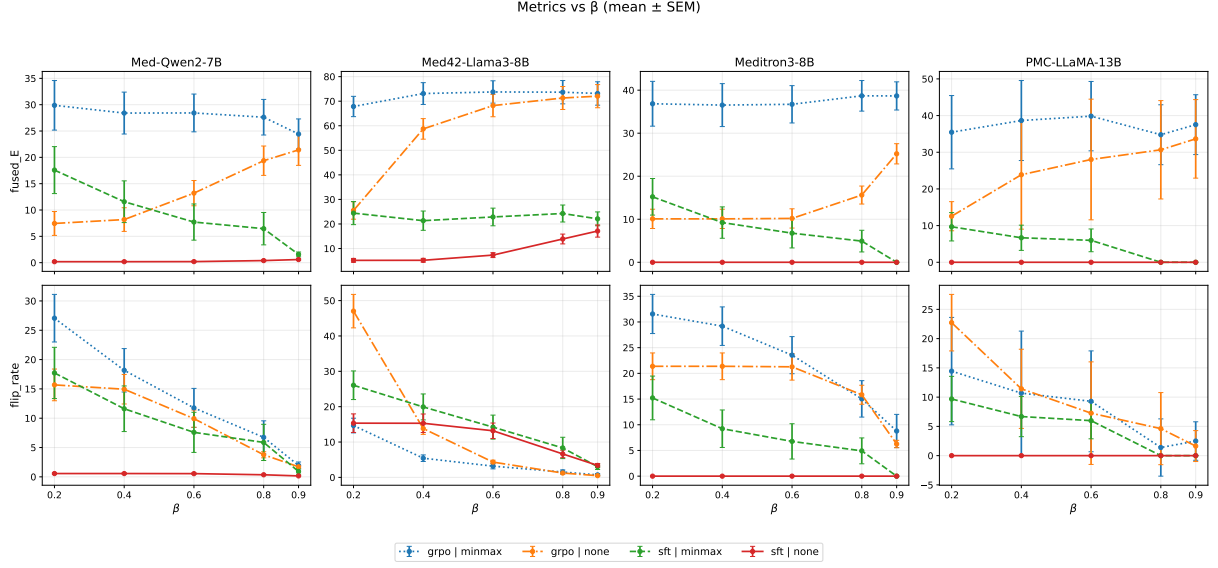


Figure 4: Sensitivity of KG–NLI fusion to the mixing weight β . Top: the fraction of claims classified as supported after fusion. Bottom: decision-flip rate relative to a near-NLI reference ($\beta=0.9$). The SFT curves correspond to the reported system; GRPO traces are shown only as an exploratory sensitivity reference. minmax and none denote calibrated and raw KG scores.

each independent human judgment against NLI-only and fused predictions to compute agreement and macro-F1.

O Retrieval settings and indices

Overview. We standardize retrieval across datasets by using (i) a FAISS index for MedRedQA runs, (ii) a pre-retrieved pickle for MedQuAD, and (iii) the MedRAG retrieval system for the remaining datasets. Table 13 summarizes the mapping from dataset to retriever and the corresponding index artifacts. Table 14 lists the index files and where they are stored.

P Student Configurations

All models are trained on a single NVIDIA A100 80GB GPU. Hyperparameters for all distillation stages are in Table 15.

F1-weighted ensemble. Different student checkers specialize in different NLI classes, so we combine their predictions using per-class dev-set F1 as reliability weights. Let $p_m(y | c_i, D)$ be checker m ’s probability for label y , and $F1_m^{(y)}$ its dev-set F1 on class y . We weight each checker by

$$w_m^{(y)} = \frac{F1_m^{(y)}}{\sum_{m'} F1_{m'}^{(y)}}, \quad (11)$$

and predict

$$\hat{y}_i = \arg \max_y \sum_m w_m^{(y)} p_m(y | c_i, D). \quad (12)$$

This lets minority-class specialists (especially CONTRADICT) contribute more to those decisions.

Table 13: Dataset-to-retriever mapping and the evidence source used for retrieval.

Dataset	Retriever	Index / evidence source
MedRedQA	MedRedQAFaissRetriever / gold	FAISS mode (Johnson et al., 2021): FAISS approximate nearest-neighbor (ANN) index built over the MedRedQA corpus (faiss.index + texts.jsonl + meta.jsonl under <code>-MedRedQA_index_dir</code>); gold mode: use provided contexts directly (no external index).
MedQuAD	MedQuADRetriever	Pre-retrieved contexts loaded from <code>-medquad_pickle</code> (a pickle file containing query-to-context mappings).
PubMedQA / LiveQA / MedRedQA (non-FAISS mode)	MedRAGRetriever	MedRAG retrieval over a selected corpus (e.g., PubMed / MedCorp) using BM25 (Robertson and Zaragoza, 2009) (<code>-retriever bm25</code>), dense (contriever/specter/medcpt), or hybrid (reciprocal rank fusion, RRF (Cormack et al., 2009)). Indices are stored under <code>-medrag_db_dir</code> .

Table 14: Index artifacts used by each retriever family.

Retriever family	Key artifacts	Where it lives / how it is used
MedRedQA FAISS	faiss.index, texts.jsonl, meta.jsonl	All under <code>-MedRedQA_index_dir</code> . Retrieved hits are mapped to text via texts.jsonl (with optional metadata in meta.jsonl).
MedQuAD pre-retrieval	retrieved_*.pkl	<code>-medquad_pickle</code> . No ANN/BM25 search at runtime; directly loads cached retrieval results.
MedRAG BM25	Lucene/Pyserini (Lin et al., 2021) BM25 index (directory)	Stored under <code>-medrag_db_dir</code> (per corpus). Used when <code>-retriever bm25</code> .
MedRAG dense / hybrid	faiss.index (+ metadata file such as metadatas.jsonl)	Stored under <code>-medrag_db_dir</code> (per corpus and retriever). Used when <code>-retriever</code> is dense (e.g., contriever/specter/medcpt) or hybrid (RRF fusion).

Table 15: Training configurations for the distilled extractor and checker models. L denotes tokenizer truncation length.

Stage	Base models	lr	ep	bs	acc	L	Gen	LoRA / Extra
Extractor SFT	Med42-Llama3-8B Meditron3-8B Med-Qwen2-7B PMC-LLaMA-13B	1e-4	3	2	16	2048	max_new_tokens=256	r=16, α =32, drop=0 targets: q,k,v,o,gate,up,down
Checker SFT	Med42-Llama3-8B Meditron3-8B Med-Qwen2-7B PMC-LLaMA-13B	5e-5	3	2	16	1024	eval: max_new_tokens=4 eval_subset=1000	r=8, α =16, drop=0 targets: q,k,v,o,gate,up,down

Q Recommended Default Settings

For practitioners deploying MEDRAGCHECKER on a new biomedical QA dataset, we recommend the following defaults based on our dev-set tuning across four benchmarks:

- **KG back-end:** BioPortal (94.2% entity-pair coverage; strongest student-ensemble results in our evaluated configurations).
- **Fusion weight β :** 0.7 (text-dominant; robust across $\beta \in [0.6, 0.8]$ per Appendix L).
- **KG calibration:** Apply min-max calibration (Appendix L.1) on dev-set KG scores before fusion.
- **Support threshold τ :** 0.5 (tune on dev if possible; stable within ± 0.1).
- **Selective gating:** Enable ($\tau_{\text{conflict}} = 0.3$, $\tau_{\text{entail}} = 0.7$) to restrict KG influence to the

evaluated high-NLI/low-KG regime.

- **Checker:** F1-weighted SFT ensemble of Med42-Llama3-8B + Med-Qwen2-7B + Meditron3-8B. If only one model is feasible, use Med42-Llama3-8B (best macro-F1 and most stable fused distribution).
- **Diagnostics to report when KG coverage is low:** Report NLI-only Faithfulness and Hallucination alongside the fused metrics. The fused versions are KG-dependent and should not be described as KG-independent.