Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot, 2025. URL https://arxiv.org/abs/2502.04413.

# Appendix

## A The Use of Large Language Models

In accordance with ICLR 2026 policy, we disclose the usage of Large Language Models (LLMs) in the process of writing this paper. Specifically, we employed LLMs to assist with the refinement and polishing of the manuscript's language. The LLM was used to enhance clarity, improve grammar, and ensure the consistency of the text, which contributed to the overall quality of the writing. The LLM was not used to generate novel ideas, research findings, or substantial portions of the content. Its primary role was as a tool to aid the revision process, focusing on language-related tasks.

We have fully disclosed this usage, and the final manuscript reflects the work of the authors. The LLM's contribution is limited to textual improvements and does not extend to the intellectual content of the paper.

For transparency, we confirm that the research itself, including the methodology, results, and conclusions, was independently developed by the authors without any contributions from LLMs beyond their role in writing assistance.

## B Detailed Performance Comparisons

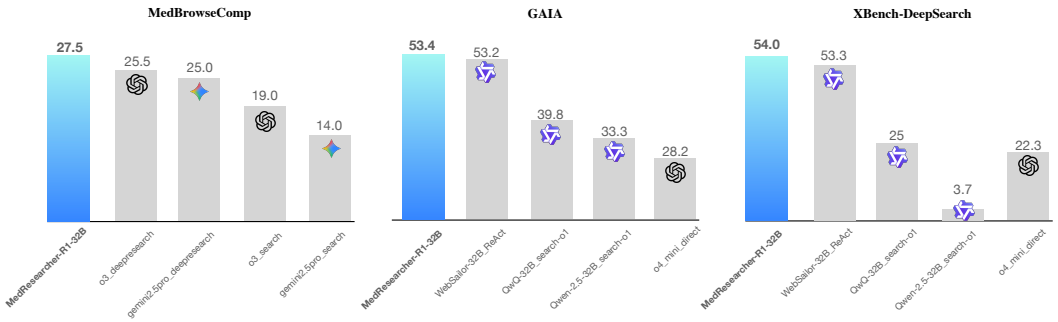


Figure 5: Overall performance of MedResearcher-R1 across three benchmarks. On MedBrowseComp, our MedResearcher-R1-32B achieves state-of-the-art performance with 27.5/50 correct answers, surpassing o3-deepresearch (25.5/50), Gemini-2.5-Pro-deepresearch (25.0/50), and significantly outperforming search-only approaches (o3-search: 19.0/50, Gemini-2.5-Pro-search: 14.0/50). On general deep research tasks, we achieve competitive results on GAIA (53.4 vs. WebSailor-32B's 53.2) and xBench (54.0 vs. WebSailor-32B's 53.3).

## C Technical Details

### C.1 PrivateMedicalRetriever

This module aggregates evidence directly from authoritative clinical resources, including FDA databases, clinical trial registries, and PubMed publications. Each candidate document $d$ is scored for a query $q$ by a weighted linear combination of semantic relevance and clinical authority:

$$\text{Score}(d, q) = \lambda \,\text{Rel}(d, q) + (1 - \lambda)\,\text{Auth}(d),$$

where $\text{Rel}(d, q)$ represents the semantic similarity to the query (computed via embedding cosine similarity), and $\text{Auth}(d)$ reflects the clinical authority (combining impact factor and guideline status). The hyperparameter $\lambda$ ($0 \leq \lambda \leq 1$) balances the importance between relevance and authority; in all experiments, we set $\lambda = 0.4$ to favor reliable and clinically significant evidence.

### C.2 CLINICALREASONINGENGINE

Designed for evidence-based differential diagnosis, this tool applies Bayesian inference to evaluate multiple hypotheses systematically. Given observed symptoms $\mathbf{s}$, candidate diagnoses $D_j$, and patient context $\mathbf{c}$, the posterior for each diagnosis is computed as:

$$P(D_j \mid \mathbf{s}, \mathbf{c}) = \frac{\prod_{i=1}^{n} P(s_i \mid D_j, \mathbf{c}) \cdot P(D_j \mid \mathbf{c})}{\sum_{k=1}^{m} \prod_{i=1}^{n} P(s_i \mid D_k, \mathbf{c}) \cdot P(D_k \mid \mathbf{c})}$$

where conditional probabilities are derived from clinical literature and iteratively updated based on newly retrieved evidence.

### C.3 DYNAMIC TOOL SELECTION STRATEGY

Our agent dynamically switches between general and medical-specific tools to ensure complete evidence chains. The tool selection is governed by a learned policy that evaluates query complexity:

$$P(t \mid s_t, q) = \begin{cases} \sigma(\mathbf{w}_m^T \phi(s_t, q)) & \text{if } t \in \mathcal{T}_{\text{medical}} \\ \sigma(\mathbf{w}_g^T \phi(s_t, q)) & \text{if } t \in \mathcal{T}_{\text{general}} \end{cases}$$

where $\phi(s_t, q)$ extracts features that include the rarity of the entity, the required reasoning hops, and the presence of medical terminology, $\mathbf{w}_m$ and $\mathbf{w}_g$ are learned weight vectors, and $\sigma(\cdot)$ is the sigmoid function. The policy learns to prioritize medical tools when encountering rare diseases or complex chemical compounds while leveraging general tools for contextual information.

## D TRAINING IMPLEMENTATION DETAILS

### D.1 SUPERVISED FINE-TUNING CONFIGURATION

**Dataset.** We train on $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ where $N = 2,137$ trajectories, with $x^{(i)}$ denoting input context and $y^{(i)}$ the expert action sequence. The objective maximizes:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|y^{(i)}|} \log p_\theta(y_k^{(i)} | x^{(i)}, y_{<k}^{(i)}) \tag{1}$$

**Robustness Augmentations.**

- **Tool failure simulation:** 5% random corruption of tool outputs to encourage error recovery
- **Intermediate thought supervision:** Explicit reasoning traces before each tool invocation
- **Multi-task sampling:** Balanced batching across diagnosis (30%), treatment (25%), guidelines (25%), rare diseases (20%)

**Optimization.**

- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.98$
- Learning rate: $\lambda = 0.01$ with cosine annealing to $\eta_{\min} = 3 \times 10^{-7}$
- Batch size: 128 (16 per GPU $\times$ 8 H800 GPUs)
- Training epochs: 3
- Gradient clipping: 1.0
- Warmup steps: 100

### D.2 REINFORCEMENT LEARNING CONFIGURATION

**Reward Components.** The composite reward function $r = \alpha r_{\text{task}} + \beta r_{\text{expert}} - \gamma r_{\text{efficiency}}$ comprises:

- $r_{\text{task}}$: Binary task completion (1.0 for correct, 0.0 for incorrect)
- $r_{\text{expert}}$: GPT-4 preference score $\in [0, 1]$ evaluating medical accuracy and completeness
- $r_{\text{efficiency}}$: Penalty for redundant tool usage, computed as:

$$r_{\text{efficiency}} = 0.1 \times n_{\text{redundant}} + 0.2 \times n_{\text{post-answer}} + 0.15 \times n_{\text{irrelevant}} \tag{2}$$

**GRPO Configuration.** The GRPO objective:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\log \pi_\theta(y|x) \cdot \left(r(x,y) - \bar{r}_{\mathcal{G}(x)}\right)\right] \tag{3}$$

where $\bar{r}_{\mathcal{G}(x)}$ is the group-level baseline (batch average).

- Group size: 4 responses per query
- Sampling temperature: 0.7
- PPO clip range: 0.2
- Value loss coefficient: 0.5
- Entropy coefficient: 0.01
- Training iterations: 500
- KL regularization: Disabled (following He et al. (2025))

**Curriculum Learning.** Task complexity increases based on moving average pass rate:

- Level 1 (pass rate $> 80\%$): Single-hop queries
- Level 2 (pass rate $> 60\%$): 2-3 hop queries
- Level 3 (pass rate $> 40\%$): 4+ hop queries with rare entities

## E Ablation Study Details

Table 3: Ablation study for MedResearcher-R1. We remove key components while keeping all other settings fixed. Statistical significance: $^*$ $p<0.05$ vs. the Full model. MedBrowseComp is reported as # correct out of 50.

| Model Configuration | MedBrowseComp (correct / 50) | GAIA (%) | XBench (%) | Avg. Tool Calls |
|---|---|---|---|---|
| **MedResearcher-R1 (Full)** | **27.5** | **53.4** | **54.0** | 4.2 |
| *Component Ablations* | | | | |
| w/o Medical Tools | 23.1 | 48.3 | 40.0 | 3.3 |
| w/o RL Training (SFT only) | 25.5 | 50.2 | 51.0 | 3.7 |
| w/o MTG | 24.2$^*$ | 44.3$^*$ | 47.8$^*$ | 3.5 |
| w/o Rare Entities | 20.1$^*$ | 27.8$^*$ | 38.2$^*$ | 3.2 |
| *Data Ablations* | | | | |
| Common Entities Only | 23.0$^*$ | 43.0$^*$ | 46.0$^*$ | 4.5 |
| No Tool Diversity | 21.0$^*$ | 38.0$^*$ | 49.0 | 3.2 |
| *Training Ablations* | | | | |
| SFT Only | 25.5$^*$ | 49.0 | 48.0$^*$ | 3.4 |
| RL Only (no SFT) | 12.0$^*$ | 34.0$^*$ | 34.0$^*$ | 3.2 |

**Tasks and metrics.**

- **MedBrowseComp** is reported as *correct/50*.
- **GAIA** and **XBench-DeepSearch** follow official (%) scoring.
- **Avg. Tool Calls** is the average number of tool invocations per example.

**Evaluation protocol.** All ablations share the same backbone, prompts, decoding parameters, tool budgets, and evaluation splits as the Full model; only the targeted component is removed/altered. Each number is the mean of three seeds.

**Significance.** We compute paired bootstrap (over instances) against the Full model with 10,000 replicates; we mark $^*$ for $p<0.05$.