PubMed LLM Evaluation

Introduction

This report presents the results of the PubMed-trained language model selection and evaluation conducted as part of the *tumour signalling pipeline optimisation project*.

The goal of this evaluation is to select a biomedical LLM capable of performing the core literature interpretation tasks required by the pipeline, including:

- mechanistic relevance assessments of citations,
- extraction of detailed mechanistic evidence,
- evaluation of evidence quality,
- structured output generation, and
- robust instruction following for controlled, reproducible outputs.

The selected LLM will form the foundation of the evidence assessment stages of the finalised pipeline.

Objective

The objective of the evaluation is to:

- Survey available PubMed-trained and biomedical language models;
- Identify 3-5 models suitable for testing;
- Evaluate these models using a representative test suite aligned to the project's literature-filtering requirements;
- Compare performance across accuracy, consistency, output quality, and practical deployment considerations; and
- Recommend a primary and backup LLM for integration into the optimised pipeline.

This ensures that the model not only performs well on biomedical NLP tasks, but also meets the project's strict requirements for structured outputs, determinism, and deployment feasibility.

Scope of Evaluation

This report covers:

- Model landscape review
- Model selection methodology
- Evaluation criteria
- Test suite design
- Performance results
- Comparative analysis
- Risks and limitations
- Final model recommendation and backup option

The evaluation aligns fully with the project requirements and is designed to support downstream integration and deployment on Hartree.

Model Landscape

Model	Parameters	Training Data	Last Update	License
microsoft/BioGPT-Large	1.5B	PubMed (15M abstracts)	September 2022	MIT
microsoft/MediPhi-PubMed	3.8B	PubMed, Medical Wikipedia, Guidelines, Clinical notes	May 2025	MIT
epfl-llm/meditron-7b	7B	PubMed, Clinical Guidelines, RedPajama-v1	December 2023	Llama 2
BioMistral/BioMistral-7B	7B	PubMed Central	February 2024	Apache-2.0
ContactDoctor/Bio-Medical-Llama-3-8B	8B	Custom biomedical dataset (500K+ entries)	August 2024	🔴 Non-commercial
stanford-crfm/BioMedLM	2.7B	PubMed abstracts and full articles (The Pile)	December 2022	BigScience RAIL-M

Published PubMedQA Benchmark Scores

PubMedQA is a standard biomedical question-answering benchmark where models must answer research questions with yes/no/maybe using corresponding abstracts. Human expert performance is **78.0%**.

Model	PubMedQA Accuracy	Notes
Human Expert	78.0%	Single-rater baseline
microsoft/BioGPT-Large	81.0%	Fine-tuned on PubMedQA training data
epfl-llm/meditron-7b	74.4%	Fine-tuned
BioMistral/BioMistral-7B	77.7%	DARE variant
stanford-crfm/BioMedLM	74.4%	-

Note: Scores are from published papers and may use different evaluation settings (fine-tuning, few-shot, etc.). Published scores not available for all models.

Evaluation Methodology

Two-Stage Evaluation Pipeline

To accommodate models that struggle with JSON formatting, a two-stage pipeline was developed:

1. **Generation**: The evaluation model generates plaintext analysis of the biomedical abstract
2. **Parsing**: A separate parser model (via Instructor) extracts structured data from the plaintext

This approach separates biomedical reasoning ability from JSON formatting capability, allowing fair evaluation of models that may have strong domain knowledge but poor instruction-following for structured outputs.

Test Suite

A minimal test suite was produced, comprising tests relevant to the project covering four main capabilities:

Test 1 - Mechanistic Relevance Classification

Models should determine whether the citation provides mechanistic evidence for the agent-pathway pair. This tests that the subject-specific knowledge of the model is sufficient for the task, and that it can effectively interpret the citations.

- **Test items**: 50
- **Metric**: Exact match accuracy

Test 2 - Mechanism Extraction

Models should extract mechanistic summaries and molecular components from the citation. This tests that models are capable of understanding the content and extracting the key evidence.

- **Test items**: 50
- **Metric**: Direction of effect accuracy (activation/inhibition)

Test 3 - Evidence Quality Classification

Models should assess the quality of the citation in relation to providing evidence for the agent-pathway pair (strong/moderate/weak/insufficient). This checks the models' accuracy in determining the quality of a citation.

- **Test items**: 50
- **Metric**: Exact match accuracy, within-1-step accuracy

Test 4 - Parsing Stability

Models should generate consistent outputs across repeated runs. This tests reproducibility and ensures absence of significant hallucinations or randomness in interpretation.

- **Test items**: 50
- **Metric**: Consistency rate across duplicate runs

Test Environment

- **Hardware**: NVIDIA RTX 5090
- **Parser provider**: Mistral (via Instructor)
- **Framework**: Python, PyTorch, HuggingFace Transformers

Results

Model	Relevance	Mechanism	Quality (±1 step)	Stability	Avg. Gen Time
microsoft/BioGPT-Large	60.0%	42.0%	86.0%	72.0%	6.9s
microsoft/MediPhi-PubMed	86.0%	70.0%	98.0%	100.0%	8.4s
epfl-llm/meditron-7b	46.0%	36.0%	60.0%	56.0%	12.4s

BioMistral/BioMistral-7B	76.0%	46.0%	82.0%	66.0%	1.1s
ContactDoctor/Bio-Medical-Llama-3-8B	72.0%	18.0%	86.0%	70.0%	1.6s
stanford-crfm/BioMedLM	66.0%	16.0%	52.0%	42.0%	14.5s

Comparative Analysis

Performance Summary

microsoft/MediPhi-PubMed demonstrated the strongest overall performance across all evaluation tasks:

Metric	MediPhi-PubMed	Next Best	Improvement
Relevance	86.0%	BioMistral (76.0%)	+10.0%
Mechanism	70.0%	BioMistral (46.0%)	+24.0%
Quality (±1 step)	98.0%	BioGPT-Large (86.0%)	+12.0%
Stability	100.0%	BioGPT-Large (72.0%)	+28.0%

Key Findings

1. **MediPhi-PubMed excels in all categories**: The model achieved the highest scores in relevance classification (86%), mechanism extraction (70%), quality assessment (98%), and perfect stability (100%). This consistent performance across diverse tasks indicates strong generalisation for biomedical literature interpretation.

2. **BioGPT-Large provides a reliable fallback**: Despite being the oldest and smallest model evaluated (1.5B parameters), BioGPT-Large achieved competitive quality assessment (86%) and reasonable stability (72%). Its MIT license and proven track record make it a suitable backup option.

3. **Larger models did not guarantee better performance**: Meditron-7B (7B params) and Bio-Medical-Llama-3-8B (8B params) underperformed relative to MediPhi-PubMed (3.8B params), suggesting that training data quality and recency outweigh parameter count for this task.

4. **Stability varies significantly**: Only MediPhi-PubMed achieved 100% stability. BioMedLM (42%) and Meditron-7B (56%) showed concerning inconsistency that would undermine reproducibility requirements.

5. **Inference speed trade-offs**: BioMistral-7B (1.1s) and Bio-Medical-Llama-3-8B (1.6s) offer fast inference but sacrifice accuracy. MediPhi-PubMed's 8.4s inference time is acceptable given its superior performance.

Licensing Considerations

Model	License	Commercial Use
microsoft/MediPhi-PubMed	MIT	Yes
microsoft/BioGPT-Large	MIT	Yes

BioMistral/BioMistral-7B	Apache-2.0	Yes
epfl-llm/meditron-7b	Llama 2	Yes (with restrictions)
ContactDoctor/Bio-Medical-Llama-3-8B	Non-commercial	No
stanford-crfm/BioMedLM	BigScience RAIL-M	Restricted

Both recommended models (MediPhi-PubMed and BioGPT-Large) use the MIT license, ensuring unrestricted commercial deployment on Hartree.

Risks and Limitations

- **Parser dependency**: The two-stage pipeline introduces dependency on the parser model's accuracy. Parser errors may propagate to final outputs.
- **Test suite size**: 50 items per test provides a representative sample but may not capture all edge cases in production use.
- **Hardware constraints**: MediPhi-PubMed (3.8B parameters) requires approximately 8GB VRAM; ensure Hartree nodes meet this requirement.
- **Model recency**: MediPhi-PubMed was updated May 2025, incorporating recent biomedical literature. Performance on older or highly specialised topics should be monitored.
- **Quality assessment subjectivity**: The "within 1 step" tolerance for quality classification may mask meaningful disagreements between model and gold labels.

Recommendation

Primary Model

microsoft/MediPhi-PubMed

MediPhi-PubMed is recommended as the primary model for the tumour signalling pipeline based on:

- **Superior accuracy**: Highest scores across all four evaluation tasks
- **Perfect stability**: 100% consistency ensures reproducible outputs
- **Permissive licensing**: MIT license allows unrestricted commercial deployment
- **Modern architecture**: Based on Phi-3.5-mini with 128K context length, suitable for processing long abstracts
- **Diverse training**: Trained on PubMed, Medical Wikipedia, Guidelines, and Clinical notes, providing broad biomedical coverage
- **Acceptable inference time**: 8.4s per item is suitable for batch processing pipelines

Backup Model

microsoft/BioGPT-Large

BioGPT-Large is recommended as the backup option based on:

- **Proven benchmark performance**: 81% on PubMedQA (exceeds human expert baseline of 78%)
- **Competitive quality assessment**: 86% on evidence quality classification matches project requirements
- **Lightweight deployment**: 1.5B parameters enables deployment on resource-constrained nodes
- **MIT license**: Same permissive licensing as primary model
- **Mature ecosystem**: Well-documented with extensive community support

Deployment Strategy

1. Deploy MediPhi-PubMed as the primary inference endpoint
2. Implement fallback logic to route to BioGPT-Large if primary model is unavailable or exceeds latency thresholds
3. Monitor both models' performance on production data and retrain/fine-tune as needed
4. Consider fine-tuning MediPhi-PubMed on project-specific examples to further improve accuracy on tumour signalling literature