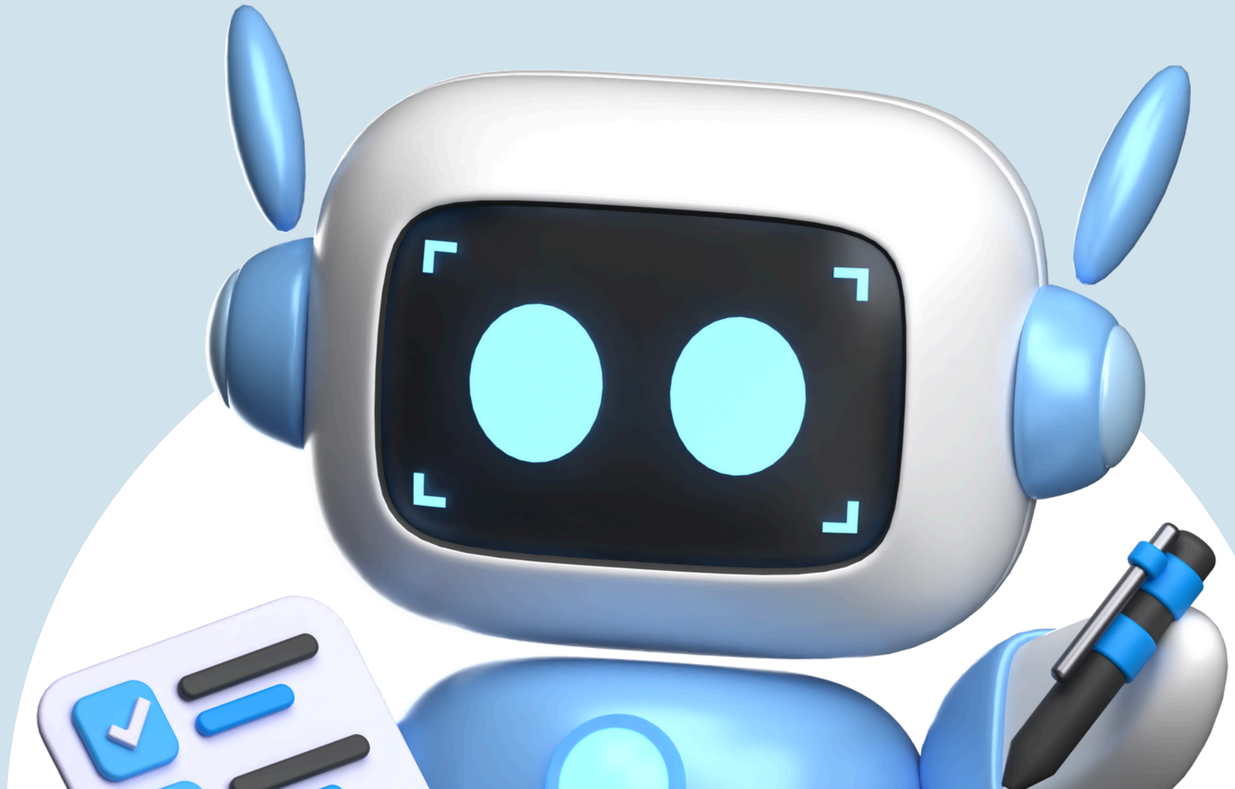
Capstone Phase 2

Niloofar Karimi,
Rimsha Kayastha,
Yanzheng Liu

# MULTI-LEVEL SUMMARIZATION OF INDIVIDUAL RESEARCH PAPERS



Multi-level summarization of individual research papers

# Problem

- Steady **annual decreases in readability** of papers from 2010 to 2024, with high complexity post-ChatGPT
- Clinical journals getting harder to read: **grade levels of college graduate** (FRE Scores below 30) [1]
- Patient education materials: grade levels of **11.2 to 13.8** (FRE Scores 30-50) [1]

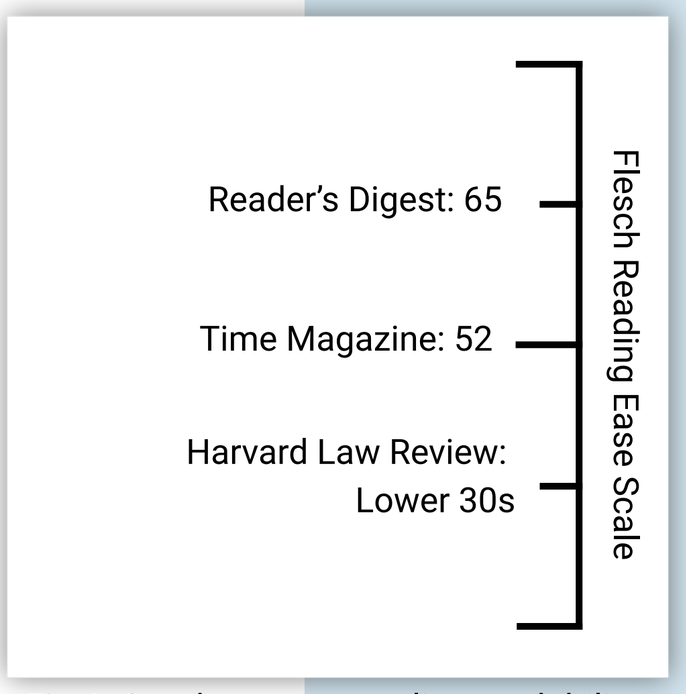



Fig 1: Graph to contexualize readability; not drawn to scale

# P2 Goal

- **Fine-tune a summarization model on biomedical research papers (M3)** and evaluate whether simplification emerges when moving from general-domain (OSE) to medical-domain (M3) fine-tuning.
- Simplification did not emerge naturally in Phase 2 due to lack of biomedical multi-level simplification data.

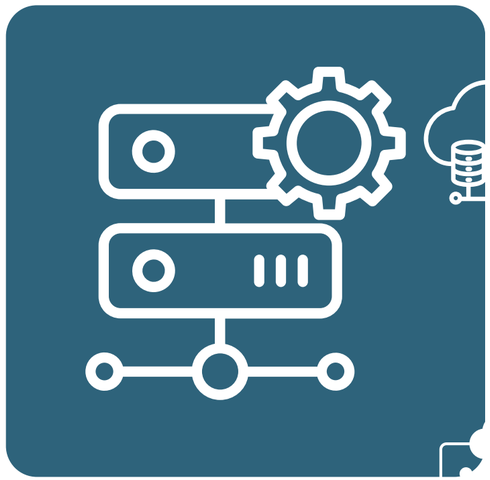
Multi-level summarization of individual research papers

# Phase 1

**Dataset: OneStopEnglish Corpus (189 articles)**

**Model Architecture**

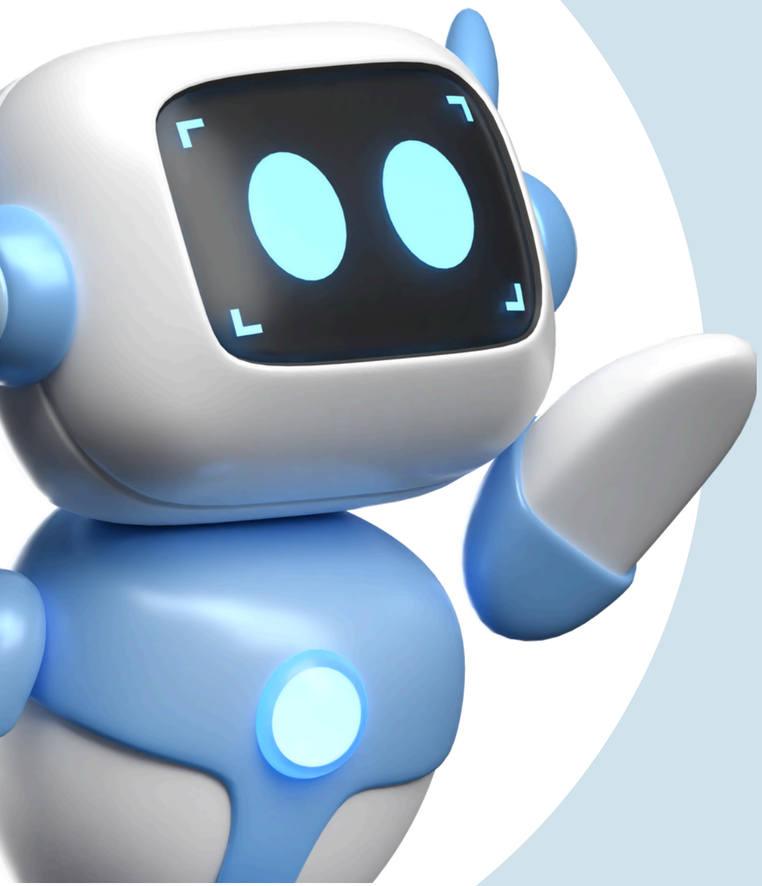
- Base BART model
- Data Collator
- Learning rate: 3e-5
- Batch size: 2 (per device)
- Gradient accumulation: 16 steps (effective batch = 32)
- Epochs: 10
- Optimizer: AdamW
- Loss function: Cross-entropy



**2.5-3 Grade Levels Simplification**

**68% Compression Rate**

**ROUGE-L scores of 0.533**



# METHODOLOGY

- Use BART (Bidirectional and Auto-Regressive Transformers) for abstractive multi-level summarization.
- Data collator for padding
- Fine-tune on general and medical texts to generate summaries for general and expert audiences.
- Control summary detail through prompt design, decoding strategies, and fine-tuning objectives.
- Phase 2 adds biomedical fine-tuning on M3 but does not include specialized simplification data.

Multi-level summarization of individual research papers

# Dataset: M3

Hierarchical biomedical summarization dataset for generating evidence-based summaries from scientific literature.

**Source:**

- Public M3 dataset (julia-nixie/m3) from biomedical meta-analyses

**Levels:**

- L1 – Document: multi-document (~3.6k words, 451)
- L2 – Sentence: abstract-level (~170 words, 315)
- L3 – Claim: evidence-level (~125 words, 381)

**Fields:**

docid, input_text, target_text, pico, input_studies

Multi-level summarization of individual research papers

# Data Cleaning & Normalization:

**M3 Dataset:**

- Removed extra spaces and non-standard punctuation.
- Unified quotes, dashes, and symbols for consistency.
- Cleaned both input_text and target_text for all three levels.

**Clean files:**

level1_clean.jsonl,

level2_clean.jsonl,

level3_clean.jsonl

Multi-level summarization of individual research papers

# Preprocessing and Preliminary Insights:

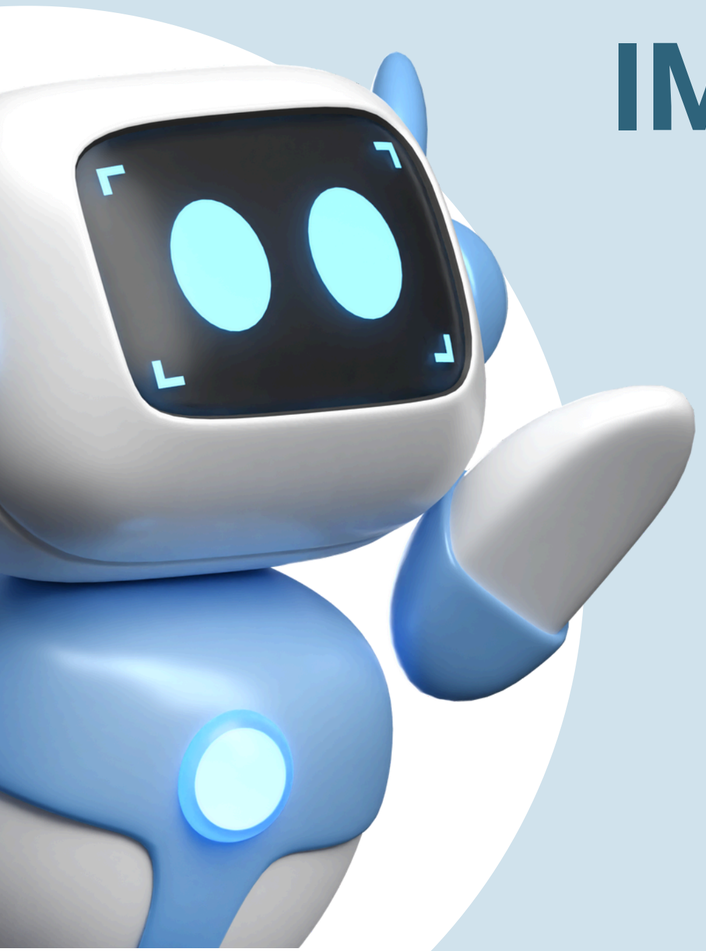
## Preprocessing:

- Merged Levels 1–3 → multi-granularity dataset
- Added control tokens: <LEVEL:DOC>, <LEVEL:SENT>, <LEVEL:CLAIM>
- Split: 974 train / 173 validation
- Tokenized with BART-base (input = 1024, target = 256)
- **Joint Dataset (OSE + M3): Later combined to test whether OSE simplification knowledge can transfer to biomedical summarization.**

## Insights:

- Texts are complex yet consistent across levels
- Level tags enrich contextual learning
- Dataset ready for fine-tuning and ROUGE evaluation

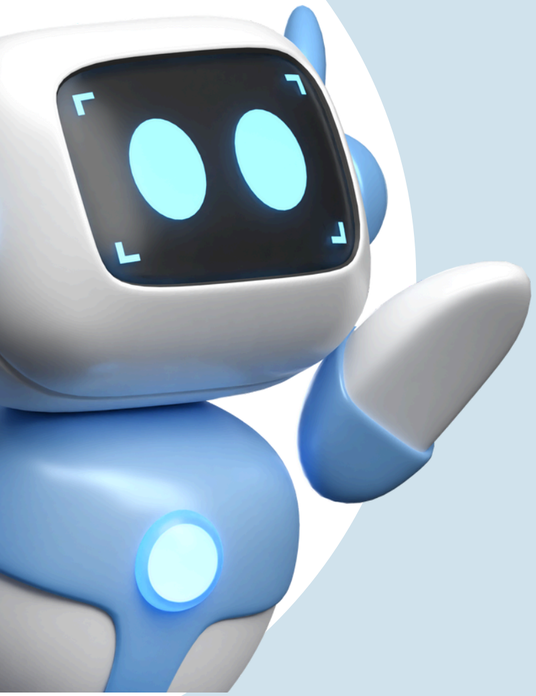
# TECHNICAL IMPROVEMENTS IN PHASE 2



- Dynamic max-length **(DOC 1024, SENT/CLAIM 256)**
- Learning-rate sweep: **1e-5, 2e-5, 3e-5, 5e-5**
- Label smoothing **(0.1)** to reduce overfitting
- Generation constraints: beam search, no-repeat n-grams, repetition penalty
- Joint **OSE+M3** dataset with task + granularity tags
- Added **special tokens + resized embeddings** for multi-task learning

# RESULTS

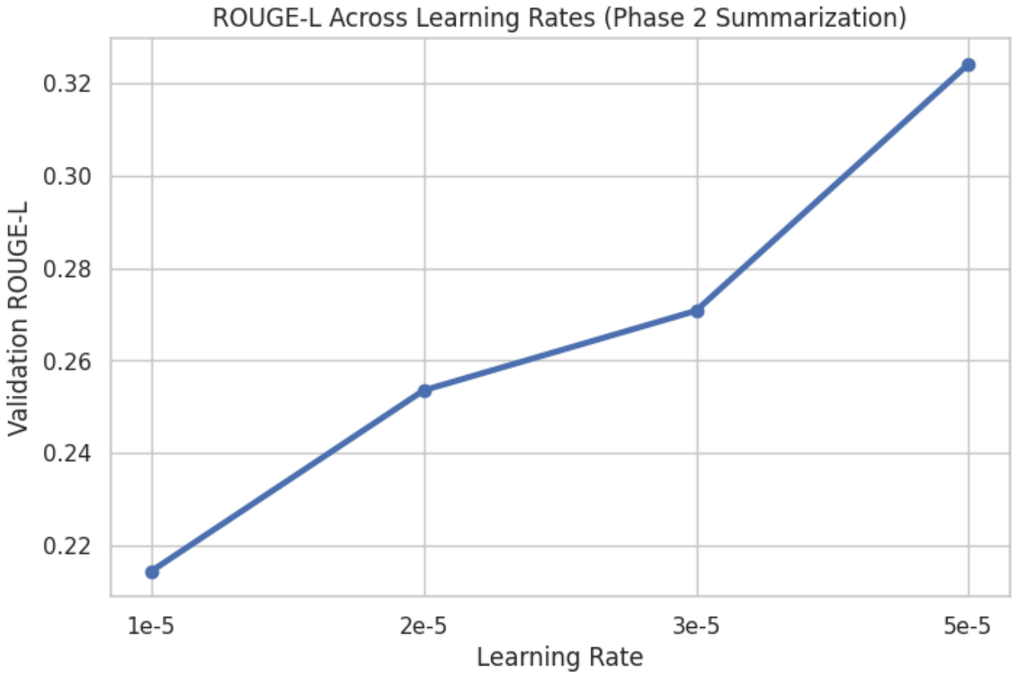
Multi-level summarization of individual research papers



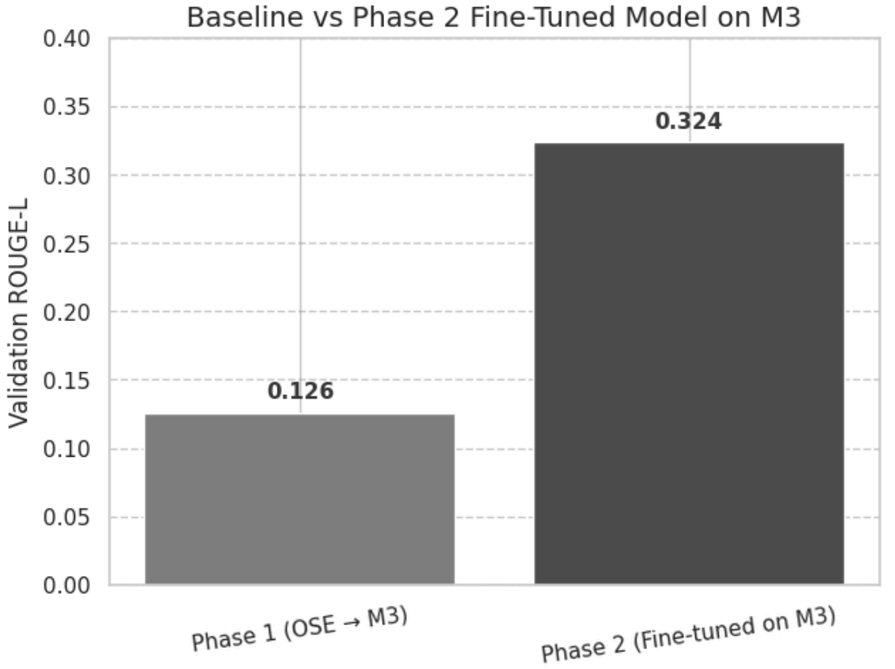
- Our model significantly outperformed the M3 baseline paper on all ROUGE metrics.
- ROUGE-L improved from **0.126 → 0.324**
- Model preserves core biomedical meaning
- **No measurable simplification** observed on M3
- Joint model adds slight structural simplification but **loses details**
- **Why Simplification Did Not Occur:**
    a. OSE teaches general simplification, not biomedical.
    b. M3 provides expert-level summaries, not simplified ones.
    c. The model learns summarization only, because M3 has no simplified targets.

| Model | ROUGE-L | Simplification | Detail Preservation |
|---|---|---|---|
| **Phase 2 (M3 only)** | 0.324 | No | Best |
| **Joint OSE+M3** | 0.297 | Some structural simplification | Loses details |

# Results




LR=5e−5 consistently dominates from epoch 3 onward.
All LRs improved significantly over the Phase 1 baseline (0.126).
Higher learning rates converge faster and reach higher ceilings.




- Phase 1 baseline evaluated on **M3 → 0.126 ROUGE-L** (very low because general-domain training cannot summarize biomedical text).
- Fine-tuning on M3 improved from **0.126 → 0.324**, a 2.5× boost in summary similarity in ROUGE-L, demonstrating successful adaptation to biomedical summarization.

# EXAMPLE (SUMMARIZATION ONLY ):

I**nput (DOC-level, truncated)**: These findings suggest that the AMD self-management program was an effective intervention to enhance well-being in older persons with poor eyesight due to AMD, particularly in those who were initially depressed. Psychosocial group intervention is a promising approach to improve the quality of life in patients suffering from ARMD. The sustained positive effects at the 6-month follow-up provide support for the effectiveness of the AMD self-management program in reducing distress and disability, improving self-efficacy, and preventing depression in poorly sighted elderly ...

**Reference Summary:** Clinical practice with patients with AMD can rely on some tailored cognitive-behavioral therapeutic protocols to improve patients' mental health, but further clinical trials will generate the necessary evidence-based knowledge to improve those therapeutic techniques and offer additional tailored interventions for patients with AMD.

**Model Summary (Phase 2 — M3 Fine-Tuned Model):** Self-management programs appear effective for older adults with AMD.

# EXAMPLE (JOINT OSE + M3 MODEL):

**LEVEL:DOC:** Genome-wide association studies (GWASs) assess correlation between traits and DNA sequence variation using large numbers of genetic variants… We sought to determine if characteristics of genomic loci associated with a trait could identify associations more likely to replicate in a second cohort…

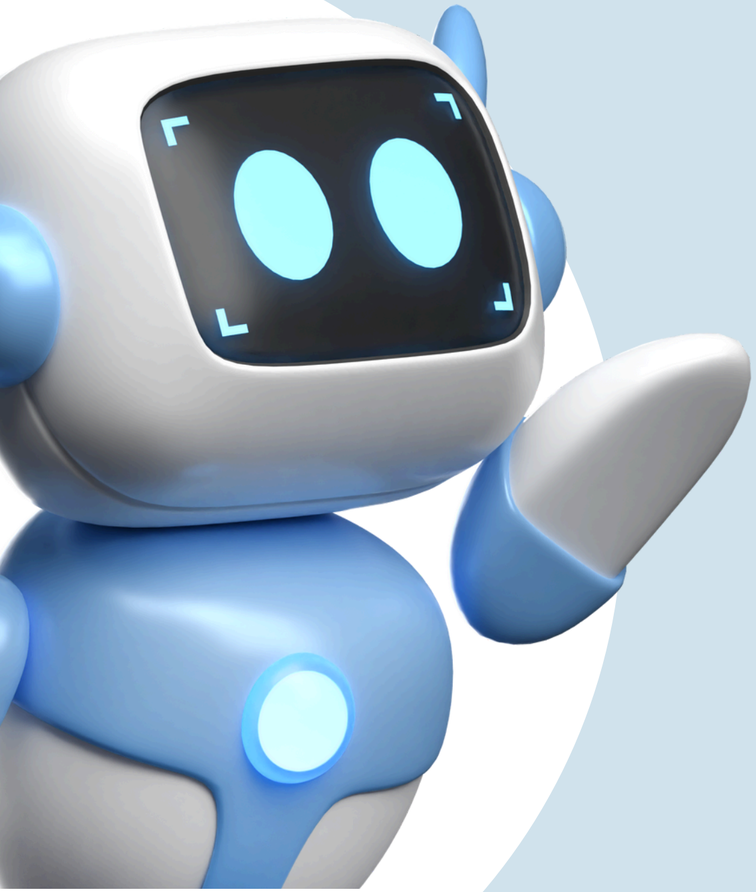
**Reference Summary:** This meta-analysis supports the association between C3 and AMD and provides a robust estimate of the genetic risk.

**Model Summary (Joint OSE + M3 Model):** This meta-analysis suggests that CFH, C3 and ARMS2 polymorphisms may be associated with AMD susceptibility.

**Information Lost:**

- Study goal (predicting replication success)
- Method details (AREDS cohort, GWAS scale)
- Motivation (replication failures)
- Nuanced interpretation of genetic risk

# DISCUSSIONS



## Conclusion

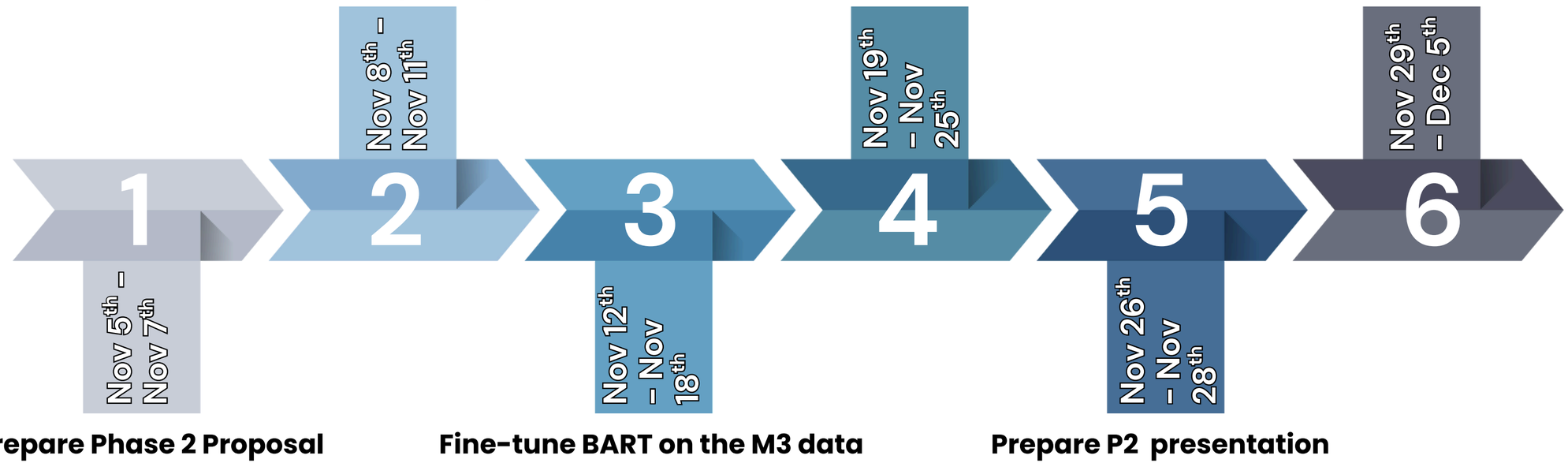
- Phase 2 improved biomedical summarization but did not simplify text.
- Joint OSE+M3 training produced shorter summaries, but not true simplification.
- Some scientific details were lost.
- Key limitation: no biomedical simplified-summary data.

## Future Works

- Collect or use other biomedical multi-level simplification data:
  BioLaySumm, MS² (Medical Student Summaries), MeQSum
- Use RLHF or reward models to target readability.
- Leverage LLM-generated simplified biomedical texts.
- Use medical ontologies to preserve accuracy while simplifying.

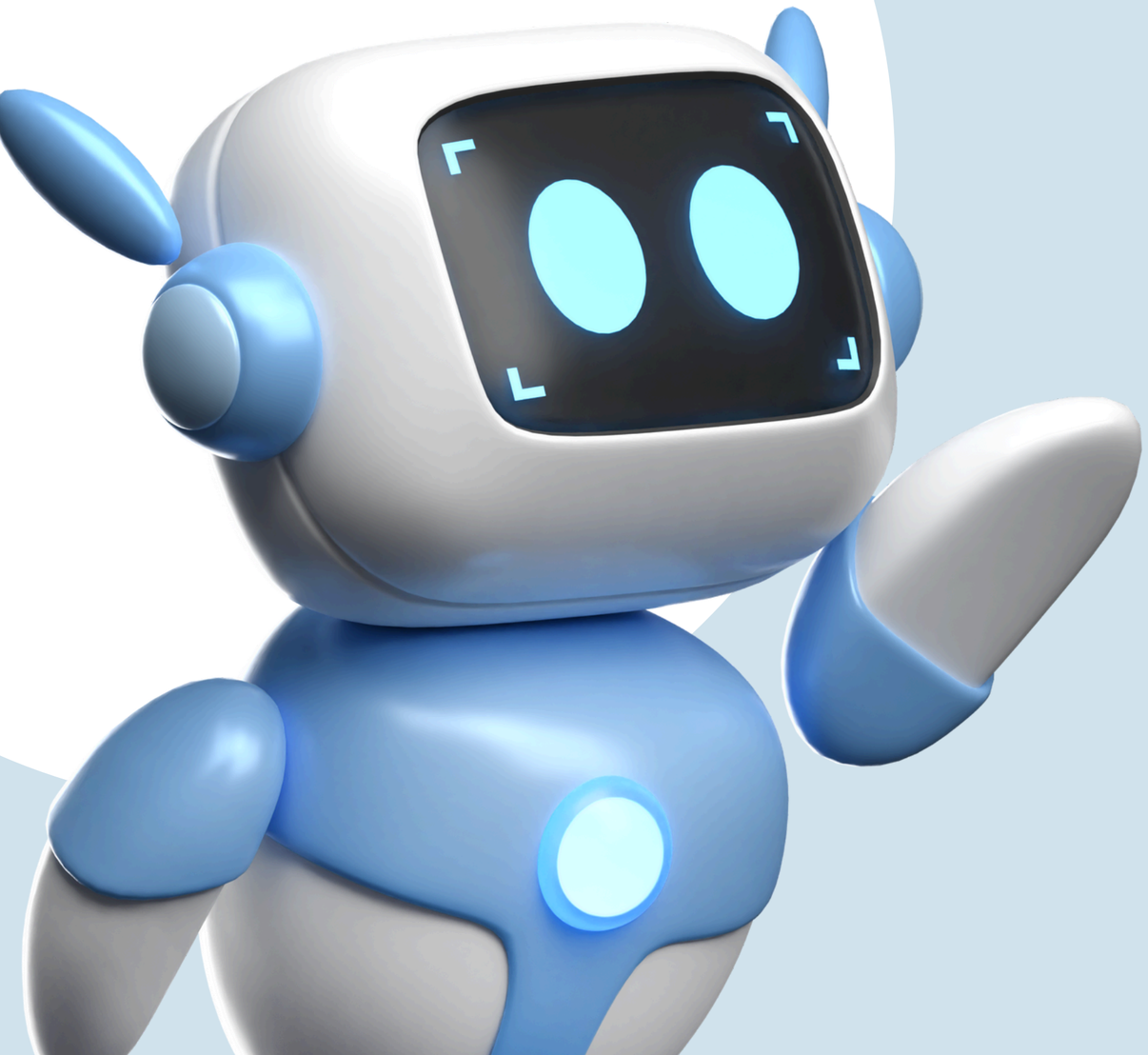
Multi-level summarization of individual research papers

# Phase 2 Timeline



# References

1. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jagsi R, Golden DW. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J Patient Exp. 2021;8:2374373521998847. doi:10.1177/2374373521998847



# THANK YOU