Exercise 1: Evidence of User Need

User Research Summary

Date	Source	Summary of findings
Sept 2025	Manual surveying	Researchers spend 23% of their time searching and filtering papers (average 8-10 hours/week). Manual literature reviews take 2-4 weeks for comprehensive coverage.
Sept 2025	Semantic Scholar, Connected Papers	Existing tools focus on citation networks but lack real-time news integration and conversational interfaces. No tool combines academic papers with industry developments.
Sept 2025	ArXiv usage statistics	200,000+ papers submitted annually in CS alone. Growth rate: 15% year-over-year. Impossible for individuals to track manually.
Sept 2025	Student/researcher pain point analysis	Graduate students report missing relevant papers due to keyword limitations. 67% say they discover important papers "too late" in their research cycle.
Sept 2025	Industry professional survey (hypothetical)	Tech professionals need to stay current but lack time. Average: 2-3 hours/week for research, want digest format with source verification.

Make a Case For and Against Your AI Feature

User Need Statement: *"How might we solve the challenge of helping AI/ML researchers, students, and professionals quickly discover, understand, and stay current with relevant academic research and industry developments without spending excessive time on manual literature searches and news monitoring?"*

AI better for:

The core experience requires recommending different content to different users.

- Different users have different research interests
- Personalization improves relevance of papers and news

The core experience requires prediction of future events.

- Predicting which emerging research areas will be important
- Identifying trending topics before they become mainstream

User experience requires natural language interactions.

- Conversational Q&A format is central to the chatbot
- Users can ask complex, nuanced questions

Need to recognize a general class of things that is too large to articulate every case.

- Vast and growing corpus of AI/ML papers (200K+ annually)
- Cannot manually catalog every research direction

Need to detect low occurrence events that are constantly evolving.

- Emerging research trends and breakthroughs
- New techniques and methodologies appearing constantly

An agent or bot experience for a particular domain.

- Specifically designed for AI/ML research domain
- Acts as research assistant

The user experience doesn't rely on predictability.

- Users expect discovery and serendipity
- Novel connections between papers add value

AI not better for:

The cost of errors is high and outweighs the benefits of a small increase in success rate.

- *Partially applicable*: Missing a paper isn't catastrophic, but hallucinated citations could mislead researchers
- **Mitigation**: Use RAG with source attribution, no generation without grounding

Conclusion Statement

We think **AI can help** solve the challenge of keeping AI/ML researchers and professionals current with relevant literature and industry developments, because**:**

1. **The domain is inherently suited for AI**: The volume, velocity, and variety of AI/ML research publications (200K+ papers annually, 15% YoY growth) make manual tracking impossible, requiring intelligent filtering and semantic understanding.
2. **Natural language interaction is essential**: Researchers think in questions, not keywords. Conversational AI allows nuanced queries like "What are the latest approaches to reduce hallucinations in LLMs?" which traditional search struggles with.
3. **Semantic understanding over keyword matching**: AI can understand conceptual relationships between papers even when they use different terminology, connecting relevant work that keyword search would miss.
4. **Real-time synthesis capability**: AI can integrate and synthesize information from multiple sources (arXiv papers + tech news) to provide comprehensive insights, while maintaining source attribution to ensure credibility.

Exercise 2: Augmentation versus Automation

Research Protocol - Contextual Inquiry Questions

For Current Workflow Understanding:

1. If you were helping to train a new coworker for a similar role, what would be the most important tasks you would teach them first?
    - "How to set up alerts, use Google Scholar, track key conferences, do manual research on the scope and find gaps in research in relevant domains."
2. Tell me more about that action you just took (conducting literature search), is that an action you repeat:
    - **Daily**
3. If you had a human assistant to work with on this task, what, if any, duties would you give them to carry out?
    - Track key conferences, do manual research on the scope and find gaps in research in relevant domains

For Concept Evaluation:

1. Describe your first impression of this feature.

    - Really useful in tracking recently published papers, summarizing research along with current trends and finding existing gaps in 90% less time than manual effort.
2. How often do you encounter the following problem: Difficulty staying current with AI/ML research and missing relevant papers?
    - Often (a few times a week)
3. How important is it to address this need or problem?
    - Very important as it saves a lot of time in literature review.

Augmentation vs Automation Decision

We feel **ResearchAI should focus on automation and not just augmentation because:**

Rationale:

- Researchers need not **evaluate** papers themselves manually. We now have an agent to do so.
- The chatbot should **surface, summarize** and **generate.**
- Users maintain control and agency in their research process
- Builds trust through transparency (citations, sources)

Exercise 3: Design Your Reward Function

Based on our requirement for **"top-k relevant papers"** using FAISS, we're optimizing for **precision over recall**.

Reward Function Template

Prediction

	Positive	Negative
Positive (Reference)	**True Positive** **Example 1:** User asks about "transformer attention mechanisms". System returns seminal "Attention Is All You Need" paper **Example 2:** Query "latest LLM alignment techniques". Returns recent RLHF papers from 2024 **Example 3:** "Computer vision for medical imaging". Returns relevant papers combining	**False Negative** **Example 1:** User asks about "BERT". The system missed the original BERT paper because the query was too generic. **Example 2:** Query uses synonym "neural machine translation". System doesn't retrieve."sequence-to-sequence" papers. **Example 3:** Recent breakthrough paper

	both domains.	published yesterday not yet in database
Negative (Reference)	**False Positive** **Example 1:** User asks about "GPT applications". System returns generic NLP papers not specifically about GPT **Example 2:** Query "image generation". Returns papers about GANs when user wanted diffusion models **Example 3:** Keyword match on "reinforcement" returns papers about structural reinforcement (engineering), not RL.	**True Negative** **Example 1:** User asks about "computer vision" related questions. System correctly excludes NLP-only papers **Example 2:** Query "2024 papers". The system correctly filters out papers from 2020. **Example 3:** Physics papers correctly not returned for ML query.

Optimization Decision

Our AI model will be optimized for precision because:

- **Users value relevance over completeness**: When a researcher asks a question, they want the top 5-10 highly relevant papers, not 100 marginally related ones.
- **Limited context window**: Users can only review a limited number of papers in a session. Irrelevant results waste time and reduce trust.
- **FAISS top-k retrieval aligns with precision**: Fetching top-k results inherently prioritizes the most semantically similar (most relevant) documents
- **Trust and credibility**: False positives (irrelevant papers) directly harm user experience and trust. Missing one paper (false negative) can be recovered through follow-up queries
- **Conversational refinement**: Users can iteratively refine queries if they don't find what they need, but can't easily filter out irrelevant results from a large set

We understand that the tradeoff for choosing this method means our model will:

- **Potentially miss some relevant papers** that use different terminology or are at the boundary of relevance.
- **Require users to rephrase or refine queries** if their initial search doesn't surface what they need.
- **Need robust synonym/semantic understanding** to ensure we don't miss papers due to vocabulary mismatch
- **Benefit from query expansion and conversation history** to improve coverage over multiple interactions

Mitigation Strategies:

1. Use semantic embeddings (not just keyword matching) to capture conceptual similarity
2. Allow users to provide feedback to improve relevance over time

Exercise 4: Define Success Criteria

Success Metrics Framework

Version 1: User Experience Metric

If user satisfaction score (measured via post-query thumbs up/down + optional feedback)

for ResearchAI's paper recommendations and summaries

drops below 80% positive feedback rate (calculated weekly)

We will conduct user interviews to identify failure modes, review false positive cases, and adjust retrieval thresholds or embedding models within 5 business days.

Version 2: Retrieval Quality Metric

If average relevance score of top-5 retrieved papers (human-evaluated on sample queries)

for ResearchAI's semantic search and recommendation engine

drops below 4.0 out of 5.0 (evaluated monthly on 100 random queries)

We will retrain or fine-tune the embedding model, audit the vector database for quality issues, and review query preprocessing logic within 2 weeks.

Version 3: System Performance Metric

If 95th percentile response time for query-to-answer

for ResearchAI end-to-end system (including FAISS retrieval + LLM generation)

goes above 8 seconds

We will investigate infrastructure bottlenecks, optimize database queries, implement caching for common queries, or scale up resources within 48 hours.

Version 4: Data Freshness Metric

If the time lag between paper publication on arXiv and availability in ResearchAI

for the arXiv ingestion pipeline

goes above 48 hours for 90% of papers

We will increase pipeline frequency, investigate API rate limiting issues, and add alerting for ingestion failures immediately.

Version 5: Source Attribution Metric

If the percentage of AI-generated responses that include proper source citations

for ResearchAI's LLM-generated summaries and answers

drops below 95%

We will review RAG grounding mechanisms, adjust prompt engineering to enforce citations, and implement automated citation validation within 1 week.

Statement Iteration Checklist

For each version, evaluate:

Is this metric meaningful for all of our users?

- Students, researchers, and professionals all care about relevance and speed
- All users need accurate citations for credibility

How might this metric negatively impact some of our users?

- Optimizing for speed might sometimes reduce answer quality
- Focusing only on recent papers might miss foundational work
- *Mitigation*: We need to balance multiple metrics, not just one

Is this what success means for our feature on day 1?

- Day 1: Focus on basic functionality, relevance, and system stability
- Initial thresholds can be more lenient (e.g., 75% satisfaction, 10-second response time)

What about day 1,000?

- Day 1,000: Expectations higher - 85%+ satisfaction, <5 second response time
- Need personalization, advanced query understanding, and proactive recommendations

Final Version (Primary Success Metric)

If the composite quality score (weighted average: 40% user satisfaction + 30% retrieval relevance + 20% response time + 10% citation accuracy)

for ResearchAI's overall system performance

drops below 80/100 (evaluated weekly with monthly deep dives)

We will discuss with the team within 3 business days to identify root causes, prioritize fixes based on impact, implement corrective actions, and re-evaluate within 2 weeks

Schedule Regular Reviews

Success Metric Review Cadence:

- **Weekly automated dashboards**: Track all metrics, alert on threshold breaches
- **Bi-weekly team sync** (30 min): Review trends, discuss minor adjustments.
- **Monthly deep dive** (2 hours): Comprehensive analysis, user feedback review, strategic adjustments.
- **Quarterly retrospective** (half day): Evaluate long-term trends, consider major pivots, update success criteria.