Option Number 1

Self-Assessment Report

Context Windows in Practice
LLMs and Multi-Agent Systems Course

Project Name:	Context Windows Lab - LLM Experiments
Repository:	https://github.com/OmerCaplan/context-windows-lab
Team:	OmerAndYogever
Students:	Omer Caplan, ID: 208753665 Yogev Cuperman, ID: 207540550
Assessment Date:	December 4, 2025
Self-Grade:	91/100

Category-by-Category Assessment

Category	Weight	My Score	Final Score
Project Documentation	20%	18/20	18
README & Code Documentation	15%	14/15	14
Project Structure & Code Quality	15%	14/15	14
Configuration & Security	10%	9/10	9
Testing & QA	15%	13/15	13
Research & Analysis	15%	14/15	14
UI/UX & Extensibility	10%	9/10	9
TOTAL	**100%**		**91**

Grade Level: 90-100 (■■■■■ - Excellent)

With a self-grade of 91, we expect thorough and meticulous scrutiny. The evaluators will check full compliance with all criteria and expect high quality standards throughout. This grade reflects excellent work at a high academic level with comprehensive documentation, real research with statistical analysis, and quality visualizations.

Justification for Self-Assessment

Strengths

The project demonstrates strong technical implementation and comprehensive documentation across multiple dimensions. The formal PRD provides clear project requirements, success metrics (KPIs), detailed functional and non-functional requirements, and system architecture documentation. The comprehensive README offers detailed step-by-step installation instructions, troubleshooting guidance, technical details with code examples, and documentation of experimental results including statistical analysis. Code quality is excellent with proper documentation through docstrings in all functions, clear module separation (experiments/, utils/, cli.py), and consistent naming conventions. The modular structure follows best practices with src/, tests/, notebooks/, and docs/ directories. Security and configuration management are exemplary with proper use of .env files, no hardcoded credentials, and comprehensive .gitignore coverage. The research component is strong with four systematic experiments investigating context window phenomena. The Context Engineering experiment yielded statistically significant results ($F=17.29$, $p=0.0007$, $\eta^2=0.866$) demonstrating that the SELECT strategy significantly degrades performance compared to other approaches. The RAG experiment demonstrated dramatic efficiency gains (92% token savings, 97% latency reduction) while maintaining accuracy. All experiments include proper statistical analysis with t-tests, ANOVA, correlations, and effect size calculations.

Weaknesses

The main limitation is that Experiments 1-3 showed ceiling effects (100% accuracy) because Claude Haiku handles the tested scenarios exceptionally well. This prevented demonstration of the expected "Lost in the Middle" accuracy degradation. While this is itself an interesting finding about modern model capabilities, it means we could not validate the original hypothesis from the literature. Additionally, rate limiting issues affected data collection for Experiment 2, resulting in fewer data points than planned. The test coverage could be improved to reach 80%+ with more integration tests. Some experiment files are lengthy and could be further decomposed for better maintainability.

Investment

Significant effort was invested in creating a complete experimental framework with four specialized experiments, implementing proper statistical analysis with effect size calculations, and integrating sentence-transformers for semantic analysis. The work includes comprehensive documentation (PRD, README with detailed sections), a test suite with unit tests and mocks, and proper project structure following Python best practices. The development process involved debugging rate limiting issues, implementing retry logic with delays, ensuring proper statistical validity, and creating professional visualizations with matplotlib. This represents substantial work beyond a minimal implementation and demonstrates commitment to delivering a polished, research-quality application with reusable components.

Innovation

The project demonstrates several innovative aspects. Testing both "easy" and "hard" parameter configurations revealed that modern models like Claude Haiku may not exhibit traditional context window limitations documented in older research - this is itself a novel finding. The discovery that the SELECT strategy significantly fails ($p<0.001$) while COMPRESS and WRITE maintain accuracy provides actionable insights for building production multi-step agents. The comprehensive statistical framework including effect sizes (Cohen's d, η^2) goes beyond basic accuracy reporting. The modular experiment framework using BaseExperiment inheritance allows easy extension for future experiments. The CLI interface with Rich formatting demonstrates attention to user experience.

Learning

The project demonstrates learning in multiple areas: LLM context window behavior and limitations, statistical experimental design with proper hypothesis testing, and professional Python packaging practices. Key learnings include understanding the importance of parameter tuning to reveal model limitations, the challenges of API rate limits in experimental research, proper statistical analysis including effect size calculations for meaningful comparisons, and the practical trade-offs between different context management strategies. The finding that modern models may have improved beyond limitations documented in earlier research papers (Liu et al., 2023) represents genuine empirical discovery. The iterative process of adjusting experimental parameters to reveal phenomena demonstrates scientific methodology.

Academic Integrity Declaration

We hereby declare that:

• Our self-assessment is honest and truthful
• We checked our work against all criteria before determining the grade
• We are aware that a high self-grade will lead to more rigorous review
• We accept that the final grade may differ from our self-assessment
• This work is the product of our work (of the team) and we are responsible for all software

Student 1: Omer Caplan (ID: 208753665)
Signature: Omer
Date: December 4, 2025

Student 2: Yogev Cuperman (ID: 207540550)
Signature: Yogev
Date: December 4, 2025