Comparative Analysis of Multi-Agent LLM Systems for Solving Polish Matura Physics Exams

Piotr Wróbel
Warsaw University of Technology
Warsaw, Poland

Abstract

Large Language Models have gained widespread recognition since OpenAI released their revolutionary model, ChatGPT 3.5. Since then, many new approaches have emerged to improve the capabilities and accuracy of these models for different tasks. One such method involves using multi-agent conversations. This article compares two multi-agent setups designed to solve the Polish standardized high school exam in physics. Comparative benchmarks were performed on several real final exams published by the Polish Central Examination Board (pl. CKE — Centralna Komisja Egzaminacyjna). The study employed ChatGPT-4 Turbo and the AutoGen framework. Benchmarks covered a total of 90 tasks from three Polish Matura physics exams (editions: 2018, 2019, 2023). The simpler multi-agent systems achieved an average score of 76.1%, while the more complex systems averaged 85.6%.

Introduction

Recent advances in AI, particularly in natural language processing and generative models, have set new benchmarks for tackling demanding STEM challenges. Standardized exams, such as the Polish Matura exam, serve as ideal evaluations for both human students and AI systems. This study focuses on the diverse physics tasks in the Matura exam.

Inspired by the impact of ChatGPT, we investigated multi-agent systems that decompose complex problems into manageable subtasks, leveraging specialized agents. Our experiments employed OpenAI's LLM with the AutoGen framework (v0.2.2) to implement and assess these systems.

The study aims to:

- Evaluate overall performance on standardized exams,
- Analyze detailed results across various physics topics and task types,
- Compare AI performance with that of students.

Evaluation Environment

- Platform & Tools:
  - Experiments utilized ChatGPT-4 (gpt-4-1106-preview) for its cost-effectiveness and accessibility.
  - The AutoGen framework and Group Chat Manager streamlined multi-agent collaboration.
- Approaches:
  - Simple System (4 Agents):
    - Admin: Initiates and concludes conversations.
    - Scientist: Solves physics tasks and writes code.
    - Manager: Oversees discussions and validates solutions.
    - Executor: Runs the Python code produced.
  - Complex System (7 Agents):
    - Maintains Admin and Executor roles, with added specialized agents:
      - Translator: Converts Polish task descriptions to English.
      - Physics Professor: Handles preliminary calculations and proposes solutions.
      - Python Programmer: Develops code from provided instructions.
      - Manager: Ensures consensus on a consistent final answer.
      - Reviewer: Critically assesses and refines the solution.

Note: In the complex system, specialized agents may independently achieve final outcomes without input from every member.

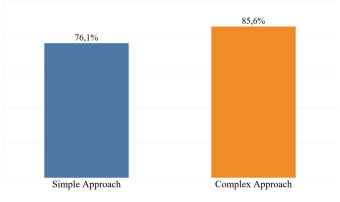

Fig. 1. Average results for both solutions

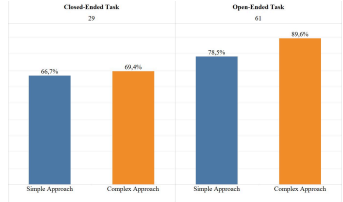

Fig. 2. Results per task type

Results and Findings

- Evaluation Context:
  - Experiments ran from December 2024 to February 2025 and were graded using official CKE guidelines.
- Overall Results:
  - Simple Approach: 76.1% average score
  - Complex Approach: 85.6% average score
  - The complex method consistently outperformed the simple approach across all exam editions (2018, 2019, and 2023), achieving an average improvement of 12.28% and scoring within at least the 86th percentile, with top performance reaching the 99th percentile.
- Detailed Results – Task Type:
  - Open-Ended Tasks (61 tasks): Complex approach scored 11.1 percentage points higher
  - Closed-Ended Tasks (29 tasks): Difference of 2.6 percentage points in favor of the complex system
  - For most subtypes, scores were higher or comparable, with only a marginal difference (1 point) in "True or False" tasks.
- Detailed Results – Task Topic:
  - The complex approach achieved higher or similar scores across topics.
  - Key Findings:
    - Harmonic Motion and Mechanical Waves: Improvement of 22.2 percentage points
    - Direct Current (DC): No change (70.6% for both systems)

This summary encapsulates the primary performance advantages of the complex multi-agent solution over the simpler approach in the context of Matura physics exam results.

Key Findings:

- Impressive Performance:
  - Multi-agent systems achieved scores comparable to the top 15% of human test-takers on the Polish Matura physics exam.
  - Some configurations reached up to the 99th percentile.
- Impact of Specialization:
  - The complex model scored 85.6% compared to 76.1% for the simpler approach.
  - Better performance was noted on open-ended tasks, highlighting the strength of these systems in complex reasoning.
- Trends & Future Directions:
  - A slight performance decline was observed on the 2023 exam, potentially linked to the training data cut-off.
  - Further advancements could be achieved with enhanced agent specialization, prompt optimization, and integrated vision capabilities.
  - Future studies should explore broader subjects and diverse standardized tests to validate these findings.

Exam Edition	Simple Approach		Complex Approach	
	Result	*Percentile*	*Result*	*Percentile*
2018	76,7%	96th	88,3%	99th
2019	80,0%	91th	90,0%	97th
2023	71,7%	86th	78,3%	90th

Table 1. Results by exam edition

Topic	Simple Approach Result	Complex Approach Result
Motion of a Material Point	83,3%	94,4%
Mechanics of a Rigid Body	73,9%	78,3%
Mechanical Energy	77,8%	88,9%
Gravitation	87,0%	100,0%
Thermodynamics	66,7%	80,0%
Harmonic Motion and Mechanical Waves	77,8%	100,0%
Electric Field	85,7%	100%
Direct Current (DC)	70,6%	70,6%
Magnetism and Magnetic Induction	64,3%	71,4%
Electromagnetic Waves and Optics	64,3%	71,4%
Atomic Physics	81,8%	86,4%

Table 2. Results by topic