Elektrotechnik
und Informatik
Faculty of Electrical Engineering
and Computer Science

Performance of Large Language Models in a Computer Science Degree Program of a University of Applied Sciences

AI for AI Education workshop at ECAI 2023 in Kraków, Poland

Tim Krüger, Michael Gref - September 30th, 2023

Niederrhein University of Applied Sciences

Motivation

- LLMs like ChatGPT have an ubiquitous presence
- They provide new ways and possibilities to teach and learn
- Useful for both students and instructors
  - Educational aids and “learning buddies”
  - Generation of supplementary materials
  - Aid in the assessment process
  - …
- **But** to accurately assess the benefits we need to know the performance capabilities and possible limitations (in our curriculum)

Methodology

Methodology

- Idea: Evaluation of various Large Language Models (LLMs) using academic content drawn from our computer science bachelor’s degree program
  - 6 LLMs tested
- Objective: Determine model performance
  - Average grade
  - Highest and lowest-scoring modules
  - Topic affinities
  - Is a completion of the degree program for the models possible?
- 10 modules tested (8 written, 2 oral)
- Data set included past and sample exams, questionnaires, and practice exercises

Workflow

- Obtaining a written approval for the module
  - Ideally, including approval for past exams
  - With provided sample solutions
- Acquisition of usable materials
- Digitization of all materials
- Creation of prompts from these materials
- Prompting of all LLMs
- Documentation of all results and any special occurrences
- Evaluation and assessment of the performance
  - Additional effort involves recompiling and testing program code

Evaluation

- Prompting: All models were prompted once with a generic pre-prompt setting exam context and response expectations
  - Only once due to severe time limitations
- Modules with written exams: Prompts with past and sample exams
  - Multimedia input: Where feasible we transformed tasks into a suitable textual format
    - Tables → Markdown; Diagrams → UML …
- Modules with oral exams: Prompts with professor-approved questionnaires
- Evaluation based on the system and point allocation provided by the supervising professor

Pre-prompt

I am now going to ask you a few questions from a hypothetical [insert topic or subject] exam of an undergraduate computer science degree program. I want you to answer the questions to the best of your knowledge and capabilities. Please answer briefly and concisely unless I explicitly ask for a more detailed answer! Please answer purely in continuous text or bullet points. If output in chart or table form is desired, I will let you know.

Tested LLMs

- ChatGPT-3.5 (OpenAI, 175B parameters)
  - chat.openai.com
- GPT-4.0 (OpenAI)
  - platform.openai.com (Playground)
- Bing Chat with AI (Microsoft, GPT-4 Foundation LLM) → “BingAI”
  - bing.com
- StableLM-Alpha (StableAI, 7B parameters)
  - github.com/Stability-AI/StableLM
- LLaMa (MetaAI, 7B + 65B parameter variants)
  - With 4-bit integer quantization → Perplexity increase of only 4.23%
  - github.com/ggerganov/llama.cpp

Results

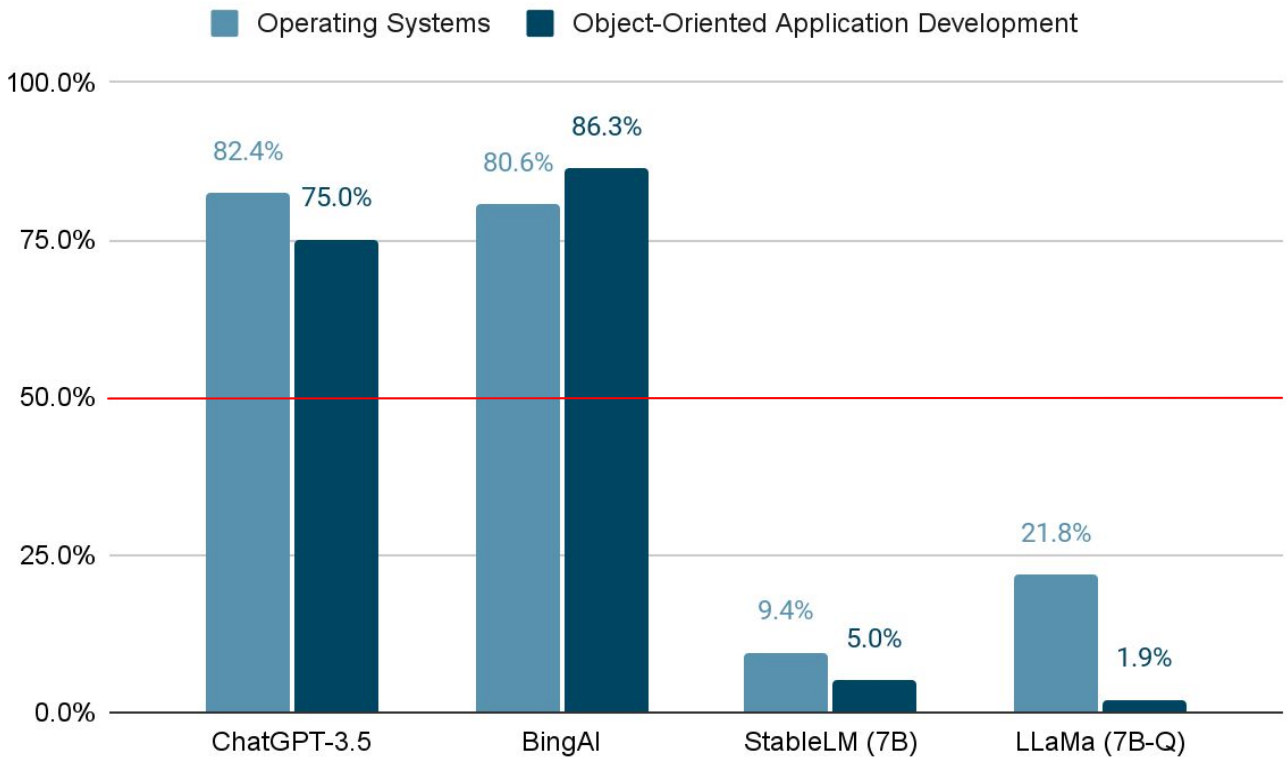
Disclaimer

- We got access to GPT-4.0 midway through the project
  - Use within our research group was monetary limited
- We weren’t able to test StableLM and the LLaMa iterations on all modules due to time constraints
  - But we gathered enough information to make a judgement about these LLMs
- Collected 40 data points
  - Data point := Performance of one LLM or LLM version in one module
- All following results are given in percent (%)
  - Standardized because each exam has a different max. score

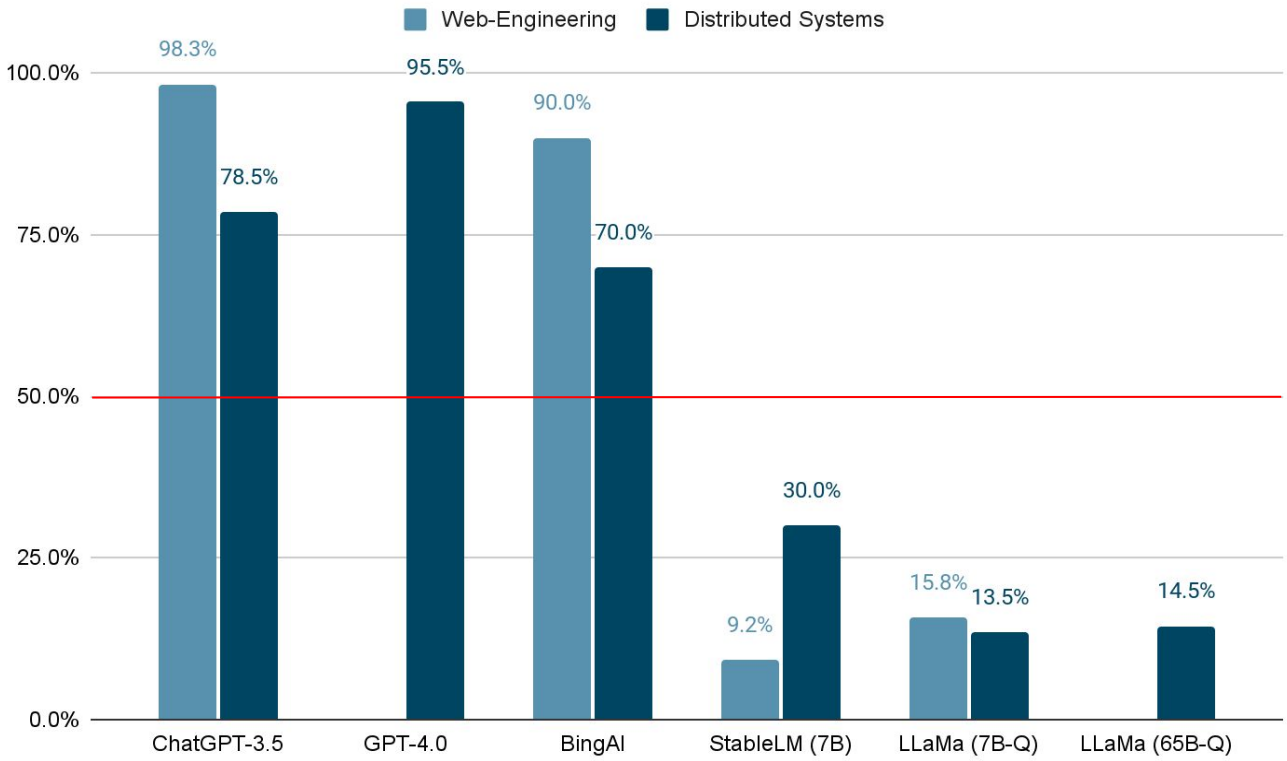
1st semester

- Unfortunately, we did not receive approval for any 1st semester module

2nd semester

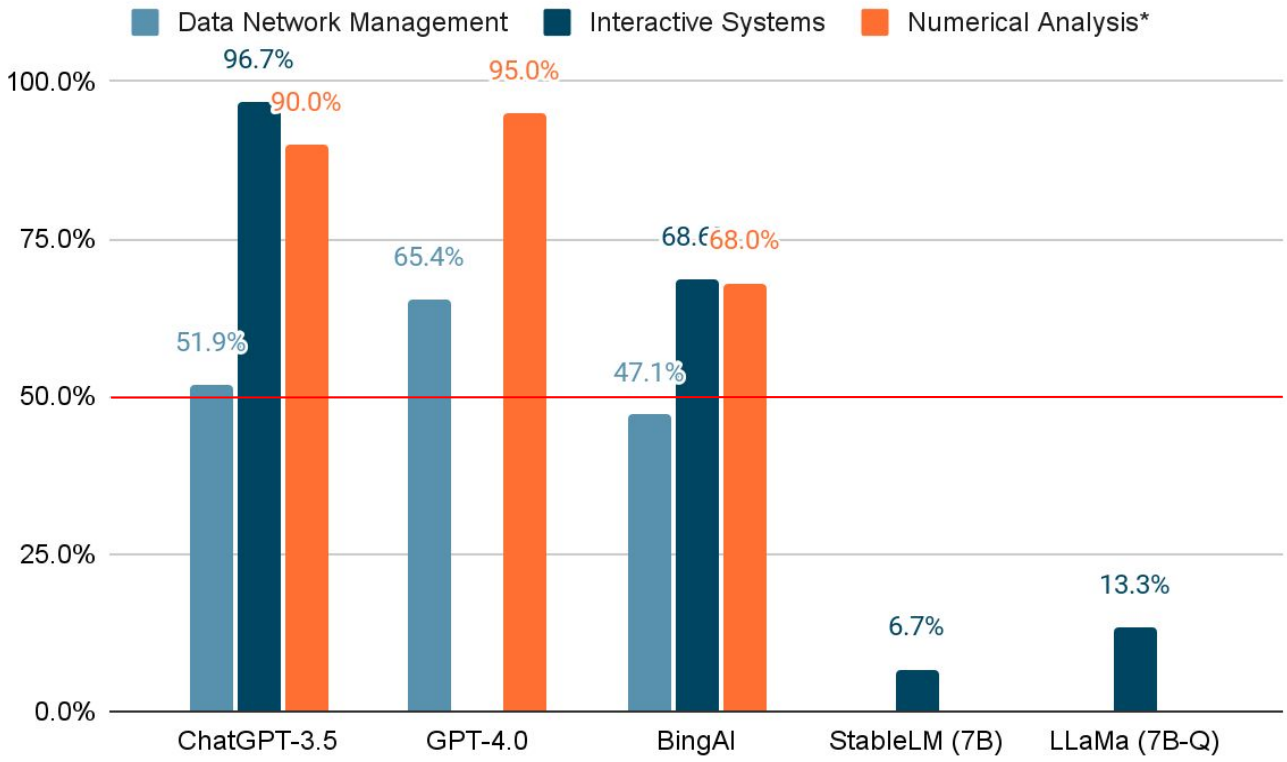


3rd semester



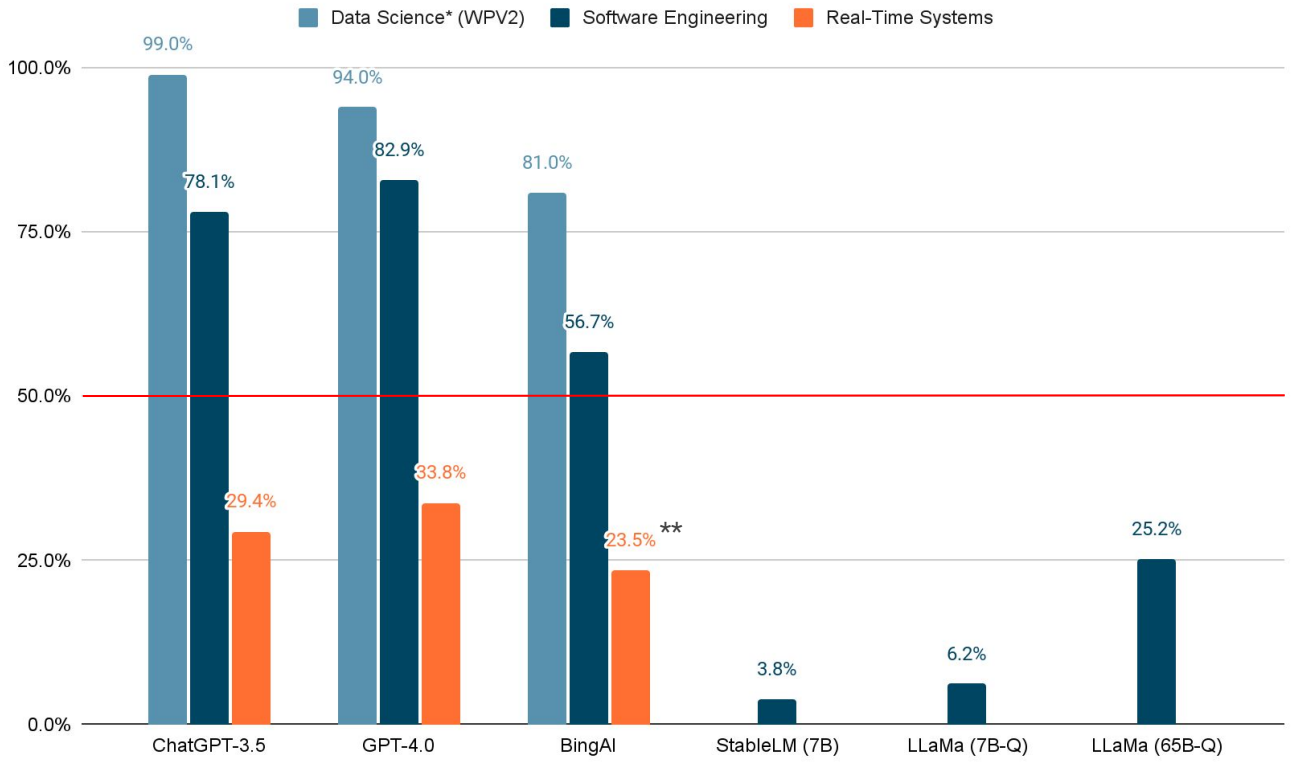
* Simulation of an oral exam

4th semester



5th semester

* Simulation of an oral exam
** BingAI copied one exercise solution 1 to 1 from studocu.com and was unable to explain it. If we would discard it because of cheating → **19.1%**



Performance summary

Model	Average score	# Modules	Passed / Failed
GPT-4.0	80.2%	6	5 / 1
ChatGPT-3.5	79.9%	10	9 / 1
BingAI	68.4%	10	8 / 2
LLaMa-65B-Q	20.0%	2	0 / 2
LLaMa-7B-Q	12.3%	6	0 / 6
StableLM-7B	10.8%	6	0 / 6

Conclusion

- Results align with existing research, highlighting strong LLM performance but limitations in some areas (e.g. mathematical computations)
  - GPT models showed strong performance
  - Smaller LLMs had significant performance deficiencies
- A prevalent worry is the potential for essays to progressively lose significance
  - Comprehensive blueprint for our whole curriculum remains elusive at this point
- But this is only a question of time
  - Patterns of past exams often remain unchanged
  - Advancing abilities of LLMs → Additions of systems like WolframAlpha?
  - Sometimes exams even allow aids → potentially perfect “Cheat Sheets”
- Compels us to reconsider and construct robust computer science examination methods

Elektrotechnik
und Informatik

Faculty of Electrical Engineering
and Computer Science

Thanks for your attention!

Tim Krüger - tim.krueger@stud.hn.de
Michael Gref - michael.gref@hs-niederrhein.de