Building MultiphaseHUB: Autonomous Data Extraction from Scientific Literature Using PlotPick and Large Language Models

Erik McKeithen, Summer 2026 NSF REU in Energy Security
Advisor: Dr. Ying Sun, Department of Mechanical Engineering and Engineering Science

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Introduction

The Challenge

Multiphase transport data are highly fragmented across academic literature, difficult to compile

The Bottleneck

Lack of accessibility hinders the use of data-driven modeling and machine learning

The Solution

MultiphaseHUB: a centralized, open-access data hub designed to aggregate, standardize, and share multiphase flow data

Objectives

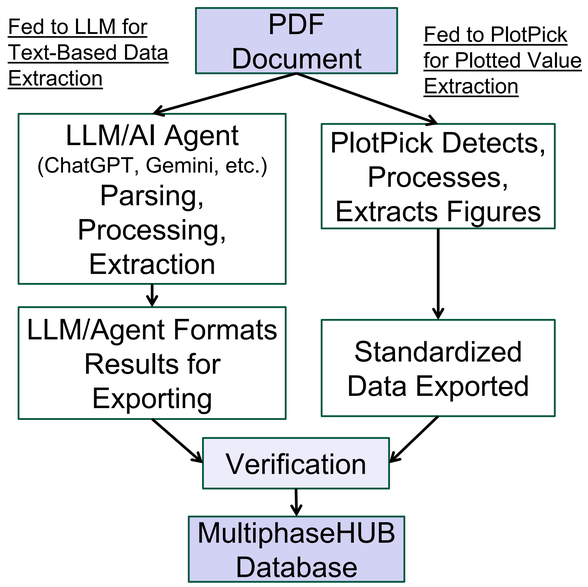
Multiphase Flow Database Creation

- **Centralize Data:** Build an open-access database for experimental multiphase flow data
- **Integrate Analytics:** Implement machine learning and interactive visualization tools to accelerate data interpretation

Automatic Data Extraction

- **Automate Extraction Workflow:** Utilize AI agents, VLMs (vision-language models) , and LLMs (large language models) to autonomously extract data from text, tables, and plots in literature

Data Extraction Pipeline



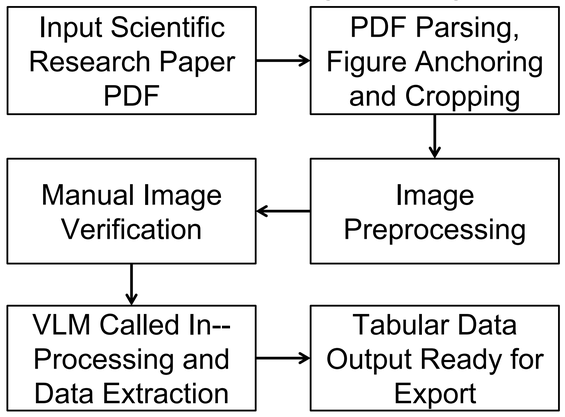
PlotPick for Data Extraction

LLM Performance and Limitations

- Advanced LLMs (Gemini Pro, ChatGPT High Reasoning, etc.) pulled text-based data with negligible error, but struggled to accurately extract visual data from plots

PlotPick for Visual Data Extraction

- Tool designed to pull structured data out of scientific plots and figures using VLMs



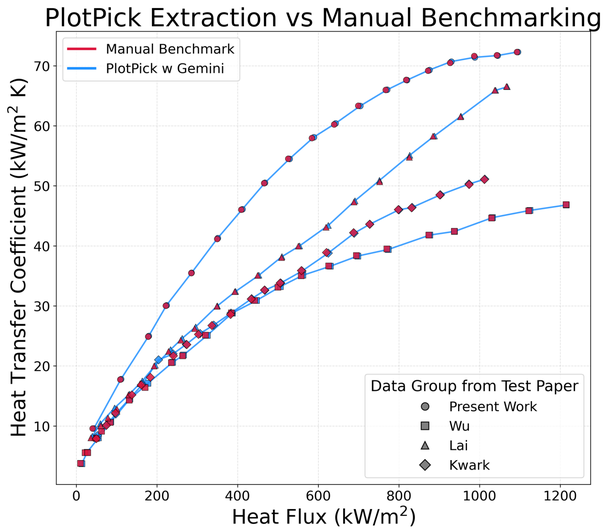
PlotPick Test Case

Testing PlotPick w/ Gemini Pro

PlotPick was used to extract a line/scatter plot from a research paper containing boiling curve data. Gemini Pro served as the VLM.

Results of This Test Case

- **Image Extraction:** Several figures in the test paper were extracted successfully; verification step allowed for individual figure selection
- **Verification:** Extracted line-and-scatter-plot data was compared directly against manual benchmarks to verify accuracy.
- **High Precision:** PlotPick achieved **< 1% overall error** and successfully detected **> 95% of plotted points**.
- **Data Reliability:** The extraction process produced **zero hallucinated data points**



PlotPick with Gemini Pro vs Manually Extracted Data for a PDF Figure (Huang et al., 2023)

PlotPick vs Manual Benchmark

Calculated using linear interpolation within matching categories

Overall Error — Data From All Four Curves Pooled

Metric	Value	Metric	Value
PlotPick # of Points	80	Actual # of Points	84
Coverage	95.2%	Mean Bias	-0.005
MAE	0.149	RMSE	0.256
MAPE	0.89%	Max Absolute Error	1.621

multiphasehub.org

Conclusions and Future Work

Conclusions

- LLM is effective for text-based extraction
- PlotPick shows significant promise for bulk visual academic data extraction, especially when paired with a powerful VLM
- Scale and scope of extraction experiments are very limited at present; further testing with PlotPick is needed

Future Work

Scale Testing: Expand the literature corpus for rigorous error and accuracy testing, and test PlotPick's reliability on complex figures beyond standard scatter plots.
Integrate Analytics: Implement machine learning and visualization tools to accelerate data interpretation.

References

T. Carstensen, "PlotPick: AI-powered batch extraction of numerical data from scientific figures," *arXiv preprint arXiv:2605.06021*, 2026.

X. Huang *et al.*, "Pool boiling experiment characteristics on the pure copper surface," *Experimental and Computational Multiphase Flow*, vol. 5, no. 2, pp. 192-198, 2023.