Synthetic Homes: A Generative AI Framework for Urban Energy Data

Jackson Eshbaugh [1], Chetan Tiwari [2 3], Jorge Silveyra [1]

[1]Department of Computer Science, Lafayette College
[2]Departments of Geosciences & Computer Science, Georgia State University
[3]Center for Disaster Informatics & Computational Epidemiology, Georgia State University

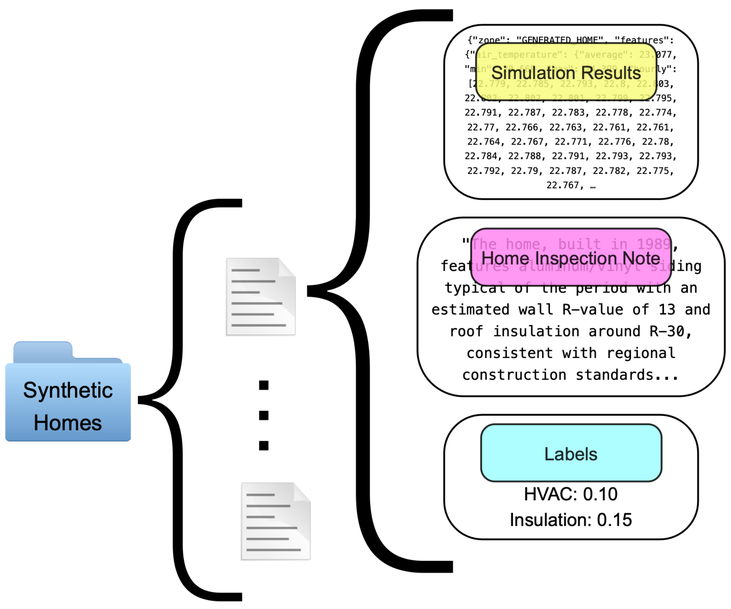
Motivation

Electricity consumption in the United States has risen sharply in recent decades, intensifying pressure on infrastructure, costs, and emissions. Accurate energy models are critical for planning, but these models often require detailed building data (materials, floor plans, microclimate conditions) that are expensive, scarce, or restricted by privacy concerns. These barriers hinder scalability and limit the effectiveness of existing modeling approaches.

Many approaches, such as synthetic data generation and digital twins, rely on costly simulation tools or extensive sensor networks. Recent advances in generative AI provide a new path: producing large volumes of low-cost, multimodal data (text, images, tabular) tailored to specific urban contexts.

We present a modular generative AI framework for producing synthetic urban building data that integrates pretrained language and vision models with energy simulation tools. Our pipeline enables scalable, low-cost data generation suitable for urban energy modeling, policy evaluation, and decision-making.

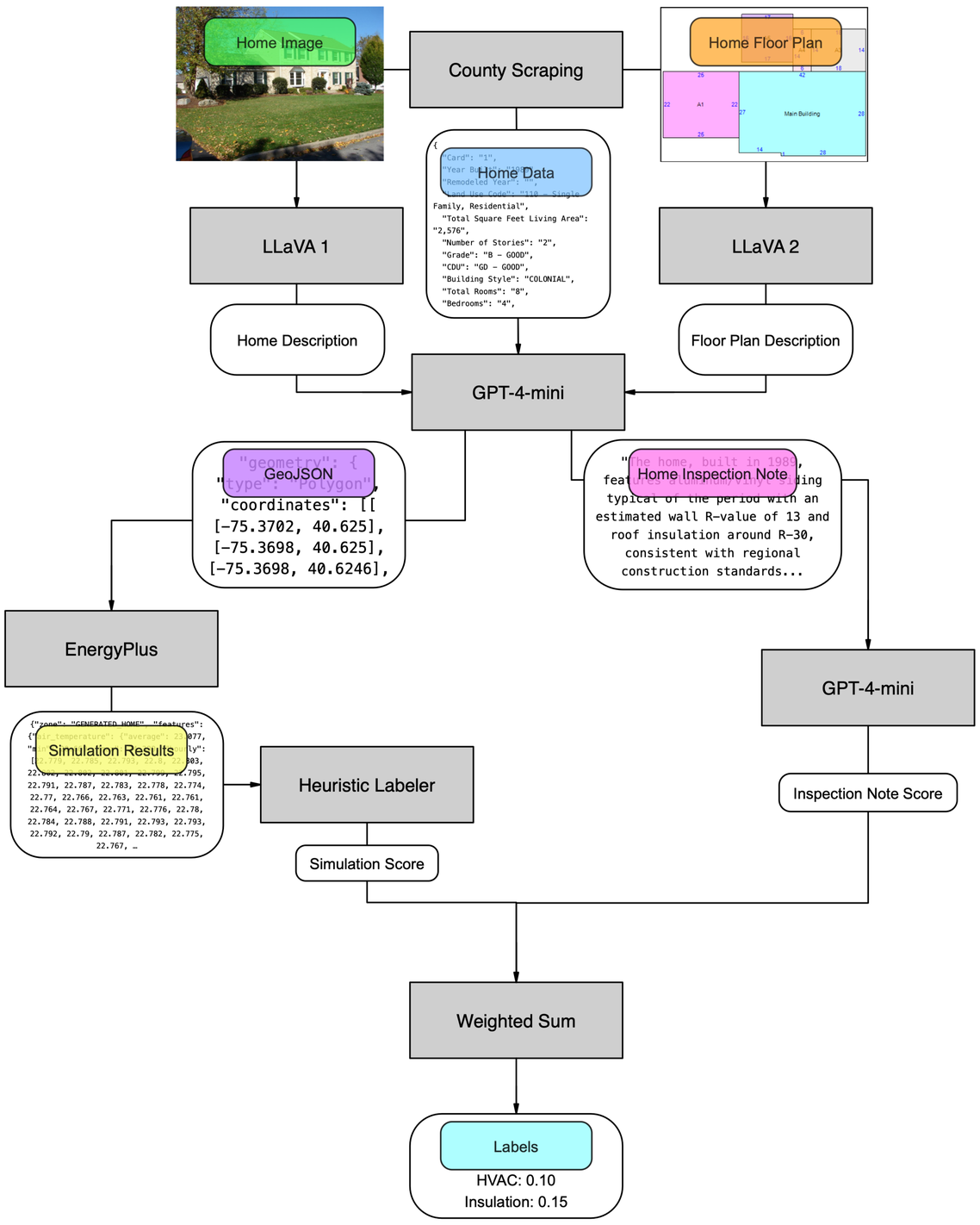
Methodology



Our pipeline on the right comprises five modular components that transform public data and generative AI outputs into simulation-ready datasets. The pipeline produces a dataset of synthetic homes, each containing simulation results, a home inspection note, and labels.

Advantages: Produces scalable, multimodal, and realistic data. Generating synthetic homes is significantly less expensive than sourcing actual data.

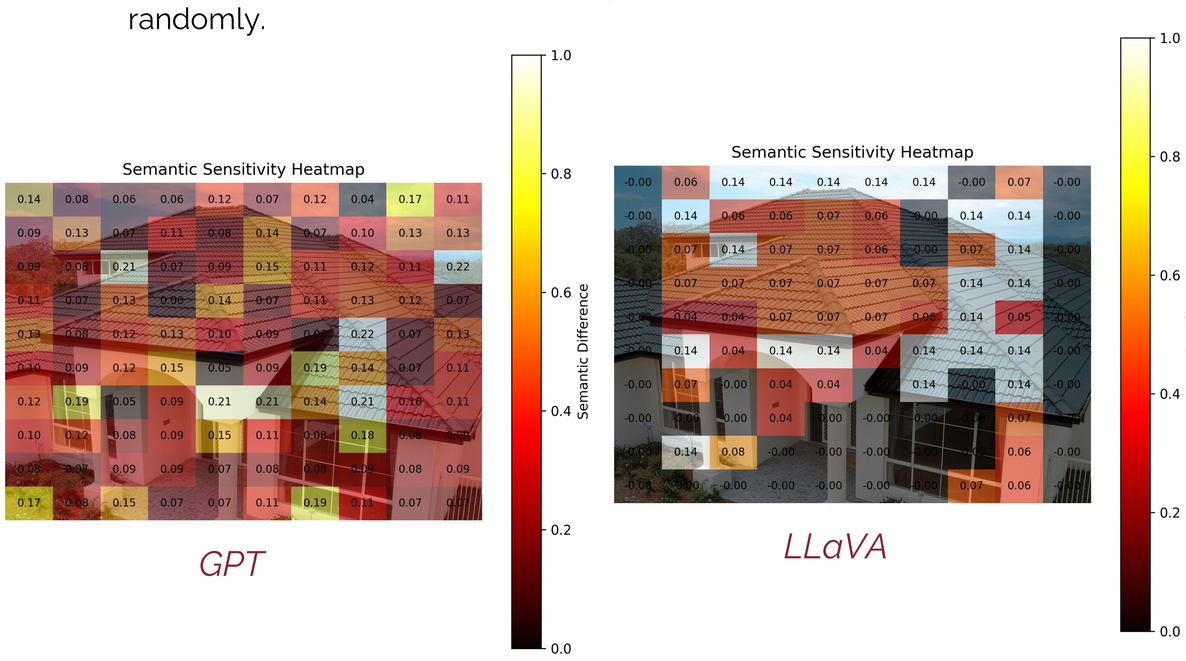
Limitations: Common LLM issues may arise. The data generated only reflects one county.



Results (Pipeline Validation)

Focus Test (Occlusion)

Evaluated two models for their focus on parts of images related to a given prompt. Found that LLaVA focused on relevant parts more on average, while GPT behaved randomly.

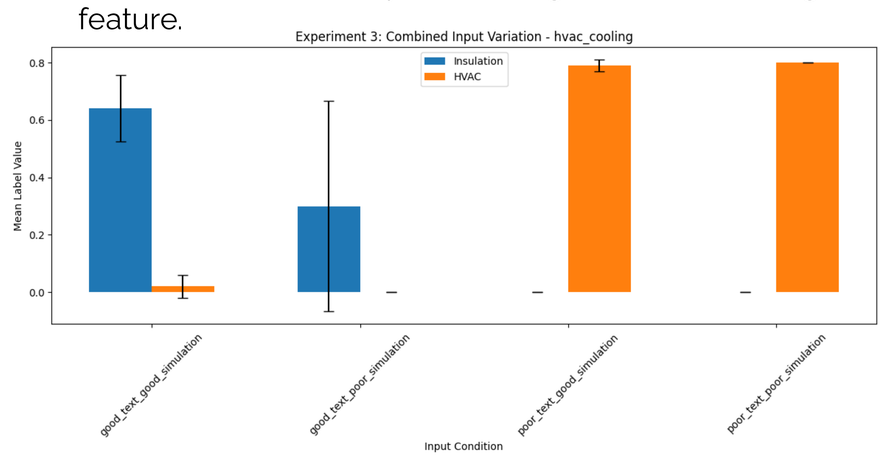


GPT *LLaVA*

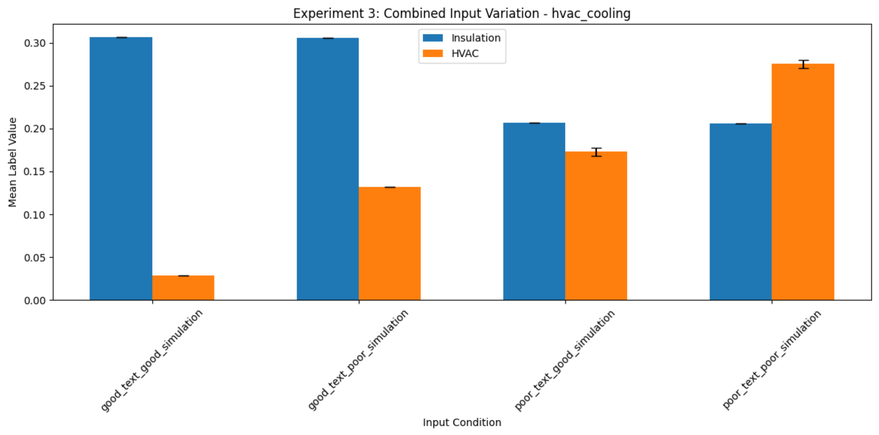
Fairness Test (Ablation)

Iteratively improved the pipeline by ensuring inputs contribute more evenly, preventing bias toward a single feature.

Before



After



Conclusions

- Our pipeline generates inexpensive, scalable synthetic data for urban energy modeling and other applications.
- We validated model focus and fairness across multiple data types, though so far testing only considers one county.
- **Future Work**: train a model to recommend energy-saving retrofits and expand the number of counties used to generate synthetic homes.

Acknowledgements

We thank the Lafayette College EXCEL Scholars program for funding this work. We also thank Lafayette for providing access to computational resources, including JetStream2 and the Firebird HPC.

References

[1] Haowen Xu, Femi Omitaomu, Soheil Sabri, Sisi Zlatanova, Xiao Li, and Yongze Song. "Leveraging Generative AI for Urban Digital Twins: A Scoping Review on the Autonomous Generation of Urban Data, Scenarios, Designs, and 3D City Models for Smart City Advancement". In: Urban Informatics 3.1 (Oct. 2024), p. 29.

[2] Mingzhe Liu, Liang Zhang, Jianli Chen, Wei-An Chen, Zhiyao Yang, L. James Lo, Jin Wen, and Zheng O'Neill. "Large Language Models for Building Energy Applications: Opportunities and Challenges". In: Building Simulation 18.2 (Feb. 2025), pp. 225–234.