On Grounded Planning for Embodied Tasks with Language Models
Research Seminar Number Theoretic and Algebraic Methods in Data Analysis Course

Muhammad Sufyan & Zahid Hussain

December 2025

Problem Motivation

- Grounded planning: generating action sequences based on a goal and an environment.
- Example: "Move the teapot from the stove to the shelf."
- Original G-PlanET paper shows that language models can use structured information (tables) to improve planning.
- Goal of this project: test these ideas on a controlled synthetic dataset.

Dataset

- 300 synthetic examples (240 train, 60 validation).
- Each example includes:
  - Environment table (object, parent, position)
  - Goal sentence
  - Four-step reference plan
- Simple kitchen-like layouts to match the structure of G-PlanET tasks.

Models Evaluated

- BART-base (goal only)
- BART-base with table
- Iterative BART (step-by-step)
- BART-large (goal only)
- BART-large with table
- TAPEX (zero-shot table QA model)
- GPT-4o (evaluated on 10 examples)

Methods: How Plans Were Generated

Single-pass

- Model outputs entire four-step plan in one shot.

Table-based

- Flattened environment table appended to the input.

Iterative decoding

- Plan is generated step-by-step, using previous steps as context.

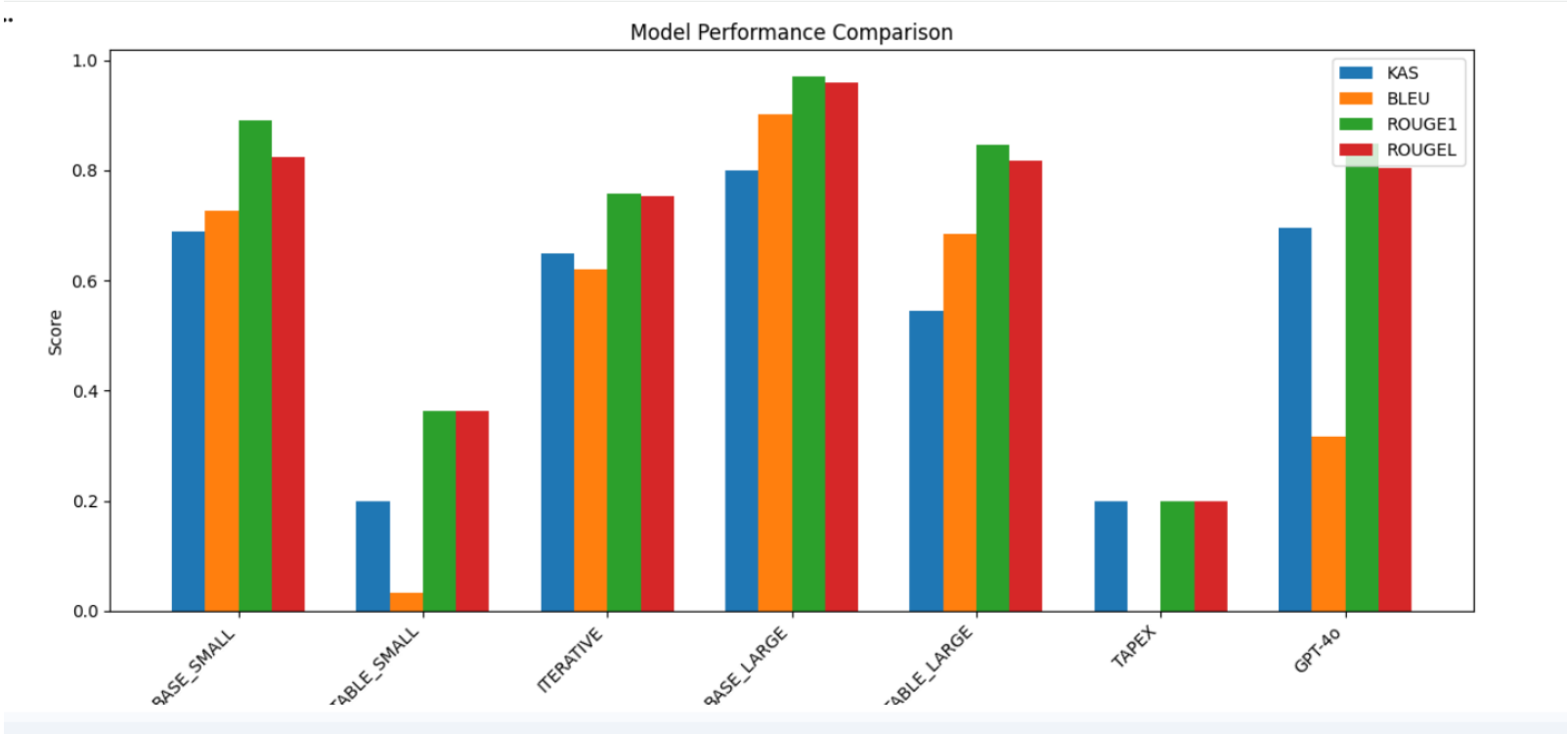
Methods: Evaluation Metrics

- **KAS**: checks overlap of actions.
- **BLEU**: precision-based text similarity.
- **ROUGE-1**: recall of single tokens.
- **ROUGE-L**: measures sequence structure (longest common subsequence).

Results: Score Table

Model	KAS	BLEU	R1	RL
BASE_SMALL	0.690	0.7276	0.8900	0.8240
TABLE_SMALL	0.200	0.0342	0.3636	0.3636
ITERATIVE	0.650	0.6210	0.7580	0.7530
BASE_LARGE	0.800	0.9015	0.9700	0.9600
TABLE_LARGE	0.545	0.6842	0.8470	0.8180
TAPEX	0.200	0.000002	0.1990	0.1990
GPT-4o	0.695	0.3158	0.8483	0.8039

Results: Graph Comparison



Large BART models dominate. TAPEX stays low across all metrics.

Key Findings

- Model size strongly affects planning quality.
- Small models struggle to use table information.
- Iterative decoding stabilizes weak models.
- TAPEX is not suitable for sequential action generation.
- GPT-4o performs well logically but does not follow strict 4-step output.

Limitations and Challenges

Limitations

- Small synthetic dataset
- Fixed four-step plan format
- Simple table structure
- TAPEX and GPT-4o evaluated zero-shot

Challenges

- Difficult for small models to use table information
- Maintaining consistent step order
- Format mismatch between models and evaluation

Conclusion

- Large models provide the most stable results.
- Structured inputs help only when the model has enough capacity.
- Iterative decoding improves weaker models.
- TAPEX does not adapt well to sequential planning.
- Results follow the main pattern of the G-PlanET study.

Thank you!