Lost in Aggregation: A Multi-Scale Diagnostic Benchmark for LLM Spatial Navigation

Yuhan Jiang [1] Peng Luo [2] Liqiu Meng [1]

[1]Technical University of Munich [2]Massachusetts Institute of Technology

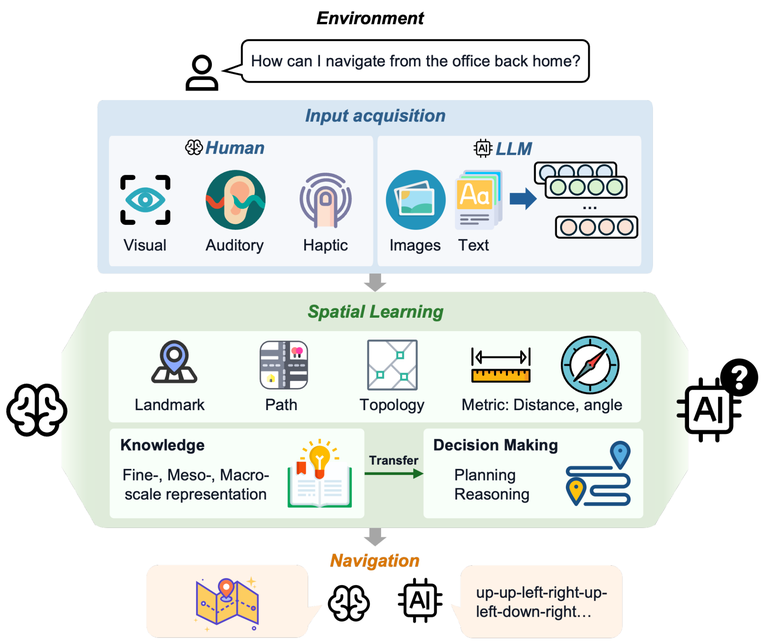
Motivation & objective

LLMs act as **planners and assistants** in tasks with inherent spatial structure (turn-by-turn navigation, route planning, map-based decisions), yet are brittle in *sequential* spatial reasoning. Knowing *that* they fail at navigation is unsurprising; the actionable question is:

Where in the spatial-cognition pipeline do LLMs get lost?

An aggregate success rate cannot distinguish these hypotheses, which imply *opposite* fixes: better perception, better topology, or a different division of labor.

Three cognitive levels



Three levels a navigator must sustain at once:

- **Fine** (local positioning): which adjacent cells are passable.
- **Meso** (topological mapping): which branch at a junction, and when a corridor is a dead-end.
- **Macro** (goal orientation): the global heading to the destination.

Research questions

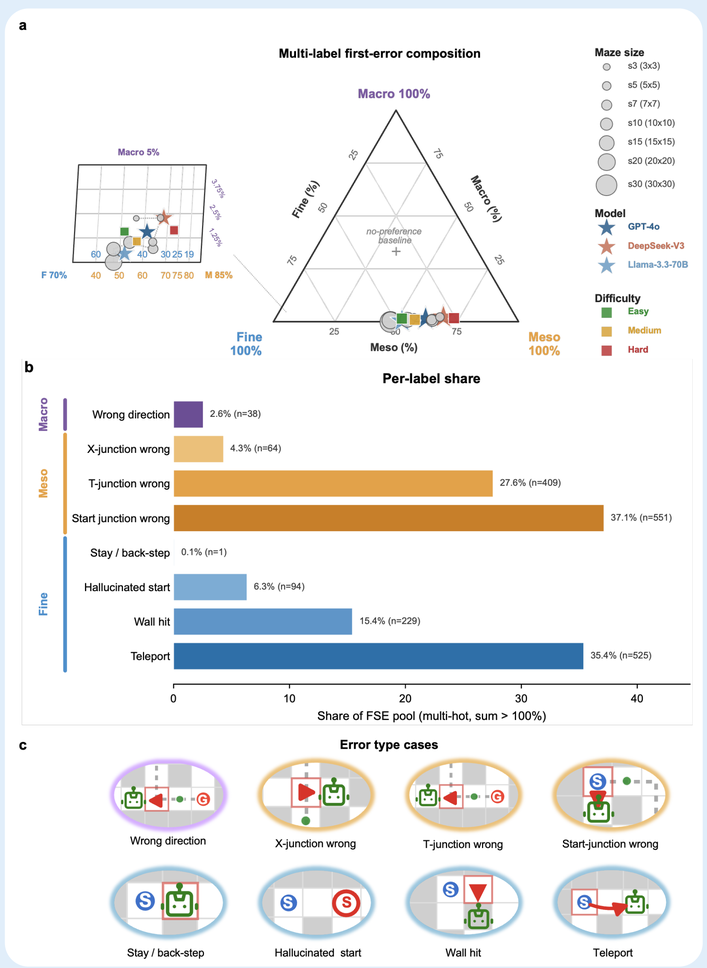
- RQ1: which spatial **input format** best supports navigation?
- RQ2: across sizes and the Fine / Meso / Macro levels, **where** do LLMs fail?
- RQ3: which spatial control should a hybrid agent **delegate** to the LLM vs. the algorithm?

Key finding: lost in aggregation

59% Meso junctions

39% Fine perception

1% Macro heading

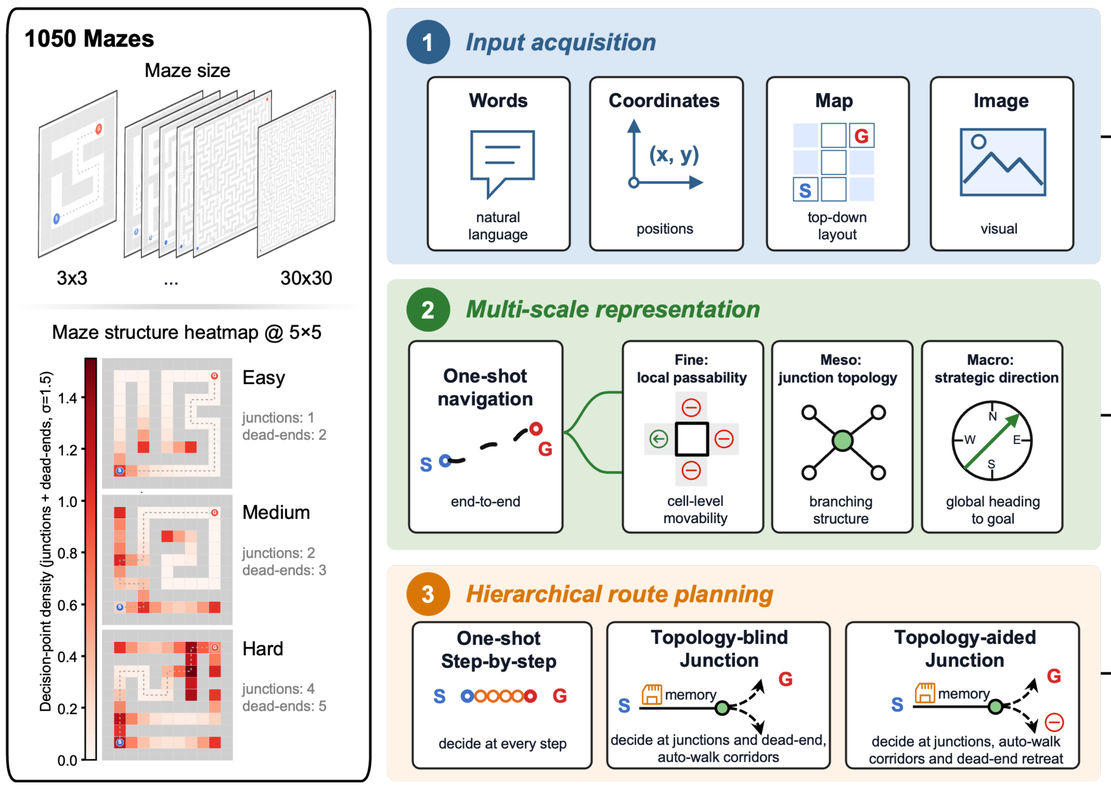


First errors fall on Meso junction choices and Fine perception; global direction almost never breaks first. The same junction is decided *less* reliably inside a long route than in isolation, and that gap grows with size.

Value & impact for urban mobility

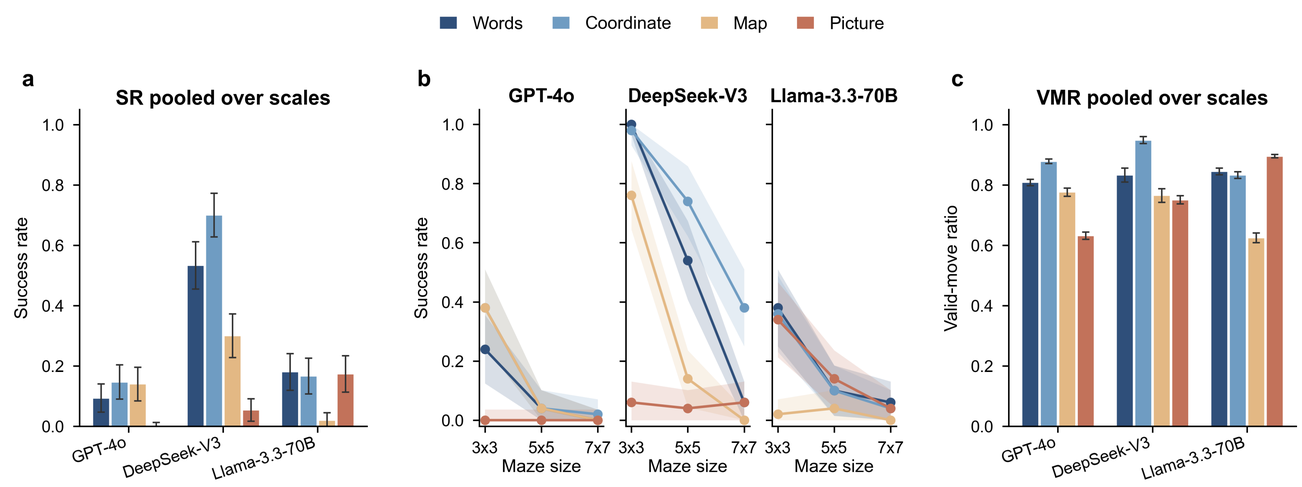
- Use the LLM as a **junction-level decision-maker inside an algorithmic skeleton**, not an end-to-end path generator.
- Surface spatial structure as explicit **coordinates**, not rendered maps.
- **Safety:** mis-navigation stays *confidently legal*, so verify each junction against ground-truth topology.
- Classical search owns metric routing; LLMs suit **topology-dominated** wayfinding.

Benchmark design



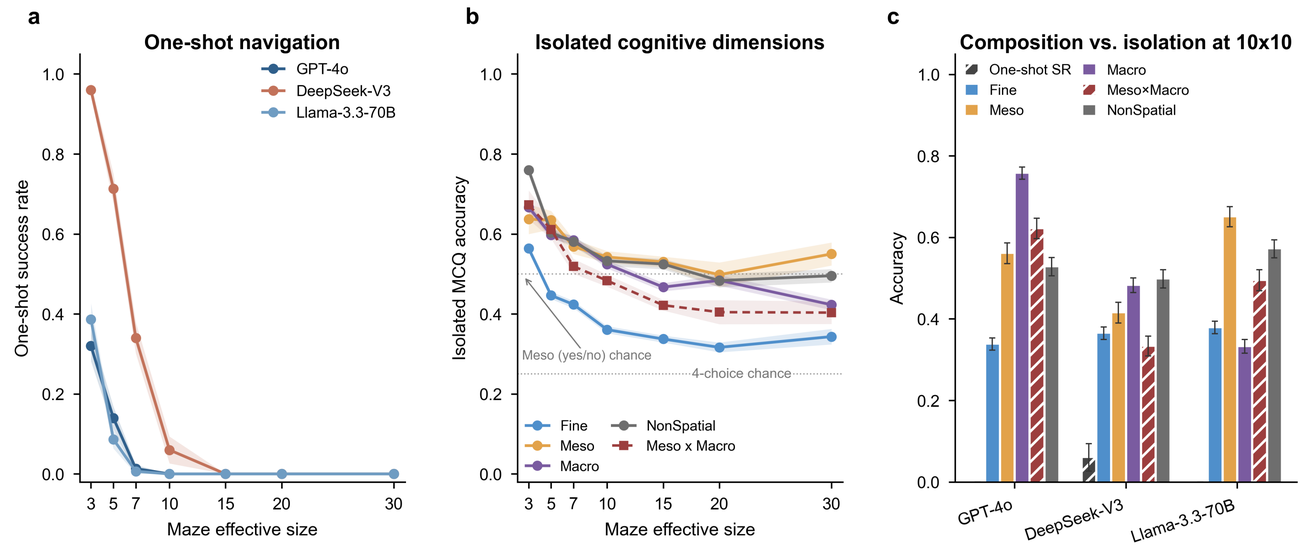
1,050 tree mazes · 7 sizes (3×3–30×30) · 3 tiers · 4 input formats · isolated probes + controls (49,372 Q) · junction-delegation harness · 3 LLMs · released as code + data.

RQ1: Structured text is the most navigable input



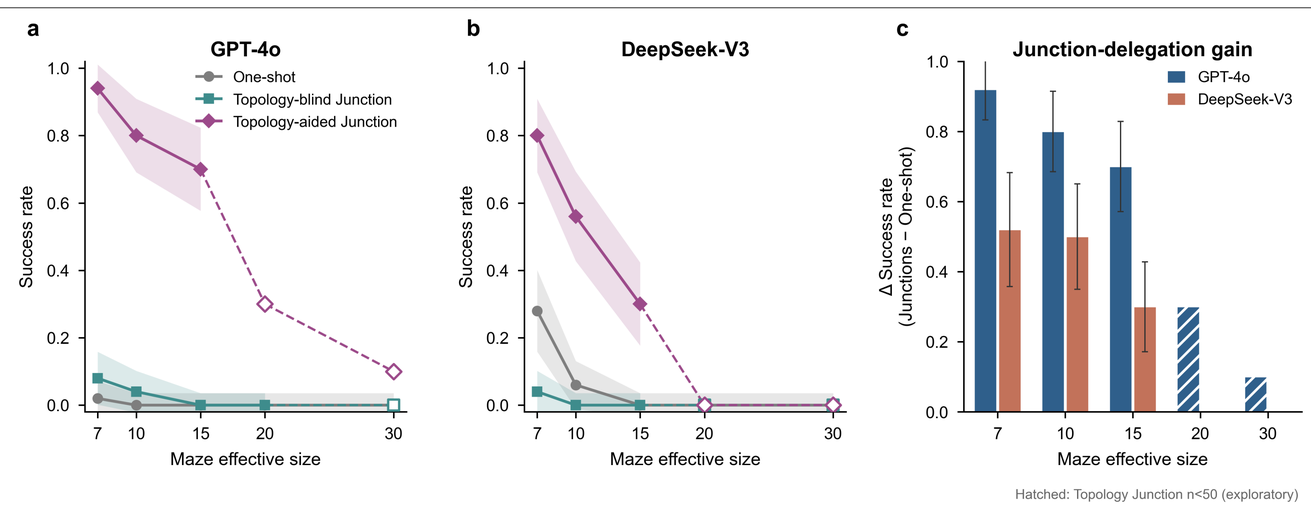
Mean SR: **Coordinate 0.34** > Words 0.27 > Map 0.15 ≫ Image 0.07. Moves stay grid-legal (VMR ≥ 0.62) even at SR ≈ 0, so the collapse is one of **global path** assembly.

RQ2: Navigation collapses; isolated skills survive



End-to-end SR collapses to ≈0 by 10×10, yet isolated Fine / Meso / Macro probes persist. The barrier is **cross-scale aggregation**, not any single competence.

RQ3: Hybrid delegation recovers mid sizes



Hatched: Topology Junction n<50 (exploratory)

Hand corridor-walking and dead-end retreat to a walker, and query the LLM only at junctions with a cell-type prompt. GPT-4o then lifts by **up to +92 points**; the gain is **topological support**, but the wall returns by 30×30.

Take-aways & resources

Size erodes the **integration** of spatial skills, not any single competence. Pair LLM topological judgement with algorithmic execution.

Benchmark, mazes & code:
yuhanjiang415.github.io/lost-in-aggregation

Project & code

LinkedIn

eit Urban Mobility

Co-funded by the European Union

TARTU 2026
Doctoral Training Network | Annual Forum