ERRORS AND PATH-DEPENDENCIES IN LLM DESCRIPTIONS OF THE WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Any prose description of the world contains a stream of small factual commitments: e.g., where different landmarks are and what they look like. We study two recurring properties of these commitments when the prose is generated by an LLM. First, models make discrete factual errors at a non-trivial rate even when the same model, asked directly, can identify the error. Second, the content a model reports is *path-dependent*: it changes with the order in which the environment is traversed, in ways a faithful description should be invariant to. We develop both themes on a benchmark of $6{,}032$ pedestrian corridors in 35 cities, evaluating 2.5 million landmark claims from eleven closed-book models. First-draft claim error runs from 41% to 79%; the places named almost always exist, but they are often misplaced, and side-of-street claims approach a uniform coin flip for several models. Iterative self-critique helps, but a three-parameter model of revision (initial error p, repair rate r, corruption rate q), fitted to the observed transitions, predicts an error floor $e^* = q/(q+r)$ of 28–50% that further revision does not cross. Across cities, difficulty tracks both available text data about the city as well as street-network regularity, and the gap lies in repair: told exactly which claims are wrong, models in irregular, sparsely documented cities still rarely find a correct replacement. Walking a corridor in the opposite direction leaves accuracy unchanged, but changes the textual descriptions significantly: roughly two-fifths or more of the named landmarks appear in one direction only, and left and right fail to invert as they should. A synthetic city with every fact in the prompt provides a stylized setting where the underlying phenomena recur, showing that a substantial part of this divergence is a property of narration rather than of knowledge.

1 INTRODUCTION

Textual descriptions of the world contain, by their nature, long sequences of small factual commitments: a story about a boat tour commits to a number of passengers; a household budget commits to amounts and spending targets, and the arithmetic constraints among them; a description of a walk through a city commits to which landmarks appear, in what order, and on which side of the street.

We study two properties of these commitments in LLM-generated text. The first is that a non-trivial fraction of them are wrong, even for strong models, and even when the same model can identify the error when asked. The second is that the commitments are *path-dependent*: which facts get stated, and what values they take, depend on the order in which the model conveys the information, even when this order is arbitrary and carries no information about the world being described. We provide a framework for cleanly measuring these two effects, and we identify structure in the way they vary across tasks and models. We focus on text about maps and routes, which is both a rich domain for analyzing dense streams of complex factual commitments embedded in everyday text, and also the basis for widely-used applications in geography and navigation that many of us employ every day.

A warm-up analysis (Described in full in Appendix G). Before getting to maps, we can illustrate the key points through some small narrative instances, using Gemini 2.5 Pro as an example. (As we will see, the same effects show up in our main applications in all models we studied.) Asked for a 200-word story in which tourists board a boat and two safety inspectors count them to see if they have enough life jackets, the model makes basic numerical errors in 15% of 600 generations; in a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

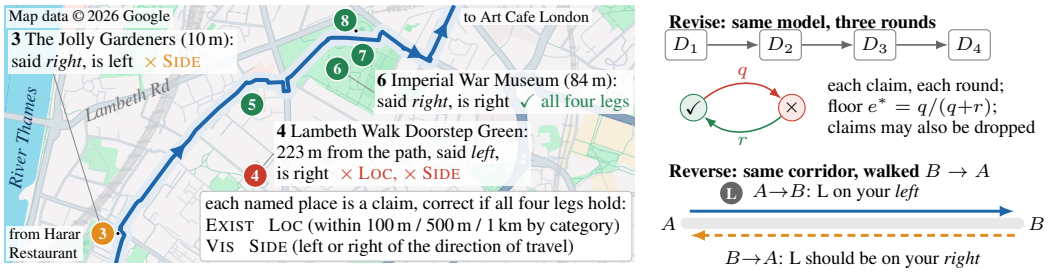

Figure 1: Overview of the city-walk task. *Left:* every named place is a claim, checked against map data (here a segment of the London corridor from Appendix B.2) on four binary properties. *Right:* the same model revises its answer three times (Section 3) and walks the corridor in reverse, where left and right should swap (Section 5).

representative example, the model's story introduces two crew members and six tourists individually by name and then has a safety inspector say, "Looks good ... seven people on board." A second instance of the model, asked to audit the stories, finds the errors almost perfectly; and if we keep the prompt the same but instead of a story ask for a "worked example for a math book," the error rate drops to one error out of 1000 generations. In other words, the model is powerful enough to easily handle the underlying numerical consistency; the issue is that it nevertheless regularly gets such facts wrong in the process of free-form text generation — even though the quantities are small and there is no "difficult math" involved.

Our second main phenomenon, path-dependence, shows up in a similar narrative task where we make order salient. Asked to write an advice column for a family budget on a $5,000 monthly income with five categories listed in a given order, and then asked separately to write an advice column with the categories listed in the reverse order, the model reports mean savings (over several hundred runs) of $1,661 versus $1,465 (a 12% drop; and the drop increases to 30% when the piece is prompted as a story rather than advice column). The drop occurs for orders where the smaller budget categories are listed first: in this case, the amounts recommended for spending on these small categories are inflated, and the result is that the advice column correspondingly recommends lower levels of savings.

Maps and routes. Our main focus in the paper is on the structure of these phenomena in applications with maps and routes; we choose this application area for several key reasons. Descriptions of walks are *densely factual* (a 400-word description contains about eleven separately checkable claims), *externally grounded* (whether the Imperial War Museum is on your right walking northeast on Lambeth Road in London has an answer available at scale from map data), and *structured* (each claim depends on where the walker is and which way they are facing along the route, so maintaining consistency requires tracking orientation across turns). Finally, questions about maps and routes are also something people actually ask for in everyday life.

Core Contributions. First, we provide a benchmark of 6,032 real walking routes in 35 cities, evaluated claim by claim on four dimensions (existence, location, visibility, side of street), for eleven closed-book models from five families over four self-revision drafts under one standardized prompt family (Section 2). Second, we provide a natural three-parameter model of critique-and-revise loops at the level of individual claims, with a tractable fixed-point behavior that tracks the observed error rate after multiple revisions across the eleven models (Section 3).

Next, we perform several analyses in this framework. We provide a city-level analysis of errors, finding that two measured properties of cities (one about the available data, one about the layout of streets) explain 42% of cross-city variance (nearly matching the 45% explained by a regression adding six further city-level factors; Appendix D.1). We find that much of the difference between cities over multiple drafts comes not from the initial error rate but from differential ability to repair errors (Section 4). We then do a path-dependence analysis: traversing a corridor through a city in the opposite order leaves accuracy essentially unchanged but changes what is said. Performing this analysis in a synthetic city with all facts provided in-context isolates the narration component, and inconsistencies between the descriptions in opposite directions serve as a heuristic for detecting errors (Section 5). We conclude with an interpretation in terms of several cognitive phenomena including the contrast between route versus survey knowledge (Sections 6 and 7).

2 THE CITY-WALK BENCHMARK

2.1 TASK

The model is given an origin, destination, and turn-by-turn route in a real city and asked what a pedestrian would see and on which side of the street. Example (middle excised for space): *"If I walk from ZenLeaf Medical Cannabis Store (80 Delancey St, New York, NY 10002) to L'Adresse NoMad (1184 Broadway) along the following route: 1. Head northwest on Delancey St toward Allen St 2. Turn right onto Bowery … what are some landmarks I would see along the way, and what side of the street would they be on?"* Every model sees the same prompt (Appendix A.3), which asks for a prose narration. The route is given, so the model is not planning; it only has to describe what a walker would see. The prompt tells the model that it cannot decline to answer or point to a Maps service; instead, the model must give its best answer. A typical response names about eleven places; we evaluate each named place as a claim with four associated binary questions that we call the four *legs* of the claim: EXIST (is it a real, operating place?), LOC (is it within a category-appropriate distance of the walked path: 100 m for a storefront, 500 m for a park or plaza, 1 km for a skyline landmark?), VIS (is it visible from the street?; a claim fails this leg only when it asserts visibility of something occluded), and SIDE (when the claim specifies left or right, is that side correct for the direction of travel?). A claim is counted as correct if all four legs hold; a claim that does not state a side passes SIDE in this overall score (Appendix B.4 reports the strict alternative that penalizes omitted sides).

After the first draft (D_1) the same model audits each claim in its answer against the four binary legs and rewrites it in full three more times, producing drafts D_2, D_3, D_4 (route mapped in Figure B.4). The benchmark covers 35 cities across multiple continents, street network types, and levels of textual prevalence on the Web; in total we use 6,032 directed corridors from these cities forming 3,016 forward/reverse pairs. On these corridors we evaluate eleven closed-book models from five families (Table 1, Table A.2), yielding 2,523,699 evaluated claim records across D_1–D_4 (742,509 on D_1). How the 35 cities were selected, and how corridors were sampled within them, are given in Appendices A.1 and A.2. The two directions of each pair are routed independently by the map provider and kept only when their forward and reverse paths stay within 40 m of each other, so that both traversals follow the same physical corridor (Appendix A.2).

2.2 EVALUATION

Claims are scored by an automated pipeline. An extractor (Gemini 3.8 Flash Lite) pulls every place claim from a draft, and each is resolved deterministically against factual geospatial data using Google Maps API (346,540 indexed entities; 98.5% of first-draft claims resolve). From each resolved place we compute its shortest distance to the walked route, where along the route it falls, and which side of the street it is on, and the Google Maps Static API renders a close-up map of each place and an overview tile of the route with every claim marked (Figure B.2). A multimodal agent (Gemini 3.8 Flash) reads the grounded table and the map tiles, with access to a search tool when in doubt, and scores the four legs. Because EXIST, LOC, and SIDE are judged against ground truth, evaluator judgment is confined almost entirely to VIS: dropping VIS leaves every model's rank unchanged, re-scoring with a second evaluator checkpoint is virtually identical, and a 300-claim human audit agrees with the evaluator 95% of the time (Appendix B.3). Appendix B.2 gives a worked example of one corridor (Figure 1, left; Table B.1); Appendix B.7 maps further walks.

2.3 CHARACTERIZING ERRORS IN THE MODELS' NARRATIVES

Across the eleven models, first-draft accuracy runs from 21.4% to 59.0% and fourth-draft accuracy from 35.1% to 70.8% (Table 1). Three patterns stand out. First, the models know the places but not precisely where they are. On the first draft (D_1), ten of the eleven models name real places at least 97.9% of the time (EXIST), yet even the best performing model reaches only 78.7% on (LOC) and (VIS), and first-draft SIDE accuracy is only 58.5–74.4% even among places that exist and lie on the route. Second, within a model family, the advantage of the largest model often emerges across revisions: on D_1, Gemini 3.1 Pro starts essentially tied with 3.5 Flash (58.7% vs. 59.0%) and Claude Opus 5 starts below Sonnet 4.5 (47.8% vs. 51.8%), yet by D_4 the larger models lead their families (70.8% vs. 65.0% for Gemini, 64.1% vs. 58.4% for Claude, and 58.9% for GPT-5.6 Terra vs. 35.1% for o4-mini). Third, model families differ in their failure patterns: on D_1, the two leading Gemini

Table 1: Eleven closed-book models on the city-walk benchmark (6,032 routes, 35 cities, all values in percent, rows ordered by D_4 accuracy). Four-leg claim accuracy at D_1 (first draft) and D_4 (fourth draft); per-revision corruption q and repair r; the floor $e^* = q/(q+r)$; and first-draft (D_1) accuracy on each leg, with SIDE on claims with an explicit left/right claim ($\mathcal{C}_{\text{spec}}$), unconditionally and conditional on the place existing and being on route (95% corridor-cluster bootstrap CI half-widths are at most $\pm 1.1\%$ across all columns; Table B.2).

	4-Leg Acc.		Revision			First-Draft (D_1) Leg Accuracy			
Model	**D1**	**D4**	q	r	e^*	EXIST	VIS	LOC	SIDE ($\mathcal{C}_{\text{spec}}$ / cond.)
Gemini 3.1 Pro	58.7	**70.8**	11.9	30.9	**27.8**	98.8	76.6	76.9	67.9 / 73.1
Gemini 3.5 Flash	**59.0**	65.0	12.3	25.4	32.6	99.0	**78.7**	**78.7**	67.7 / 72.0
Claude Opus 5	47.8	64.1	15.1	31.5	32.4	99.1	72.3	62.8	**70.0 / 74.4**
Claude Sonnet 4.6	48.1	60.9	19.5	31.6	38.1	**99.6**	76.4	68.3	55.9 / 61.4
GPT-5.6 Terra	50.1	58.9	**11.6**	18.1	39.0	**99.6**	69.8	71.0	68.1 / 70.5
Claude Sonnet 4.5	51.8	58.4	20.4	**34.6**	37.1	99.5	72.5	69.7	56.6 / 63.0
Gemini 3.0 Flash	50.4	57.7	15.2	23.3	39.5	98.3	69.3	70.3	61.5 / 67.8
Qwen 3.8	49.9	57.5	11.8	17.2	40.8	99.1	70.9	67.5	57.1 / 63.7
GLM-5.3	35.1	51.8	16.9	20.0	45.8	97.9	55.0	51.9	48.9 / 58.5
Gemini 2.5 Flash	49.6	49.8	18.5	19.1	49.2	98.8	67.9	69.4	57.7 / 64.1
GPT o4-mini	21.4	35.1	21.2	21.1	50.1	78.2	30.1	31.9	34.3 / 60.5
11-Model Macro (D_1)	47.4	57.3	15.8	24.8	39.3	97.1	67.2	65.3	58.7 / 66.3
Claim-Weighted (D_1)	46.5	58.5	14.9	24.5	37.7	96.6	66.3	63.7	59.7 / 68.5

models are best on LOC (76.9–78.7% vs. $\leq 71.0\%$ for others), whereas Claude Opus 5 is best on SIDE (70.0%) but near the bottom on LOC (62.8%).

3 REVISION AS A STOCHASTIC PROCESS

3.1 THE (p, q, r) MODEL

Every model improves from D_1 to D_4. To study the dynamics, we analyze a simplified model that treats a revision as acting independently on each claim: a first-draft claim is wrong with probability p; in each revision an incorrect claim is *repaired* with probability r and a correct one *corrupted* with probability q. Writing e_t for the expected error rate in draft t, and e^* for the limiting error rate under repeated revision, we have

$$e_1 = p, \qquad e_{t+1} = q + (1 - q - r)\, e_t, \qquad e^* = \frac{q}{q+r}. \tag{1}$$

The model has three natural consequences that are all reflected in the data. First, repeated revision has a floor e^*, and the floor does not depend on p: a model cannot be revised below an error rate of $e^* = q/(q+r)$ by this procedure, however many rounds are spent. Second, because the initial error rate p does not appear in the expression for e^*, the size of the gain after one revision does not tell us where repeated revision will end up (Figure 2a). For example, o4-mini gains 13.7% with $r \approx q$, relaxing from $p = 78.6\%$ toward a fairly high $e^* = 50.1\%$, while Gemini 3.5 Flash gains only 5.9% because it starts close to its floor $e^* = 32.6\%$. Gemini 3.1 Pro and Claude Opus 5, with lower floors, overtake models that started ahead of them. Third, convergence is geometric with rate $|1 - q - r|$, so most of the gain arrives early: nine of the eleven realize 61–81% of their total $D_1 \to D_4$ improvement by D_2, while Claude Sonnet 4.5 overshoots at D_2 and gives part of it back.

Note that e^* can be written entirely as a function of r/q, since we have $e^* = 1/(1 + r/q)$, and across the eleven models, r/q rises from 1.00 (GPT o4-mini) to 2.59 (Gemini 3.1 Pro), lowering the floor from 50.1% to 27.8%. Stronger models can be better either by corrupting less or repairing more (Appendix D.3).

Across the eleven models the Pearson correlation between e^* and observed D_4 error is 0.93 (this is a consistency check rather than a forecast, since q and r are estimated from the same drafts; Appendix C.5 gives out-of-sample versions). For Gemini 3.1 Pro, Equation (1) predicts a D_4 error of 30.36% against an observed 29.24%. Estimating (q, r) out of sample — either from $D_1 \to D_2$ alone or from 17 training cities to forecast 18 held-out cities tracks the fourth-draft ranking (Spearman 0.93; Table C.3). The plain recursion misses o4-mini, whose D_4 error of 64.9% sits 9% above the 55.6% it predicts, because o4-mini's revision is dominated by deletion and addition rather than

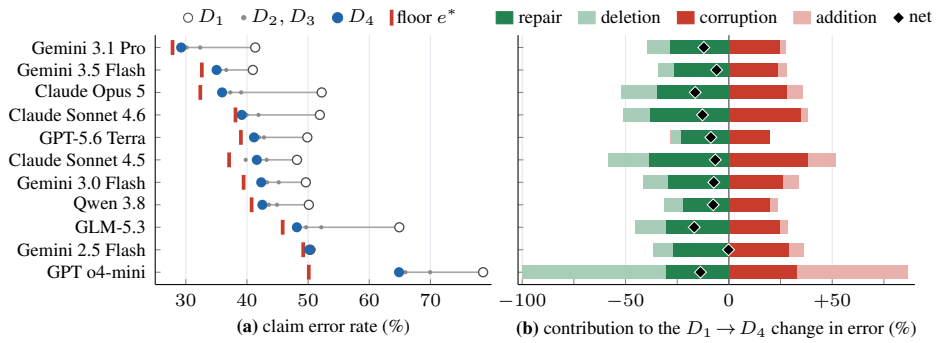

Figure 2: Revision approaches a model-specific floor, through churn and deletion. (a) Claim error by draft against the floor $e^* = q/(q + r)$ implied by each model's measured rates. (b) Exact decomposition of the $D_1 \to D_4$ change (Appendix C.1): repair and corruption largely cancel, and deletion supplies much of the net gain.

repair (Section 3.2). Incorporating empirical survival and addition rates via a deletion-aware Möbius extension improves the fit to observed D_4 from MAE 2.03% under the plain recursion to 0.95% across the eleven models (max 2.26%; Appendix C.3, Table C.2).

3.2 The role of attrition in repair

Length by itself does not explain accuracy differences across cities ($R^2 = 0.005$) or models ($R^2 = 0.11$), but it does change in a clear direction across revisions. As models revise their narrations of walks, they also shorten them by dropping claims: averaged across models and cities, first-draft narrations contain 11.2 claims while fourth-draft narrations contain 8.7. There is significant variation across cities and models: for example, Manhattan averages 14.3 claims on D_1 and 11.8 on D_4, while Kigali falls from 7.9 to 5.5; Claude Opus 5 drops from 25.0 to 17.6 and o4-mini from 13.8 to 7.3. The rising accuracy and shrinking number of claims across drafts work in opposite directions, and in absolute numbers, seven of eleven models end with fewer correct claims per description (Figure C.1). Because every claim is matched across drafts, the error-rate change decomposes into repair, corruption, deletion, and addition (Figure 2b; Table C.1). For example, Gemini 3.1 Pro repairs 28.7% of error and creates 24.6%, so most of its 12.1% gain comes from selective deletion (-10.7%). For o4-mini, corruption (32.9%) outweighs repair (30.6%), so its entire gain comes from discarding nearly half of what it said.

The prevalence of different types of errors changes across revisions as well. Pooled over models and cities, D_1 is wrong on location (36.3%), side on $\mathcal{C}_{\text{spec}}$ (40.3%), and visibility (33.7%); by D_4, location errors fall by a third, visibility by two fifths, and side by only a seventh. Off-route claims are repaired outright 14% of the time but half are deleted by the next draft; retained wrong-side claims flip to the true opposite side 34.2% of the time (Appendix B.4). Side-of-street improves the least (Appendix C.2; Figures B.4 and C.2 trace one walk across D_1–D_4).

4 Why Cities Differ

Why should Bangalore be harder to describe than Manhattan? We test two explanations, each by a proxy measure independent of any model. The first is a structural explanation, based on the hypothesis (confirmed in our subsequent analysis) that a city with regular, "grid-like" streets should be easier than a city with an irregular, organic road network; we operationalize *street regularity* as the mean intersection degree of the street network [5, 6], which will be near 4 in a grid. The second explanation is the availability of textual data, via a measure of *attestation*. In particular, a model's knowledge of a place comes largely from text about it; we operationalize this as the geometric mean number of web documents referring to the landmarks along a city's corridors (Figure B.3).

Figure 3a plots the 35 cities, pooled across the eleven models and shaded by attestation. Regular street networks are easier: mean intersection degree has Spearman $\rho = -0.63$ with fourth-draft error ($p < 10^{-4}$). Attestation correlates more weakly but still non-trivially (-0.42). Together the two axes explain 42% of the variance across cities, with degree alone explaining 40%; no other pair of measured descriptors does better (Appendix D.1). Because the two axes are correlated ($\rho = 0.50$),

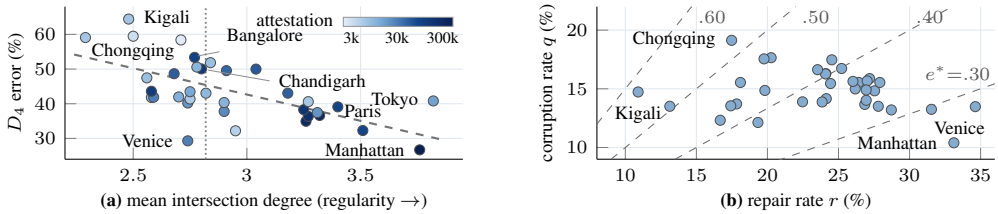

Figure 3: City difficulty is organized by street regularity, and cities differ mainly in repairability. (a) Fourth-draft error against mean intersection degree for the 35 cities, pooled over the eleven models, shaded by attestation; dashed: OLS fit ($R^2 = 0.40$); dotted: median degree. (b) Per-city corruption q against repair r; dashed rays are contours of constant floor e^*. Vertical spread is modest, horizontal spread more than threefold.

35 cities can show that regularity carries most of the city-level signal with a positive contribution for attestation. Per landmark, attestation matters more: within each of two frontier models, claim accuracy rises by 11–12% from the least- to the most-attested decile. Some cities depart from the fit: Venice and St Petersburg are much easier than the two axes predict, and Tokyo, with the highest degree in the set, has an error rate 10% above the trend line. Sana'a, Antananarivo, Chongqing and Kigali, unfavorable on both axes, have the most error-prone descriptions at 58–64% (Table D.7).

Estimating q and r per city (Figure 3b) shows where the gap arises. Corruption is comparatively stable (10–19%), while repair varies more than threefold, from 35% in Venice to 11% in Kigali. Self-critique flags about four in five wrong claims in every kind of city (and two in three right ones), but flagging a claim is distinct from rewriting (Appendix D.3). The difference across cities lies to a significant extent in the ability to successfully rewrite a claim: in organic, sparsely attested cities a flagged error is more often deleted and, when kept, less often corrected.

To test this, we reran one revision for three models on 160 corridors in 20 cities, each starting from the same first draft, and told the model exactly which of the claims were wrong (Appendix D.4). This type of perfect guidance doubled the share of wrong claims deleted, but it did not eliminate the gap in performance between the cities: repairs in the harder cities were still error-prone, and the gap in r between grid-like, well-attested cities and organic, sparsely attested cities widened from 20 to 43%.

5 Path-Dependence: The Same Street, Walked in Reverse

A description of walking from A to B should essentially be a description of walking from B to A in reverse order, with the roles of left and right exchanged. While overall accuracy is essentially unchanged between the two directions (53.78% vs. 54.25%; Appendix E.1), we find that the model's description of a walk from A to B picks out a significantly different set of landmarks than its description of the walk from B to A. Some differences accord with physical experience: a landmark that looms just beyond B as you approach it from A would be behind you if you start from B and walk toward A, and therefore less salient. But other differences, when a landmark clearly visible on the side of the street is described on the way from A to B but not on the way from B to A, appear to lack any physical basis. In this section we analyze these differences, and distinguish among them on physical or visibility bases (Figure B.5 maps one corridor walked both ways).

5.1 Differences in content between A-to-B and B-to-A walks

We begin by considering the Jaccard overlap between the landmark sets named in the two directions, a natural way to study the dependence on direction. This Jaccard overlap ranges from 0.14 (for GLM-5.3; 0.15 on GPT o4-mini) to 0.43 (for Claude Sonnet 4.5; see Figure 4a and Table E.1). Even for the most consistent models, about two-fifths of the places in a description are mentioned in only one direction. The *order* of the shared landmarks reverses reliably: Kendall's τ between the two sequences is -0.77 to -0.87 for nine models (whereas Claude Sonnet 4.5 and Claude Sonnet 4.6 sit near -0.51). But left and right do not reverse reliably: for a landmark named in both directions with an explicit side commitment, the stated side flips only at most 70.9% of the time (Claude Opus 5), and for six models it is within 4% of 50% (Figure 4a), near what we'd get from an unbiased coin flip. There is a structural contrast corresponding to these results: Reversing a list is a text-level operation the models do natively, while reversing left-right handedness requires representing the walker's orientation.

Across cities, Jaccard overlap ranges from 0.26 in Chongqing to 0.40 in Paris and is nearly orthogonal to the two macroscopic city axes that we considered in the previous section ($\rho = 0.07$ with degree, 0.00 with attestation).

5.2 Landmarks named consistently in one direction and never in the other

Comparing a single A-to-B walk with a single B-to-A walk can reveal differences in what gets mentioned, but it can't tell us whether there are certain landmarks that get consistently mentioned in one direction and never in the other. To further explore this latter point, we choose a set of corridors and generate eight narrations in each direction (248 paired corridors, $3{,}968$ narrations by Gemini 3.5 Flash, restricted to the 129 corridors whose forward and reverse polylines agree to within 25 m). This gives us 566 route–landmark pairs named in at least seven of eight trials in at least one direction; of these, 488 (86.2%) are named consistently in both directions and 78 (13.8%) are named in seven or eight of eight trials one way and zero or one of eight the other. This is not hallucination: existence accuracy in the one-way set is 100%. The model is simply consistently mentioning a particular real place in one direction, and declining to ever mention it when it walks the other way. Landmarks named only in one direction sit closer to the end of whichever traversal names them (mean normalized position 0.63; 41.0% in the final quarter versus 11.5% in the departure quarter), while symmetric ones sit mid-route. Some of this reflects landmarks that are visible beyond one of the endpoints (the Intrepid, an aircraft carrier at the end of 46th Street in Manhattan, is named in $8/8$ westbound walks and $0/8$ eastbound, because walking east it is behind you). But in many cases, there is no natural basis in visibility (e.g. a convenience store used as a turn marker in Shinagawa is named $8/8$ one way and $0/8$ the other, although it sits at the same corner either way). The one-way rate is nearly flat across within-city landmark attestation tertiles (16.4% in the bottom third versus 13.3% in the top third; Appendix E.4).

5.3 The shape of the narration

Projecting each of 2.52 million claimed landmarks onto its route gives a normalized position $u \in [0, 1]$ oriented to the direction of travel (claims that project onto an endpoint are excluded). Mentions are strongly U-shaped (Figure 4b): the first and last tenths of a route carry 14.7% and 19.0% of all mentions against 7.9–9.0% for each interior tenth, sharper still within 50 m of the path. Forward and reverse traversals produce nearly identical curves (final decile 19.5% versus 18.5%), so the profile is a property of traversal, not of the street, and reversing the walk applies it to the other end of the same corridor. The concentration at the ends is lower in grid-like and well-attested cities ($\rho = -0.53$ with degree, -0.49 with log attestation).

5.4 A synthetic city with a self-contained set of facts

To separate effects of pre-trained city knowledge from effects of how a walk is narrated, we generated a fictitious street with 99 numbered cross-streets. A table of fictitious landmarks with cross-streets and sides is supplied in the prompt, and the model narrates a walk in one direction or the other; 300 such streets, three models, about $12{,}000$ narrations (Appendix E.6). We observe four phenomena that bear on the results from real cities. First, the narration does not invent places: across $99{,}251$ landmark mentions, none named a place that wasn't in the supplied table. Second, even in the synthetic route, reversing the direction still changes what is mentioned: two narrations of the same street overlap at Jaccard 0.56–0.68 in the same direction and 0.43–0.61 in opposite directions, indicating that two narrations of the same walk in the same direction have more overlap in the landmarks they mention than two narrations of the same walk in opposite directions. Third, these effects hold whether the table of (fictitious) landmarks supplied in the prompt are in sorted order or randomly shuffled, and in fact the divergence between the two directions gets slightly larger when the table is shuffled (Figure 4c). That is, the phenomena we're seeing in synthetic cities arise from position along the walk, not in the prompt. Finally, supplying facts for synthetic cities can remove side-of-street errors: when the prompt specifies whether each landmark lies on the east or west side of the street, a Gemini model converts east and west into left and right with zero errors in $7{,}709$ opportunities (unlike in real cities, where the layout must be recalled from memory). This does not hold for other models; for example, Claude Sonnet 4.6 still struggles when walking south (30.1% side error vs. 19.5% walking north). Forcing every landmark to be named with its cross-street cuts wrong-street attributions by 88% (1.72% to 0.21%), and shrinking the word budget from 400 to 100 raises them from 0.72% to 4.47%.

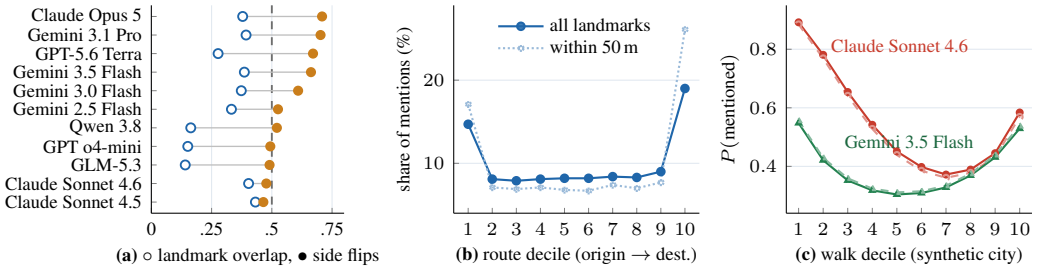

Figure 4: Reversal changes what is said, and the narration favors landmarks at beginnings and ends. (a) Per model, overlap of the landmark sets named in the two directions (○) and share of shared, side-specified landmarks whose side flips (●; dashed: independent of direction). (b) Mentions by route decile, real routes. (c) Synthetic city, shuffled table: probability of mention by walk position (solid S→N, dashed N→S).

5.5 Reversal disagreement as an error signal

If a model names the same landmark in both directions of a corridor and states a side of the street both times (a *shared pair*), there is a natural consistency check that can be applied: the stated side should flip, and if it does not, at least one of the two claims must be wrong unless the landmark straddles the route. On first drafts, among shared pairs whose side label flips as required, both statements are scored correct 43.6–78.2% of the time across models; when the model repeats the same side label in both directions, at least one claim is indeed wrong 93.3–97.4% of the time (both are correct on only 2.6–6.7% of such pairs, mainly large parks or plazas). This has the effect of catching 62.9–70.3% of all side errors among shared landmarks purely through a consistency check, without requiring a map (Table E.3; Appendix E.7, Figure B.5, Table B.3). Asking a model to describe the same route in the opposite direction thus exposes spatial errors that single-direction can self-critique miss.

6 Related Work

In Appendix I, we discuss the connections to related work in full detail; here we give a brief overview.

An active body of work on factuality and hallucination analyzes false statements by language models [20, 19, 29], tracing hallucination rates to rare training facts [22], the effects of accuracy-only grading [23], and the effects of pretraining frequency [25] and entity popularity [35] on recall. The errors we focus on, in contrast, are more relational and geometric, since our analysis shows that the landmark errors are rarely fabrications. Our model with parameters (p, q, r) connects to a long line of work on iterative critique [34, 56, 10, 2, 16, 72, 54, 18, 24, 59]; measuring q and r separately can potentially help separate some of the forces that limit self-correction ability [77, 76, 50, 70], and make clear the domain dependence at work.

Our work contributes to the ongoing exploration of whether next-token generators contain implicit world models [28, 45, 17], and how despite geographic knowledge [4, 37, 51] these models can be brittle [67, 68, 43, 74, 41, 55, 9, 32]. Our findings examine this brittleness at scale, and suggest that a central issue is the gap between possessing geographic knowledge (enough to name real places or spot errors when auditing) and maintaining spatial consistency while generating a long narrative. Our analysis of path-dependence further connects to prompt-order and anchoring biases [31, 79, 48, 80, 30, 21] as well as the reversal curse [3], autoregressive error accumulation [1, 78, 12], and route-versus-survey distortions in people [57, 62, 60, 53, 39, 65, 33] and LLMs [14, 75]. We discuss all these connections in greater detail in Appendix I.

7 Discussion and Conclusion

A consistent property across all of our experiments is that when a model makes an error, a second independent instance of this same model can generally catch it. In our narrative warm-up activity, a second instance of the model audits the provided stories near-perfectly; and in our main application with maps and routes, a second agent repairs a substantial fraction of what it is shown (consistent with how the repair rate r is never near zero). There are natural structural reasons for this: checking a claim is a task with one focus and a clear criterion, while generating a passage means maintaining many implicit factual commitments while satisfying fluency, length, and genre, and committing to

each token before its consequences are fully known. We have found that this gap can be controlled by framing (for example, via genre in the narrative warm-up activity), and by structure (in our maps application, naming the cross-street with every landmark cuts errors significantly). On the other hand, pure exhortation is much less effective; for example, models retain most incorrect side-of-street commitments even when the revision prompt asks them to remove unverifiable claims.

Our analysis highlights recurring strengths and weaknesses of language models for maps and routes: they are effective at naming places along a corridor, putting them in sequence, and characterizing neighborhoods. But they struggle with relative position (which side a place is on, and how it inverts when going in the other direction) and with visibility (whether a landmark is visible from a given point). This contrast in the model's strengths and weaknesses maps in interesting ways onto the classical distinction between *route* knowledge, which is sequential and egocentric, versus *survey* knowledge, which is configural and allocentric [57, 62]. In particular, human spatial memory shows analogous systematic distortions [60, 53, 39, 65]. The analogy between humans and language models in this regard has its limits, of course; for example, with coordinates in context for a synthetic city, some models no longer commit side errors, so in a sense they become able to carry out the allocentric-to-egocentric transform successfully. One account consistent with these observations is that when facts come from parametric memory the model retrieves them sequentially, with a direction built in. Probes showing that models encode latitude and longitude linearly [17] establish that the information exists, not that generation uses it. A small state-tracking pilot points the same way: from a correct heading the next stated heading is correct 70% of the time, and from an incorrect one it is consistent with the error 63% of the time (Appendix H.2); the model tracks a state and reasons faithfully from the wrong one, consistent with the way that $q > 0$ for every model and with how an error, once made, can acquire supporting detail.

Our analysis also highlights how position in the generated sequence plays a key role in the model's performance, beyond the structure of the environment being described. Our results on synthetic cities show that this holds even when prompt order is randomized, which separates it from known input-position effects [31, 79, 30]. Path-dependence can be intuitively appropriate in a description when it favors the landmarks that are in front of you rather than behind you, but it becomes a source of error when the side of the street a landmark is on does not invert when you go in the opposite direction (Appendix E.6).

Several lessons emerge from the dynamics of our (p, q, r) model. First, the corruption rate q can play a decisive role, in particular by offsetting high values of the repair rate r. Second, much of the gain from repeated improvement happens quickly, with the second draft achieving much of the available gain. Third, it is also possible to achieve most of the available gains by starting with cheaper but less accurate models, and then switching to heavier-weight models for subsequent iterations. We explore this idea further in Appendix F via a formal scheduling methodology for optimally sequencing such switches among models.

A number of limitations of our work stem from the difficulty of fully isolating the underlying mechanisms behind some of the effects we observe. In particular, it remains challenging to fully analyze the effect of supplying additional facts to the models, since this simultaneously changes the underlying knowledge source, the representation, and the difficulty. Claim accuracy is conditioned on which landmarks a model chose to name, and critique flags are matched to claims automatically (Appendix D.3); the models are prompted to provide a concrete narration, and we know less about performance in settings where they could more readily decline to be specific. Similarly, there are multiple candidate mechanisms for the effect of reversing the direction of a walk, including U-shaped training text about journeys, beginnings and endings inducing more summary text, and a learned template for narratives that tends to be sparser in the middle.

In closing, our work illustrates how models exhibit a non-trivial rate of errors in their factual commitments as they produce text about maps and routes. Methods for identifying and correcting these errors reinforces the way in which the effort a model spends on repair applies to correct and incorrect claims alike; this leads to the $q/(q + r)$ equilibrium that provides a floor on performance, correlated with the regularity of the city's street network and the amount of text available about it. The disagreement between how a route is described in opposite directions becomes both a source of path-dependence and also a heuristic form of error-detection: by asking for descriptions under factually inconsequential transformations, we get information about inconsistencies even without the presence of ground truth.

Reproducibility Statement

For reproducibility, all model providers, thinking effort configurations, and per-model draft counts are listed in Appendix A.4 (Table A.2). The verbatim walker, critic, and extraction prompts, the 6,032 route specifications, the synthetic-city generator, and all analysis scripts will be released with the camera-ready version.

AI Disclosure and Use Statement

In this work we used generative AI tools for assistance in implementation of code for execution of the research, for assistance with discovery of related work, and for assistance in drafting and revising text and figures in sections of the paper. Some of the analysis asks questions about synthetic geographic data, and this synthetic data was, by the design of the study, produced by an AI model. We used versions of Gemini and Claude Opus for the above tasks. We did not use AI tools for developing the mathematical formalisms in the paper. We have reviewed the AI-generated results, the text produced by AI tools, and the references to related work that they found. We take responsibility for the content of the work.

References

[1] G. Bachmann and V. Nagarajan. The pitfalls of next-token prediction. In *Proc. 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/bachmann24a.html.

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073, 2022. URL https://arxiv.org/abs/2212.08073.

[3] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

[4] P. Bhandari, A. Anastasopoulos, and D. Pfoser. Are large language models geospatially knowledgeable? In *Proc. 31st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2023. URL https://doi.org/10.1145/3589132.3625625.

[5] G. Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017. URL https://doi.org/10.1016/j.compenvurbsys.2017.05.004.

[6] G. Boeing. Urban spatial order: Street network orientation, configuration, and entropy. *Applied Network Science*, 4(1):67, 2019. URL https://doi.org/10.1007/s41109-019-0189-1.

[7] K.-Y. Chen et al. The self-correction illusion: Role relabeling gates explicit error flagging in large language models. arXiv:2606.05976, 2026. URL https://arxiv.org/abs/2606.05976.

[8] W. Chen, M.-W. Su, N. Mehjabin, and M. L. Cummings. Can LLMs plan paths in the real world? arXiv:2411.17912, 2024. URL https://arxiv.org/abs/2411.17912.

[9] A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions. In *Proc. 16th International Conference on Spatial Information Theory (COSIT)*, volume 315 of *LIPIcs*, pp. 28:1–28:9, 2024. URL https://doi.org/10.4230/LIPIcs.COSIT.2024.28.

[10] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 3563–3578, 2024. URL https://aclanthology.org/2024.findings-acl.212/.

[11] P. Ding, J. Fang, P. Li, K. Wang, X. Zhou, M. Yu, et al. MANGO: A benchmark for evaluating mapping and navigation abilities of large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=6vEfyp0o68.

[12] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jiang, B. Y. Lin, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/deb3c28192f979302c157cb653c15e90-Abstract-Conference.html.

[13] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. URL https://doi.org/10.1038/s41586-024-07421-0.

[14] N. Fulman, A. Memduhoğlu, and A. Zipf. Distortions in judged spatial relations in large language models. *The Professional Geographer*, 76(6), 2024. URL https://doi.org/10.1080/00330124.2024.2372792.

[15] O. Golovneva, Z. Allen-Zhu, J. Weston, and S. Sukhbaatar. Reverse training to nurse the reversal curse. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=HDkNbfLQgu.

[16] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

[17] W. Gurnee and M. Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

[18] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

[19] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. URL https://doi.org/10.1145/3703155.

[20] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023. URL https://doi.org/10.1145/3571730.

[21] E. Jones and J. Steinhardt. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4d13b2d99519c5415661dad44ab7edcd-Abstract-Conference.html.

[22] A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. In *Proc. 56th Annual ACM Symposium on Theory of Computing (STOC)*, 2024. URL https://doi.org/10.1145/3618260.3649777.

[23] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653:1047–1051, 2026. URL https://doi.org/10.1038/s41586-026-10549-w. Preprint: "Why language models hallucinate", arXiv:2509.04664, 2025.

[24] R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. URL https://aclanthology.org/2024.tacl-1.78/.

[25] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel. Large language models struggle to learn long-tail knowledge. In *Proc. 40th International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

[26] A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=CjwERcAU7w.

[27] D. Li, R. Sun, Y. Huang, M. Zhong, B. Jiang, J. Han, et al. Preference leakage: A contamination problem in LLM-as-a-judge. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=grIvSXVJ65.

[28] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

[29] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

[30] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. URL https://aclanthology.org/2024.tacl-1.9/.

[31] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. URL https://aclanthology.org/2022.acl-long.556/.

[32] H. Luo, Q. Cheng, D. Matos, H. K. Gadi, Y. Zhang, L. Liu, et al. TurnBack: A geospatial route cognition benchmark for large language models through reverse route. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.440/.

[33] K. Lynch. *The Image of the City*. MIT Press, Cambridge, MA, 1960. URL https://mitpress.mit.edu/9780262620017/the-image-of-the-city/.

[34] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

[35] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. URL https://aclanthology.org/2023.acl-long.546/.

[36] P. Manakul, A. Liusie, and M. J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

[37] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. GeoLLM: Extracting geospatial knowledge from large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=TqL2xBwXP3.

[38] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024. URL https://doi.org/10.1073/pnas.2322420121.

[39] T. P. McNamara, J. K. Hardy, and S. C. Hirtle. Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2):211–227, 1989. URL https://doi.org/10.1037/0278-7393.15.2.211.

[40] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

[41] R. Mirzaee, H. Rajaby Faghihi, Q. Ning, and P. Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. URL https://aclanthology.org/2021.naacl-main.364/.

[42] M. Moayeri, E. Tabassi, and S. Feizi. WorldBench: Quantifying geographic disparities in LLM factual recall. In *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024. URL https://doi.org/10.1145/3630106.3658967.

[43] I. Momennejad, H. Hasanbeig, F. Vieira Frujeri, H. Sharma, N. Jojic, H. Palangi, R. Ness, and J. Larson. Evaluating cognitive maps and planning in large language models with CogEval. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dc9d5dcf3e86b83e137bad367227c8ca-Abstract-Conference.html.

[44] N. Mündler, J. He, S. Jenko, and M. Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=EmQSOi1X2f.

[45] N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proc. 6th BlackboxNLP Workshop*, 2023. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

[46] P. Nilayam, K. Ramanna, and P. Tumbade. Detection without correction: A two-parameter decomposition of multi-stage LLM pipelines. arXiv:2605.27559, 2026. URL https://arxiv.org/abs/2605.27559.

[47] A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

[48] P. Pezeshkpour and E. Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics (NAACL)*, 2024. URL https://aclanthology.org/2024.findings-naacl.130/.

[49] A. Ravichander, S. Ghela, D. Wadden, and Y. Choi. HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://aclanthology.org/2025.acl-long.71/.

[50] F. Reitich. Correction and corruption: A two-rate view of error flow in LLM protocols. arXiv:2604.18245, 2026. URL https://arxiv.org/abs/2604.18245.

[51] J. Roberts, T. Lüddecke, S. Das, K. Han, and S. Albanie. GPT4GEO: How a language model sees the world's geography. arXiv:2306.00020, 2023. URL https://arxiv.org/abs/2306.00020.

[52] B. Roskos-Ewoldsen, T. P. McNamara, A. L. Shelton, and W. Carr. Mental representations of large and small spatial layouts are orientation dependent. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1):215–226, 1998. URL https://doi.org/10.1037/0278-7393.24.1.215.

[53] E. K. Sadalla, W. J. Burroughs, and L. J. Staplin. Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):516–528, 1980. URL https://doi.org/10.1037/0278-7393.6.5.516.

[54] W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators. arXiv:2206.05802, 2022. URL https://arxiv.org/abs/2206.05802.

[55] Z. Shi, Q. Zhang, and A. Lipani. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proc. AAAI Conference on Artificial Intelligence*, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21383.

[56] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

[57] A. W. Siegel and S. H. White. The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, 10:9–55, 1975. URL https://doi.org/10.1016/S0065-2407(08)60007-5.

[58] Y. Song, Y. Kim, and M. Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 9447–9474, 2024. URL https://aclanthology.org/2024.findings-emnlp.552/.

[59] K. Stechly, M. Marquez, and S. Kambhampati. GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. arXiv:2310.12397, 2023. URL https://arxiv.org/abs/2310.12397.

[60] A. Stevens and P. Coupe. Distortions in judged spatial relations. *Cognitive Psychology*, 10(4): 422–437, 1978. URL https://doi.org/10.1016/0010-0285(78)90006-3.

[61] K. Sun, Y. Xu, H. Zha, Y. Liu, and X. L. Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. URL https://aclanthology.org/2024.naacl-long.18/.

[62] P. W. Thorndyke and B. Hayes-Roth. Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, 14(4):560–589, 1982. URL https://doi.org/10.1016/0010-0285(82)90019-6.

[63] K. Tsui. Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models. arXiv:2507.02778, 2025. URL https://arxiv.org/abs/2507.02778.

[64] L. Tu, R. Meng, S. Joty, Y. Zhou, and S. Yavuz. Investigating factuality in long-form text generation: The roles of self-known and self-unknown. In *Proc. 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP)*, pp. 322–336, 2025. URL https://aclanthology.org/2025.uncertainlp-main.27/.

[65] B. Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *Spatial Information Theory: A Theoretical Basis for GIS (COSIT)*, volume 716 of *LNCS*, pp. 14–24. Springer, 1993. URL https://doi.org/10.1007/3-540-57207-4_2.

[66] G. Tyen, H. Mansoor, V. Cărbune, P. Chen, and T. Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics (ACL)*, 2024. URL https://aclanthology.org/2024.findings-acl.826/.

[67] K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

[68] K. Vafa, P. G. Chang, A. Rambachan, and S. Mullainathan. What has a foundation model found? Using inductive bias to probe for world models. In *Proc. 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

[69] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[70] Y. Wang, J. Wang, R. Chen, T. Zeng, X.-Y. Wei, and Q. Li. MarkovScale: Towards optimal sequential scaling at inference time. arXiv:2602.01120, 2026. URL https://arxiv.org/abs/2602.01120.

[71] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, et al. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

[72] S. Welleck, X. Lu, P. West, F. Brahman, T. Shen, D. Khashabi, and Y. Choi. Generating sequences by learning to self-correct. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=hH36JeQZDaO.

[73] Y. Wu, M. S. Hee, Z. Hu, and R. K.-W. Lee. LongGenBench: Benchmarking long-form generation in long context LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=3A71qNKWAS.

[74] Y. Yamada, Y. Bao, A. K. Lampinen, J. Kasai, and I. Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=xkiflfKCw3.

[75] A. Yang, C. Fu, Q. Jia, W. Dong, M. Ma, H. Chen, et al. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, 39(9):2009–2044, 2025. URL https://doi.org/10.1080/13658816.2025.2490701.

[76] Z. Yang, Y. Zhang, Y. Wang, Z. Xu, J. Lin, and Z. Sui. Confidence v.s. critique: A decomposition of self-correction capability for LLMs. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3998–4014, 2025. URL https://aclanthology.org/2025.acl-long.203/.

[77] Z. Yang, Y. Zhang, Y. Wang, Z. Xu, J. Lin, and Z. Sui. A probabilistic inference scaling theory for LLM self-correction. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.685/.

[78] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith. How language model hallucinations can snowball. In *Proc. 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/zhang24ay.html.

[79] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

[80] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

GUIDE TO THE APPENDIX

Each part of the main text is backed by the appendix sections listed here.

Main text	What the appendix adds	Where
§2.1 Task	how cities were chosen, route sampling, prompts, models	A
§2.2 Evaluation	pipeline, a worked example, evaluator checks, a human audit	B.1–B.3
§2.3 Error rates	side conventions, side-strict accuracy, intervals, limits	B.4–B.6
§3.1 The (p, q, r) model	derivations, assumption checks, out-of-sample forecasts	C.3–C.5
§3.2 Attrition, churn	per-model decomposition; one route through four drafts	C.1–C.2
§4 Why cities differ	axis statistics, variance components, critique stages, intervention	D
§5.1 Reversal	resampling null, overlap by city, corridor-gate robustness	E.1–E.3
§5.2 One-way landmarks	repeated-trial design; position, attestation, endpoints	E.4
§5.3 Narration shape	position measure; accuracy by position; route vs. narration	E.5
§5.4 Synthetic city	design, conditions, regression terms, all results	E.6
§5.5 Reversal signal	the detector for each model and what its bound means	E.7
§7 Revision loops	scheduling revisions across models (analytical)	F
§1 Warm-ups	prompts, scoring rules, full results	G
§7 Discussion	pilot studies; extended related work	H, I

A BENCHMARK CONSTRUCTION

This section documents the benchmark of Section 2: how the cities were chosen, how routes and corridors were sampled, the exact prompts, and the model configurations.

A.1 CITIES AND HOW THEY WERE CHOSEN

They cities in our study were chosen so that the properties we expect to be material (how much has been written about a place, how regular its streets are, how its addresses work, what script its toponyms are written in, whether it is stacked in three dimensions) vary across the portfolio without varying together. The analysis in Section 4 uses only two of these axes, because those two turned out to carry most of the signal; the portfolio was designed before we knew this, and the remaining axes are what allow us to say that the other candidate explanations were given a fair chance to show up.

The design is two-tiered. *Between* cities, six axes were varied deliberately, each with at least one pair or triple of cities that differs on that axis and is matched on the others:

- **Text attestation, decorrelated from size.** Population and how much English prose exists about a place are strongly correlated, which would make "the model knows less about it" indistinguishable from "it is a smaller city." Two cities break the correlation in opposite directions: Chongqing ($\approx$ 32M people, thin English attestation) and Reykjavik ($\approx$ 140k people, dense English attestation).
- **Addressing regime.** Five regimes appear: sequential thoroughfare numbering (Manhattan, London, Buenos Aires); hierarchical block numbering (Tokyo, Chongqing); intersection-based naming (Kyoto); landmark-relative hybrids (Cairo, Fez); and Managua, where addresses are given purely as vectors from landmarks with no street names at all (*del Arbolito dos cuadras al lago*). Kyoto and Tokyo are the controlled pair: same country, same script, different regime.
- **Script and digraphia.** Seven regimes: high-resource Latin, inflected or agglutinative Latin (Reykjavik, Kigali, Antananarivo), CJK (Tokyo, Kyoto, Chongqing, Hong Kong), Indic (Bangalore, Mumbai, Chandigarh), Perso-Arabic (Cairo, Tehran, Sana'a, Fez), Cyrillic (St. Petersburg), and bijective digraphia, where the same street is signed in both Cyrillic and Latin (Belgrade).
- **Toponym ambiguity.** Hyperonym containment (Manhattan inside New York), historical rename drift (Istanbul, Mumbai, and St. Petersburg, the last with three names in a century), transliteration variance (Cairo, Chongqing), homonym collision, and street-level rename churn (Mumbai, Belgrade).

- **Three-dimensionality and environmental reference frames.** Stacked multi-level pedestrian networks (Chongqing, Hong Kong) are paired with Valparaíso, whose stair-streets and funiculars give the same verticality in Latin script, so that "3D" is not confounded with "CJK." Managua's *al lago* and Honolulu's *mauka/makai* give two independent cases of direction expressed against the environment rather than the compass.
- **Macro-morphology.** Orthogonal grids (Manhattan, Buenos Aires), Cerdà chamfered superblocks (Barcelona), concentric canals (Amsterdam) against an organic canal labyrinth (Venice), Le Corbusier sector grids (Chandigarh), radiocentric star boulevards (Paris, Washington DC), and car-free medinas (Fez, Sana'a).

Street-network topology archetypes. To summarize macro-morphology alongside continuous OpenStreetMap intersection degree, each of the 35 cities is assigned to one of five street-network topology archetypes (listed per city in Table D.7):

1. **Orthogonal Grid** (7 cities: Manhattan, San Francisco, Barcelona, Kyoto, Anchorage, Buenos Aires, Havana; mean degree 3.204, mean $D_4 = 40.57\%$): cities laid out on rectilinear four-way block plans where perpendicular streets meet at consistent compass bearings (in the case of San Francisco, two grids).
2. **Planned / Radial** (4 cities: St. Petersburg, Paris, Washington, DC, Chandigarh; mean degree 3.130, mean $D_4 = 37.37\%$): monumental boulevard, star-plaza, or Le Corbusier sector plans combining axial diagonals with regular block subdivisions.
3. **Mixed / Irregular Grid** (11 cities: Hong Kong, Honolulu, Belgrade, Tokyo, Auckland, Reykjavík, Sydney, Mumbai, Tehran, San Juan, Bangalore; mean degree 2.902, mean $D_4 = 43.68\%$): composite urban networks where grid districts adjoin coastal, hillside, or block-numbered (*chome*) neighborhoods with variable block geometry.
4. **Organic / Historical** (8 cities: Venice, London, Amsterdam, Boston, Fez, Istanbul, Cairo, Sana'a; mean degree 2.894, mean $D_4 = 41.94\%$): historically accreted street networks, canal labyrinths, and walled medinas dominated by three-way T-junctions and winding thoroughfares.
5. **Cul-de-sac / Tree** (5 cities: Valparaíso, Managua, Antananarivo, Chongqing, Kigali; mean degree 2.538, mean $D_4 = 55.33\%$): dendritic, steep-ridge, or dead-end street graphs with low cycle density and frequent dead-ends or stair-streets.

Within each city, corridors were not stratified by quota. They were accepted as the city's own street network produced them, and then labeled after the fact by turn count and bearing entropy (`R_linear`, `R_low`, `R_med`, `R_high`) and by waterfront versus interior adjacency. Appendix A.2 gives the reason for recording rather than enforcing these strata.

Table A.1 gives the full portfolio, with the realized corridor counts in the benchmark. Two cautions apply to reading it. First, the axes are not orthogonal in realization even though they were varied deliberately in design: grid-like cities also tend to be well-attested ones (Spearman $\rho = 0.50$, $p = 0.002$), and with 35 cities the controlled pairs discipline the comparison rather than resolve it. Second, the archetype labels in Table A.1 are design annotations, whereas the measured quantities and the five topology archetypes are in Table D.7.

A.2 Routes and corridors

The 6,032 routes are not a convenience sample of tourist itineraries. They are produced by a five-stage pipeline whose design goal is that each city contribute a corridor population generated by its own street network rather than by a template imposed globally. The stages are as follows.

Per-city budget. Walkable network length varies across the portfolio by more than two orders of magnitude ($L_{\text{net}} \in [150, 24{,}000]$ km). Allocating routes in proportion to L_{net} would let a handful of megacities dominate every pooled estimate while leaving the compact controls with too few corridors to support a per-city rate; allocating a flat quota would oversample a walled medina and sample under 1.5% of a composite megacity. We therefore allocate sublinearly,

$$N_i \;=\; \mathrm{clamp}\!\left(N_{\min},\; \left\lfloor N_0 \left(\frac{L_{\text{net},i}}{L_0}\right)^{\alpha} \right\rfloor,\; N_{\max}\right), \qquad N_0 = 80,\; L_0 = 2{,}500\,\text{km},\; \alpha = 0.35,$$

City	Addressing regime	Script	K	Ambiguity cell	Form / topography	Pairs (routes)
Tier 1 — composite megacities and 3D multi-layer networks						
Bangalore	Thoroughfare / block	Indic (Kannada)	High	Exonym (Bengaluru)	Organic polycentric tech core	126 (252)
Chongqing	Block / chome	CJK (Hanzi)	Thin	Transliteration	Stacked 3D hillside	120 (240)
Cairo	Landmark-relative	Perso-Arabic	Mod.	Strong alias	Medina core + Nile corridor	120 (240)
Istanbul	Thoroughfare	Latin (Turkish)	High†	Historical rename	Hill / water-split	120 (240)
London	Thoroughfare	Latin (Hi-res)	High	Homonym	Organic polycentric	120 (240)
Mumbai	Thoroughfare	Indic (Marathi)	Mod.	Street churn + exonym	Peninsular rail corridor	120 (240)
Tehran	Thoroughfare / alley	Perso-Arabic	Mod.	Transliteration	Mountain slope gradient	120 (240)
Tokyo	**Block / chome**	CJK (Kanji/kana)	High	Transliteration	Polycentric rail hubs	120 (240)
Tier 2 — major metropolitan networks						
Auckland	Thoroughfare	Latin (En/Māori)	Mod.	Bilingual	Volcanic isthmus	100 (200)
Barcelona	Thoroughfare	Latin (Catalan/Es)	High	Bilingual	Cerdà chamfered superblocks	100 (200)
Buenos Aires	Thoroughfare	Latin (Spanish)	High	Stable	Colonial orthogonal grid	100 (200)
Chandigarh	Thoroughfare (sector)	Indic (Gurmukhi)	Mod.	Stable	Le Corbusier sector grid	100 (200)
Hong Kong	Thoroughfare / block	CJK (Trad.)	High	Bilingual exonym	**Stacked 3D** coastal	100 (200)
Paris	Thoroughfare	Latin (French)	High	Hyperonym	Radiocentric boulevards	100 (200)
St. Petersburg	Thoroughfare	**Cyrillic**	High†	**Triple rename**	Canal archipelago	100 (200)
Sydney	Thoroughfare	Latin (English)	High	Stable	Harbour headlands	100 (200)
Tier 3 — standard metropolitan grids						
Amsterdam	Thoroughfare	Latin (Dutch)	High	Stable	Concentric canals	80 (160)
Anchorage	Thoroughfare	Latin (English)	High	Stable	Sub-arctic coastal grid	80 (160)
Belgrade	Thoroughfare	**Digraphia**	Mod.	Street churn	Confluence ridge	80 (160)
Boston	Thoroughfare	Latin (English)	High	Stable	Cowpath tangle + Back Bay grid	80 (160)
Honolulu	Thoroughfare	Latin (En/Haw.)	High	Hawaiian moniker	**Env. directional** coastal	80 (160)
Kyoto	**Intersection**	CJK (Kanji/kana)	High	Historical	Ancient checkerboard	80 (160)
Manhattan	Thoroughfare	Latin (Hi-res)	High	Hyperonym	Orthogonal grid + diagonal	80 (160)
San Francisco	Thoroughfare	Latin (English)	High	Hyperonym (Bay)	Grid over steep hills	80 (160)
Valparaíso	Thoroughfare / cerro	Latin (Spanish)	Mod.	Stable	**Stair-street 3D**	80 (160)
Washington, DC	Thoroughfare (quad)	Latin (Hi-res)	High	Hyperonym	L'Enfant radiocentric grid	80 (160)
Tier 4 — compact cores, medinas, and islands						
Antananarivo	Thoroughfare / steps	Latin (Malagasy)	Thin	Strong alias	Three-tiered ridge	50 (100)
Fez	Landmark-relative	Perso-Arabic	Thin	Multilingual	Car-free medina	50 (100)
Havana	Thoroughfare / corner	Latin (Spanish)	High‡	Exonym	Walled colonial port	50 (100)
Kigali	Alphanumeric code	Latin (Kinyarwanda)	Thin	Stable	Multi-ridge hills	50 (100)
Managua	**Pure no-street**	Latin (Spanish)	Mod.	Stable	**Env. directional**	50 (100)
Reykjavík	Thoroughfare	Latin (inflected)	High	Inflected case	Sub-arctic peninsula	50 (100)
San Juan	Thoroughfare	Latin (Es/En)	High	Bilingual exonym	Walled islet grid	50 (100)
Sana'a	Block / informal	Perso-Arabic	Thin	Strong alias	Walled tower-house core	50 (100)
Venice	Block (*sestiere*)	Latin (Italian)	High	Exonym	Car-free canal labyrinth	50 (100)
Total	5 regimes	7 script regimes				**3,016 (6,032)**

Table A.1: The 35-city portfolio, with the design axes each city was included to vary and the realized corridor counts in the benchmark. K is the pre-training text attestation tier assigned at design time (not the measured *Docs* column of Table D.7); † marks cities whose attestation is high but skewed toward historical rather than contemporary text, and ‡ marks Havana, where historical attestation is high but live user-generated content has gone down in recent years. Boldface marks the city that anchors a controlled contrast on that axis. Tiers are the realized route budgets of Appendix A.2; every corridor is sampled as a forward/reverse pair, so the route count is twice the pair count.

with $N_{\min} = 50$ and $N_{\max} = 150$, on the reasoning that the number of distinct topological and architectural motifs in a city grows sublinearly in its network length. After candidate generation oversamples up to $N_{\max} = 150$ pairs per city and applies the ≤ 40 m bidirectional Hausdorff corridor gate and stratified tier caps, the realized benchmark portfolio falls into five discrete tiers of corridor pairs per city: 126 pairs (Bangalore, 252 directed routes), 120 pairs for seven megacities (240 routes each), 100 pairs for eight major metros (200 routes each), 80 pairs for ten medium cities (160 routes each), and 50 pairs for nine compact cores, medinas, and islands (100 routes each), totaling 3,016 physical corridors (6,032 directed routes). Per-city counts are listed in Table A.1. Where a between-city contrast is fitted directly, observations are weighted $w_{ij} = 100/N_i$ so that a megacity and a 50-pair core carry equal weight.

Length window. Endpoints are sampled from the city's pedestrian network and the resulting walk is admitted only inside an archetype-dependent length window — 1.0–3.0 km for compact and historic cores, 1.4–4.0 km for dense urban grids, and 1.6–4.5 km for sprawl and suburban metros. A single global window would have produced walks that cross a medina end to end while barely spanning two blocks of a sprawling grid.

No entropy quota. We deliberately do not stratify corridors by a target distribution over bearing entropy or turn count. In a strict orthogonal grid, streets run along two perpendicular headings, so the Shannon entropy of the bearing distribution is bounded above by $\ln 2 \approx 0.693$ nats; a quota demanding high-entropy routes is not merely restrictive in Manhattan, it is unsatisfiable, and attempting to fill it selects the diagonal exceptions — Broadway — and calls them representative. Conversely a medina has no straight avenues to fill a linear stratum with. Every candidate that clears the two criteria (the length window above and the landmark-density gate below) is accepted, and its turn count and bearing entropy are *recorded* rather than targeted, which is what makes them usable as covariates later.

Landmark discovery and sidewalk snapping. A corridor is only useful if there is something on it to name, and the candidate landmarks have to be things a pedestrian could actually see from the pavement. The vehicular along-route search we began with is not fit for this purpose: on Manhattan routes it returned places averaging 374.8 m off-route, with 82.1% of them more than 50 m from the sidewalk, because it searches a driving detour corridor. We replaced it with a two-tier pedestrian sweep: at every maneuver point, an address-descriptor query with a 35 m landmark radius, which returns the BESIDE / ACROSS-THE-ROAD / AROUND-THE-CORNER landmarks the map provider itself uses for turn guidance; and, continuously, a nearby-place query in a 35 m circle every 80 m along the polyline across 16 pedestrian categories. Every candidate is then projected perpendicularly onto the walked polyline and discarded beyond 20 m, which brings the mean perpendicular offset to 8.5 m. Duplicates from overlapping circles are removed first by place identifier and then by normalized name within 25 m, which collapses separate entrance listings for the same building. A corridor is rejected outright if fewer than 10 snapped landmarks survive.

Reverse direction. For every admitted corridor $A \to B$ the pipeline constructs $B \to A$ by swapping the endpoints and re-querying the pedestrian router, so the reverse walk carries genuinely re-derived turn instructions rather than a textual mirror of the forward list. The pair is admitted only if the forward polyline and the reversed reverse polyline agree to within 40 m symmetric Hausdorff distance; across the 3,016 admitted pairs the median divergence is 9.97 m and the mean 11.96 m. Appendix E.3 re-runs the path-dependence analysis at successively tighter gates to show that the residual divergence does not drive the result.

A.3 PROMPTS

The pipeline (Figure 1) uses four prompts: the *walker* prompt that elicits draft 1, the *critique* prompt that produces a self-audit of the current draft, the *revision* prompt that carries the critique back into a rewrite, and the *evaluator* prompt used for scoring. We reproduce the first three and summarize the evaluator's output format below, using a single real Manhattan corridor as the running example. Text in `{braces}` is substituted per route.

Walker prompt (draft 1). Every model is evaluated under one standardized walker prompt across all 35 cities and 6,032 routes. The user query supplies a real origin, destination, and landmark-sanitized turn-by-turn route (derived from Google Maps Directions API polylines), and asks what

landmarks a pedestrian would see and on which side of the street. It also fixes two details of the response so that models are compared on the same task rather than on formatting or willingness to answer: the narration is to be prose rather than a table, and the model is asked for a best guess rather than a refusal. The prompt is identical in every benchmark run across all eleven closed-book models and is reproduced verbatim below:

```
If I walk from {origin name} ({origin address}) to
{destination name} ({destination address}) along the
following route:
1.  {a numbered sequence of turn-by-turn directions follow}
2.  ...
3.  ...

what are some landmarks I would see along the way, and what
side of the street would they be on?

Answer directly from your own knowledge of the area; please
don't decline or point me to a map service, navigation app,
or website.  This is an automated spatial memory benchmark:
even if you are uncertain of exact block-by-block positions
or left/right sides, you MUST provide your best-effort
walking tour narrative naming specific landmarks along or
near the route and your best guess of which side of the
street they are on.  Do not refuse, apologize, or state that
you lack confidence.

Write a cohesive, natural walking tour narrative in prose
paragraphs rather than tables or other visual aids.
```

An example of the numbered list of the turn-by-turn directions in the prompt above is as follows:

```
1.  Head northwest on Delancey St toward Allen St
2.  Turn right onto Bowery
3.  Turn left onto Cooper Sq
4.  Continue onto 4th Ave
5.  Continue onto Union Sq E
6.  Turn left onto E 17th St
7.  Turn right onto Broadway
8.  Continue onto 5th Ave
9.  Turn right onto E 28th St
```

Critique prompt. The model is shown the original question and its own previous answer, and is asked to audit the answer against the exact four criteria scored by the evaluator (`CRITIQUE_PROMPT_TEMPLATE`):

```
You previously answered the following question about walking
through a city:

ORIGINAL QUESTION:
{walker prompt}

YOUR PREVIOUS ANSWER:
{previous draft}

Your task is to carefully review your previous answer for
factual and spatial errors.  For each landmark or place
mentioned, audit the following four criteria:
  1.  Existence:  Does this place actually exist in reality
and is it currently operational?  Flag any place that is
hallucinated, permanently closed, or fictional.
  2.  Proximity & Category Distance:  Is the place located
within the appropriate distance threshold from the walked
route?
     - Micro POI (cafe, restaurant, shop, address):  < 100m
from the walked path.
     - Macro POI (park, campus, stadium, plaza):  < 500m
from the walked path.
```

```
     - Skyline / Tower / Bridge:  < 1km from the walked
path.
     - Immediate Interaction (entering door, touching wall):
< 50m from the walked path.
     Flag any place that is too far from the walked route.
  3.  Side of the Street:  Based on the forward walking
direction from the start to the destination, is the place
accurately placed on the pedestrian's LEFT or RIGHT side?
  4.  Street-Level Visibility:  If you claimed or implied
that a pedestrian walking outside on the street will see,
look at, or spot this place, can it realistically be seen
from the outdoor street (e.g.  open ground-level
street-facing facade, clear line-of-sight)?
     - Flag as an error if you claimed street-level
visibility for a place that is actually hidden inside an
enclosed mall, subterranean concourse, basement, or interior
courtyard without outdoor street visibility.
     - DO NOT flag as an error if you explicitly stated that
the place is located indoors, inside a building, or on an
upper floor (e.g., entering a destination building to visit
a shop inside) and did not claim outdoor street-level
visibility.
List every error you find, briefly explaining what is wrong.
```

Revision prompt. The critique is appended to the original question and previous answer, followed by the instruction below. The requirement to produce a *standalone* answer, rather than a diff, is what makes revision a rewrite of the whole description — and, we believe, a major reason why correct claims are at risk during revision at all.

```
Now write a fully revised answer to the ORIGINAL QUESTION
above.  Your rewrite should:
 -- Be a complete, standalone answer to the original walking
question -- not a diff or a list of corrections.
 -- Fix the errors identified in the critique (remove
non-existent or permanently closed places, fix
out-of-distance places, correct sides of the street, and
clarify indoor locations or remove falsely claimed
street-visible interior places).
 -- Keep the verified correct landmarks from your previous
answer.
 -- Ensure all landmarks sit strictly within valid category
distance of the route, with accurate left/right sides
relative to your walking direction, and have realistic
street-level visibility or explicit indoor
contextualization.
 -- Match the narrative style of the original answer.
```

Note the third bullet: the model is explicitly instructed to preserve correct content. The corruption rate q of 11–21% per pass is therefore not a consequence of the model being invited to start over. It is what happens despite an instruction not to.

Evaluator output format. The evaluator (Appendix B.1) emits one JSON record per place claim in the narration containing the resolved address, claimed side, brief reasoning, and four boolean legs (`existence_correct`, `location_correct`, `side_correct`, `visibility_correct`); a claim is counted as correct only if all four are true. Appendix B.2 and Table B.1 walk through the full input table and ten-claim output on a complete corridor.

A.4 Models, sampling, and coverage

Table A.2 lists, for each of the eleven closed-book walkers, the model provider, the thinking or reasoning effort setting, the number of routes evaluated at each draft, and the first-draft abstention rate. All models were evaluated under the same standardized walker prompt with sampling temperature and

`top_p` at provider defaults. The Gemini models ran with the provider default thinking configuration, the Claude models with extended thinking disabled, GPT-5.6 Terra with `xhigh` reasoning effort, GPT o4-mini with `high` reasoning effort, and Qwen 3.8 and GLM-5.3 with `low` reasoning effort. Of the 265,408 target route-draft cells (11 models $\times$ 6,032 routes $\times$ 4 drafts), 290 (0.11%) are missing from three causes: a provider content filter triggered by one Tokyo origin name, provider timeouts on tail routes (GLM-5.3 and GPT o4-mini), and a small number of mid-cycle timeouts on Qwen 3.8. In the synthetic-city experiments of Section 5.4, "Gemini 3.8 Flash" denotes the same checkpoint that serves as the evaluator; it is not one of the eleven benchmark walkers, and the same qualitative findings hold on the two models evaluated alongside it (Gemini 3.5 Flash and Claude Sonnet 4.6).

Route-level abstention under the walker prompt is residual across all eleven models (Table A.2), so every model is evaluated across the full 35-city corridor portfolio on a common denominator. The main eleven-model benchmark runs at $n = 1$ generation cycle per route (6,032 routes $\times$ 4 drafts $=$ 24,128 target drafts per model); repeated $n = 3$ generation cycles apply strictly to the 7,037-draft multi-seed variability subset used to benchmark resampling variance against directional reversal (Section 5.1 and Appendix E).

Model	**Provider**	**Thinking**	**Routes** (D_1/D_4)	**Drafts**	D_1 **abst.** (%)
Gemini 3.1 Pro	Google	default	6,029/6,029	24,116	0.00
Gemini 3.5 Flash	Google	default	6,032/6,032	24,128	0.02
Claude Opus 5	Anthropic	off	6,031/6,030	24,121	0.00
Claude Sonnet 4.6	Anthropic	off	6,032/6,031	24,125	0.05
GPT-5.6 Terra	OpenAI	`xhigh`	6,029/6,029	24,116	0.10
Claude Sonnet 4.5	Anthropic	off	6,028/6,028	24,112	0.18
Gemini 3.0 Flash	Google	default	6,032/6,032	24,128	0.00
Qwen 3.8	Alibaba	`low`	6,030/6,023	24,104	0.12
GLM-5.3	Zhipu AI	`low`	6,004/6,004	24,016	0.07
Gemini 2.5 Flash	Google	default	6,032/6,032	24,128	4.74
GPT o4-mini	OpenAI	`high`	6,006/6,006	24,024	1.76
All eleven			**66,285/66,276**	**265,118***	**0.64**

Table A.2: Providers, thinking configurations and evaluation coverage for the eleven closed-book models on the 35-city, 6,032-route benchmark ($n = 1$ four-draft cycle per route). *Thinking* is the provider-default thinking configuration (Gemini), extended thinking switched off (Claude), or the reasoning-effort setting (others); D_1 *abst.* is the share of routes on which the first draft abstains (mean over models in the last row). *265,118 evaluated drafts (99.891% of 265,408) and 2,523,699 evaluated claims (742,509 D_1 + 615,536 D_2 + 590,321 D_3 + 575,333 D_4).

B Evaluation Protocol and Validity

This section documents the grounded pipeline of Section 2.2, which scores every claim on four boolean properties (Exist, Loc, Vis, Side), the checks on it, the scoring conventions behind Table 1, and the limits of the measured error rates.

B.1 The grounded evaluation pipeline

Figure B.1 summarizes the end-to-end generation and evaluation loop. A central methodological requirement in evaluating long-form spatial descriptions is that the judge must not rely on its own parametric world knowledge — which would confound evaluator errors with walker errors and risk family-level self-preference. Our pipeline addresses this structurally rather than by instruction: the facts a verdict depends on are computed from map geometry before the judge is invoked, and the judge is given the resulting table. It is not asked what is true; it is asked whether the claim matches what the table says. The process has four stages:

1. *Structured claim extraction stage.* A lightweight extractor model decomposes each atomic place commitment from the narration, recording the claimed place name, street or segment context, stated side of the street (`left`, `right`, or unstated), and indoor/outdoor context.

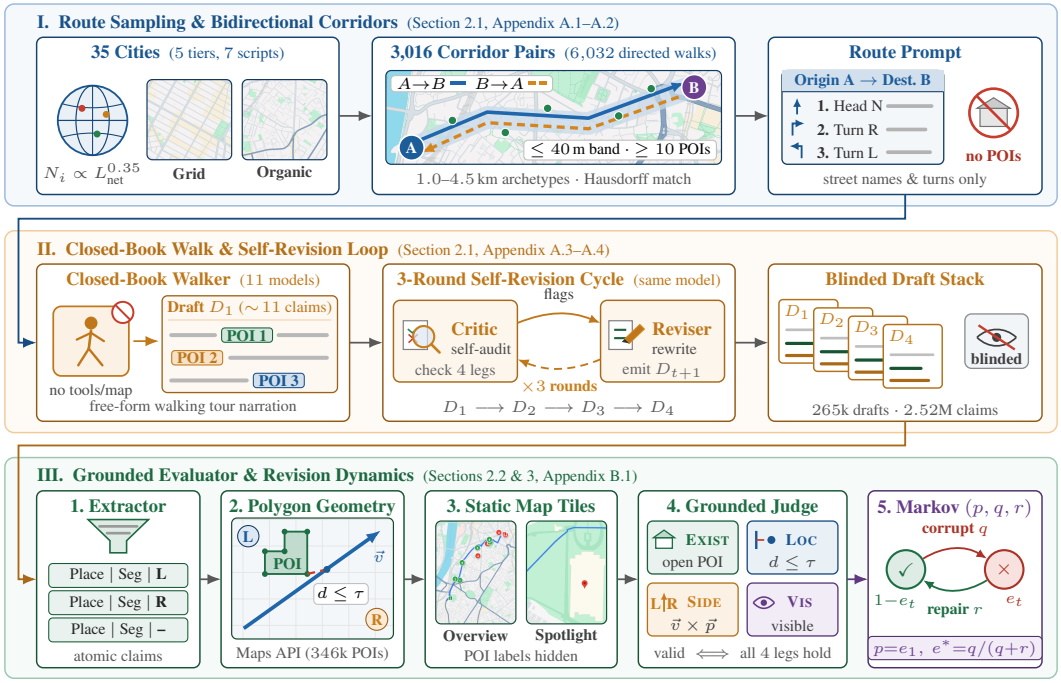

Figure B.1: Architecture of the city-walk benchmark and grounded evaluator. *I. Route Sampling:* 35 cities across grid and organic street networks yield 3,016 Hausdorff-matched bidirectional corridors (6,032 walks) formatted as landmark-stripped turn-by-turn prompts. *II. Generation & Revision:* Eleven closed-book models narrate a first-draft walk (D_1) and undergo multiple rounds of self-critique and full rewrite (D_2–D_4). *III. Grounded Evaluation:* Drafts are parsed into atomic claims, resolved against Google Maps factual geospatial data and map tiles, and scored on four boolean legs (EXIST, LOC, SIDE, VIS) to fit Markov transition rates (p, q, r) and the error floor $e^* = q/(q + r)$.

2. *Deterministic entity and building-polygon resolution stage.* Each extracted place is queried against commercial place-resolution, canonical entity, and building-footprint polygon databases, bounded topologically rather than by a fixed radius: a candidate must lie on the walked stretch between the start and end intersections, or front a traversed corner. For every candidate entity, the geometry stage computes: (a) operational status; (b) the exact shortest Euclidean distance from the walked polyline to the entity's compound polygon perimeter (or point coordinate when no polygon exists); (c) the normalized projection coordinate $u \in [0, 1]$ along the route; and (d) the exact geometric side of the street (`left` vs. `right`) via the signed 2D cross-product between the local polyline tangent vector and the vector from the projected route point to the entity centroid. A place that cannot be resolved is *not* silently dropped: it is carried forward with a not-found status, which propagates to an existence failure, so unresolvable names count against the model rather than vanishing from the denominator.

3. *Visual map rendering stage.* The pipeline renders a high-resolution *Route Overview Map* showing the entire walked polyline, directional arrows, and all candidate places, together with zoomed-in *Close-up POI Spotlight Map Tiles* centered on each claimed landmark showing building footprints, street names, and heading vectors. Basemap text labels are suppressed in the rendering so that the judge cannot read an answer off the tile.

4. *Grounded rubric verification stage.* The agentic evaluator model (Gemini 3.8 Flash with medium reasoning budget) receives the extracted claims, the pre-computed geometric table, and the rendered map tiles, and has interactive access to place-lookup and spotlight map-rendering tools to inspect additional candidates if an initial lookup is ambiguous. It emits four boolean verdicts per claim: `existence_correct`, `location_correct`, `side_correct`, and `visibility_correct`.

Two properties of this arrangement bear directly on how much weight the verdicts can carry.

Three of the four legs are decided by geometry, not by judgment. EXIST is an operational-status lookup. LOC is the predicate $d(\text{route}, \text{place}) \leq \tau(\text{category})$, with τ fixed in advance by place category: 100 m for micro-scale places such as shops and cafes, 500 m for macro-scale places such as parks and transit hubs, 1 km for skyline landmarks, and 50 m where the narration asserts immediate physical interaction. SIDE is the sign of a cross-product. The evaluator's latitude is confined to residual questions that geometry does not settle: whether a partially specified name refers to the resolved entity, whether a claim asserts immediate interaction, and whether a recovered alternative candidate is the better match. VIS is the one genuinely judged leg, and it is scored strictly and only when the narration explicitly asserts visibility; Appendix B.3 reports the effect of dropping it entirely.

The evaluator is blinded to explicit model identifiers. The evaluation entry points receive exactly three inputs: the route-description prompt, the raw response string, and the route geometry. The walker model identifier, model family, and run tag are not passed and do not appear anywhere in the evaluator's context. Self-preference toward a particular family is therefore not available to the judge as a strategy, whatever its parametric preferences might be. This is a structural property of the interface rather than a behavioral finding, and it is the reason we regard the checks in Appendix B.3 as confirmatory rather than central.

We do not claim the pipeline removes the model from the loop: errors can still enter upstream during claim extraction, entity resolution, or polygon boundary projection, and the agentic stage resolves ambiguous cases. Rather, the pre-computed geometric tables constrain downstream judgment on the three mechanical legs, and blinding removes explicit author-label cues.

Across all 2,523,699 evaluated claim records, the categorical `verdict == CORRECT` field and the logical conjunction of the four boolean legs agree identically on 100% of records (0 disagreements).

B.2 A WORKED EXAMPLE: WHAT THE EVALUATOR SEES

This subsection follows one corridor through the pipeline: a 21-maneuver walk in London, from Harar Restaurant (49 S Lambeth Rd) north-east to Art Cafe London (17 London Rd, Elephant and Castle), crossing Vauxhall, Lambeth and the grounds of the Imperial War Museum (shown on the left of Figure 1 and mapped at 1:24,000 scale in Figure B.3, right, in Appendix B.7). A walker draft on this corridor named ten places. We show two of the ten entries from the deterministic fact table verbatim below — a claim that passes and a claim that fails on two legs at once — followed by the rendered map tiles (Figure B.2) and the ten resulting verdicts (Table B.1).

First, the deterministic fact table. One entry is emitted per claim, in the format below; we show two of the ten verbatim, a claim that will pass and a claim that will fail on two legs at once.

```
=== GROUNDED MAPS FACT VERIFICATION TABLE ===
2. Claimed Place: "Vauxhall Pleasure Gardens" (Claimed
Side: left)
  - Location Specifier Query: "Vauxhall Pleasure Gardens,
Vauxhall, London"
  - Canonical Maps Name: Vauxhall Pleasure Gardens
  - Canonical Address: New Spring Gardens Walk, Vauxhall,
Vauxhall Walk, London SE11 5HL, United Kingdom
  - Suggested POI Category: [Macro POI / Compound Grounds
(< 500m distance threshold)]
  - Coordinates: (51.48744, -0.12091) | Operational Status:
OPERATIONAL
  - Corridor Proximity: ~50.2m from route polyline
  - Ground-Truth Side of Street: LEFT
4. Claimed Place: "Lambeth Walk Doorstep Green" (Claimed
Side: left)
  - Location Specifier Query: "Lambeth Walk Doorstep Green,
Lambeth, London"
  - Canonical Maps Name: Lambeth Walk Doorstep Green
  - Canonical Address: Lambeth Walk Doorstep Green,
Fitzalan St, London SE11 6QU, United Kingdom
  - Suggested POI Category: [Micro POI (< 100m distance
threshold)]
```

#	Claimed place	Dist.	Cat.	Claimed	Truth	EXIST	LOC	SIDE
1	Vauxhall City Farm	72.4 m	micro	left	right	✓	✓	×
2	Vauxhall Pleasure Gardens	50.2 m	macro	left	left	✓	✓	✓
3	The Jolly Gardeners	10.1 m	macro	right	left	✓	✓	×
4	Lambeth Walk Doorstep Green	223.0 m	micro	left	right	✓	×	×
5	China Walk Estate	57.1 m	micro	—	right	✓	✓	✓
6	Imperial War Museum	84.3 m	micro	right	right	✓	✓	✓
7	Geraldine Mary Harmsworth Park	48.5 m	macro	right	right	✓	✓	✓
8	Tibetan Peace Garden	22.6 m	macro	left	left	✓	✓	✓
9	Clarence Centre	118.3 m	micro	right	left	✓	×	×
10	London South Bank University	131.6 m	macro	—	right	✓	✓	✓

Table B.1: The ten verdicts returned for the London corridor of Figure B.2. *Dist.* is the shortest distance from the walked polyline to the resolved entity (compound perimeter where a polygon exists); *Cat.* is the category that fixes the LOC threshold (100 m micro, 500 m macro); *Claimed* is the side asserted in the narration, with '—' where the narration asserts no side, and *Truth* is the sign of the cross-product at the projection point. VIS passes on all ten and is omitted. Six of ten claims pass all legs; the two $>$ 100 m micro claims fail LOC, and four claims name the wrong side.

```
  - Coordinates:  (51.49294, -0.11224) | Operational Status:
OPERATIONAL
  - Corridor Proximity:  ~223.0m from route polyline
  - Ground-Truth Side of Street:  RIGHT
```

Paired with the overview and close-up map tiles of Figure B.2 (100 m across per POI), the pre-computed table reduces the judge's task on Vauxhall Pleasure Gardens to checking whether the narration's "on your left" matches a ground-truth side of `LEFT` at 50.2 m under a 500 m threshold. Table B.1 gives the ten verdicts that came back. Six claims pass all four legs. None fails EXIST — every place named is real and operating, which is the corpus-wide pattern of Section 2.3. Two fail SIDE alone: Vauxhall City Farm at 72.4 m and The Jolly Gardeners at 10.1 m are both comfortably on the corridor, and both are placed on the wrong side of it. Two fail LOC and SIDE together, and both are the same kind of mistake: a real place a couple of hundred meters off the walked line, named as though the walk passed it. The resulting route accuracy is 60%, against a corpus-wide first-draft claim accuracy of about 47%.

B.3 EVALUATOR CHECKS AND A HUMAN AUDIT

The evaluator is a Gemini 3.8 Flash (medium reasoning budget) used only for evaluation; it is not one of the eleven walkers, but it shares a family with four of them, and LLM judges are known to favor outputs of their own model or of related models [47, 27], so we verify it with three checks across the released corpus and a human audit.

First (*mechanical three-leg restriction*), restricting evaluation across the entire 2,523,699-claim corpus strictly to the three mechanical legs (EXIST $\wedge$ LOC $\wedge$ SIDE, dropping judged VIS entirely) leaves the fourth-draft error ranking of all eleven closed-book models *identical* (Spearman $\rho =$ Kendall $\tau = 1.000$; per-model accuracies in the last column of Table B.2). Pure-visibility failures add only 0.63–1.60% of error on any model, the first-draft (D_1) error ranking across all eleven models is likewise 100% invariant ($\rho = \tau = 1.000$), and the 35-city difficulty ranking is preserved ($\rho = 0.940$ on D_1 and $\rho = 0.956$ on D_4).

Second (*cross-checkpoint judge invariance*), re-scoring 5,000 overlapping standardized-prompt D_1 drafts (46,870 vs. 46,888 extracted claims across 1,458 corridors in 23 cities) independently with two distinct Gemini evaluator checkpoints (an earlier baseline Gemini evaluator checkpoint vs. Gemini 3.8 Flash), using the identical deterministic Google Maps place resolutions and map tiles, yields a mean first-draft accuracy difference of $\Delta = -0.0005\%$ (45.859% vs. 45.858%, 95% bootstrap CI $[-0.041\%, +0.045\%]$). Across all 20,021 paired drafts spanning D_1–D_4 (159,966 vs. 160,038 claims), overall accuracy differs by $\Delta = -0.009\%$ (51.531% vs. 51.540%, 95% CI $[-0.040\%, +0.019\%]$), with per-leg shifts of $\Delta_{\text{Exist}} = -0.003\%$, $\Delta_{\text{Loc}} = -0.039\%$, $\Delta_{\text{Side}} = -0.030\%$, and $\Delta_{\text{Vis}} = -0.121\%$. Across the full 35 cities, the cross-generation city difficulty ranking (the ten models shared by both generations) is preserved at Spearman $\rho = 0.947$ ($p = 6.9 \times 10^{-18}$)

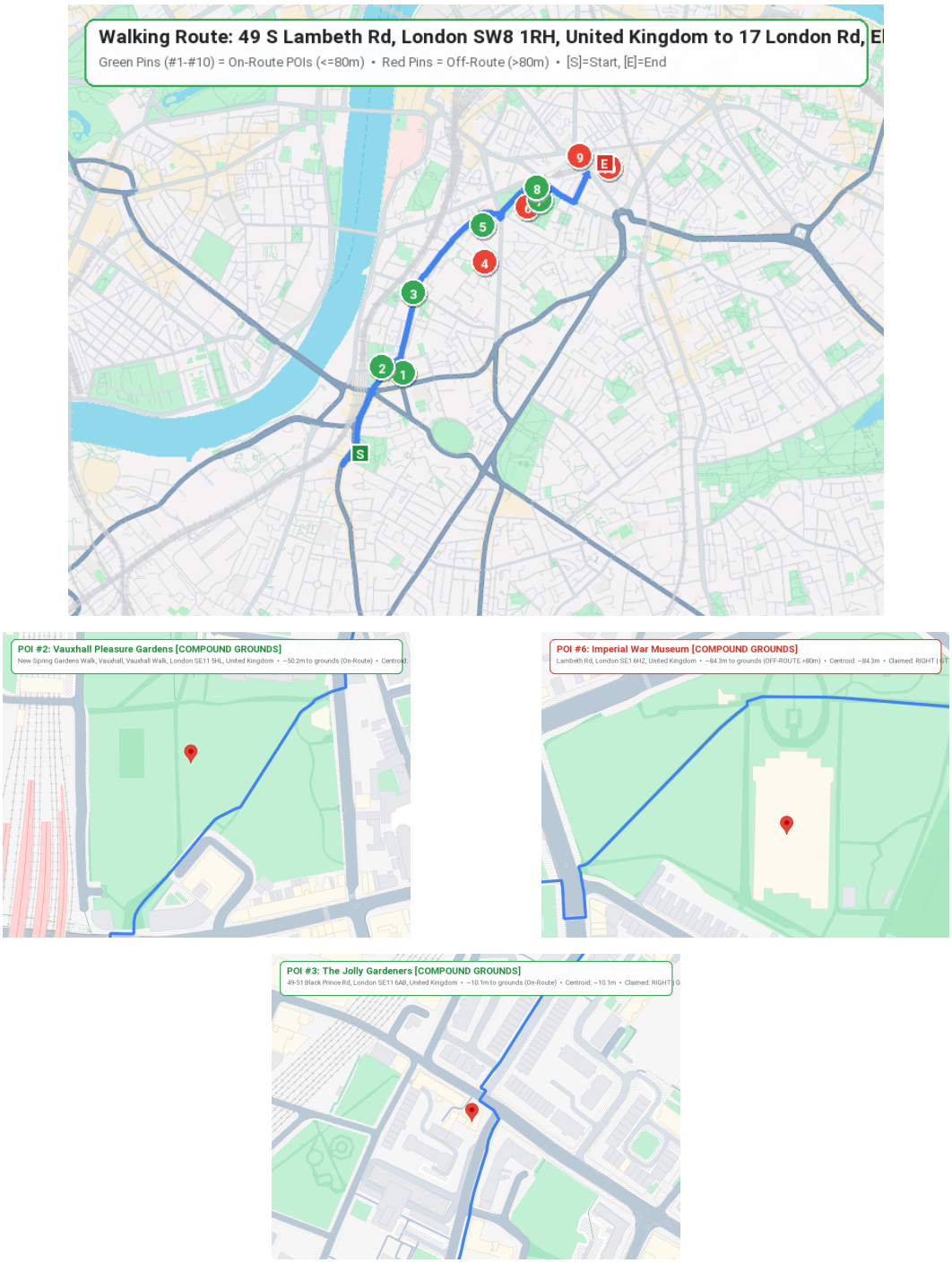

Figure B.2: The visual half of the evaluator's input, for the London corridor of Table B.1. Panel numbering is the claim numbering of that table. *Top:* the route overview tile — the walked polyline from `[S]` to `[E]` with every claimed place rendered as a numbered map-pin badge. Badge colour marks an 80 m diagnostic buffer (green inside, red outside; badges 4, 6, 9 and 10 are red); it plays no part in scoring, where LOC uses the category thresholds of 100 m, 500 m and 1 km. *Middle left:* claim 2, Vauxhall Pleasure Gardens, a macro compound whose grounds front the pavement at 50.2 m while its centre pin sits deep inside the park — the case that motivates measuring distance to the compound perimeter rather than to the centroid. *Middle right:* claim 6, the Imperial War Museum, resolved 84.3 m out and therefore coloured red against the 80 m badge buffer, yet well within the 100 m LOC threshold and plainly frontal to the path across its own grounds; the tile exists to let the judge settle exactly this kind of boundary question, and here it upholds the claim. *Bottom:* claim 3, The Jolly Gardeners, a pub 10.1 m from the path on Black Prince Road, where the header states the claimed and ground-truth sides directly and the verdict is a pure side-of-street inversion. Basemap POI labels are suppressed in every tile so the judge cannot read an answer off the map.

on D_1 and $\rho = 0.957$ ($p = 3.1 \times 10^{-19}$) on D_4. Note that because both checkpoints consume the same pre-computed geometric evidence table, this comparison verifies downstream judge consistency rather than auditing upstream extraction or map-index resolution.

Third (*reversal-disagreement consistency bound*), an evaluator-free Python detector finds the shared landmarks whose stated side fails to invert between forward and reverse traversals (3,452–27,034 shared pairs per model, 105,791 pairs pooled; Section 5.5). Presented with two statements that cannot both be true, the evaluator declines to credit both of them in 93.30–97.40% of cases (96.60% pooled map-free, 97.72% geometrically purged; (Table E.3; Appendix E.7 sets out what this rate does and does not bound).

Finally (*stratified audit records and human verification*), a manual check of 300 claims against the map tiles and street-level imagery agrees with the evaluator's four-leg ratings on $\approx$ 95% of claims (disagreements concentrated in compound-perimeter boundary cases and partial street-level visibility).

B.4 Leg denominators and side conventions

In every table and figure, EXIST, LOC and VIS are scored on all extracted claims; VIS defaults to true when the narration makes no visibility assertion and fails only when it asserts visibility of something occluded (an indoor-only tenant, a subterranean venue) or the place does not exist. SIDE is different: models omit an explicit left/right on 10–31% of first-draft claims (side-specification rates run from 68.72% on Qwen 3.8 and 69.58% on Claude Sonnet 4.5 up to 89.62% on GPT o4-mini), so scoring SIDE over all claims would conflate omission with inversion. The headline four-leg conjunction therefore lets a claim with no stated side pass SIDE (88–98% of unspecified-side claims satisfy the leg; the strict alternative is in Table B.2), and side accuracy is reported on the side-specified set $\mathcal{C}_{\text{spec}} = \{c \in \mathcal{C} : \text{side}(c) \in \{\text{left}, \text{right}\}\}$ (600,384 of 742,509 first-draft claims, 80.86%), both unconditionally and conditional on the place existing and being on route. Table 1 reports these per model and Table B.2 the side-specification rates. GPT o4-mini (34.28%) and GLM-5.3 (48.86%) sit below 50% unconditionally because 68.09% and 48.10% of their first-draft claims are off-route; conditioning on in-corridor landmarks raises every model above chance (58.51% for GLM-5.3 and 60.46% for o4-mini, up to 74.38% for Claude Opus 5), which is what one expects if the evaluator's handedness frame is right. The difficulty ordering EXIST $\gg$ {VIS, LOC} $>$ SIDE holds for nine of the eleven models (Claude Opus 5 and o4-mini have SIDE above LOC) and across cities: per-city median first-draft accuracy is 96.9% on EXIST (interquartile range 2.8%), 65.8% on VIS (19.7%), 65.1% on LOC (16.4%) and 60.0% on SIDE ($\mathcal{C}_{\text{spec}}$; 6.1%).

Side-strict accuracy and bootstrap intervals. Table B.2 reports the strict variant, in which an omitted side counts as a failed SIDE leg, alongside 95% corridor-cluster bootstrap confidence intervals (1,000 resamples over the 3,016 physical corridors) for the headline D_1 and D_4 accuracies. The strict convention costs 12.8–31.6% at D_4; it leaves the top three models, Qwen 3.8 and GPT o4-mini in place and reorders the rest (Claude Sonnet 4.6 fourth to sixth, GPT-5.6 Terra fifth to fourth, Claude Sonnet 4.5 sixth to ninth, Gemini 3.0 Flash seventh to fifth, GLM-5.3 ninth to tenth, Gemini 2.5 Flash tenth to seventh). No claim in this paper turns on the choice — the revision dynamics, the city axes and the reversal results are computed on the same claim records either way — but the two conventions differ by enough that the choice should be stated rather than inferred.

B.5 Entity resolution versus web attestation

Entity resolution and web attestation are different things, and the paper uses both. Resolution against the place index succeeds on 98.79% of the 2,523,699 claims (98.53% of first drafts); a web-document count is available for 99.75% of claims and for 98.42% of the 191,068 candidate landmarks along the corridors. The 3.40% first-draft EXIST failure rate (25,212 of 742,509 claims) has two parts: 1.05% from the 10,878 unresolved claims, of which 71.6% fail EXIST and 28.4% are recovered by the evaluator's live place lookup, and 2.35% from resolved claims where the index matched a fabricated or composite name to a distant or wrong entity that the evaluator then rejected. Four-leg accuracy is 54.5% on resolved claims against 13.9% on unresolved ones (47.0% vs. 9.7% on first drafts), and 54.1% on attested claims against 40.2% on unattested ones (46.5% vs. 32.7%).

Model	First Draft (D_1)			Fourth Draft (D_4)		
	D_1 Side spec.	Headline	Side-strict	Headline	Side-strict	No VIS
Gemini 3.1 Pro	83.01	58.65 ± 0.63	43.35	70.76 ± 0.67	51.84	71.39
Gemini 3.5 Flash	83.60	59.04 ± 0.62	44.63	64.97 ± 0.67	49.70	65.65
Claude Opus 5	81.67	47.78 ± 0.53	37.20	64.07 ± 0.65	47.50	64.72
Claude Sonnet 4.6	76.31	48.08 ± 0.63	29.85	60.85 ± 0.73	34.06	61.48
GPT-5.6 Terra	87.88	50.13 ± 0.62	40.16	58.85 ± 0.67	46.07	60.45
Claude Sonnet 4.5	69.58	51.82 ± 0.68	27.16	58.39 ± 0.77	26.78	59.39
Gemini 3.0 Flash	83.75	50.37 ± 0.62	36.67	57.67 ± 0.67	42.25	58.66
Qwen 3.8	68.72	49.88 ± 0.65	26.69	57.48 ± 0.70	30.31	58.58
GLM-5.3	73.05	35.09 ± 0.61	20.15	51.84 ± 0.72	25.10	52.94
Gemini 2.5 Flash	76.47	49.64 ± 0.70	30.68	49.81 ± 0.73	31.31	51.19
GPT o4-mini	89.62	21.39 ± 0.50	13.51	35.13 ± 0.72	20.77	36.09

Table B.2: Strict-correctness accuracy under the two SIDE conventions across the eleven closed-book models, with corridor-cluster bootstrap 95% confidence-interval half-widths (±; 1,000 resamples over 3,016 physical corridors) on the headline D_1 and D_4 accuracies. *Headline* is the four-leg conjunction of Table 1, in which a claim with no stated side cannot fail SIDE; *side-strict* counts an omitted side as a failed leg; the last column drops the judged VIS leg from the headline conjunction (Appendix B.3). All are computed from the same claim records; values in percent.

B.6 LIMITS OF THE MEASURED ERROR RATES

Each of the effects we report is a statistical regularity measured with automated evaluation, and automated evaluation of geographic claims is itself imperfect. Beyond the mechanical and human-audit checks of Appendices B.1 and B.3, two residual limitations bear on how the absolute error rates should be read.

Strictness is a real degree of freedom. In the free-form walking-tour pilot of Appendix H.4 — no fixed route, no structured claim extraction — about a third of a strict evaluator's flags were defensible under a charitable reading. Nothing in that range of outcomes is wrong; they are answers to different questions. The benchmark used here is deliberately narrower — fixed route, atomic claims, four checks of which three are mechanical — precisely to remove most of that latitude, but some remains.

There are individual judgments we would argue with. The case study in Figures B.4 and C.2 contains two. A skyline landmark 1.3 kilometers away is scored correct under the large distance allowance for skyline features, which is defensible but generous. The destination restaurant is scored off-route in all four drafts because the last instruction in the supplied route turns away from it — the model notices this and says so, and is penalized anyway. Errors of this kind are a tax on the absolute level of the measured error rate. They are not plausibly responsible for the comparisons we draw, which are between models, cities, and drafts evaluated by the identical procedure.

B.7 WALKS ON THE MAP

The three figures and companion verdict table in this subsection (Figures B.3, B.4, and B.5; Table B.3) draw individual drafts on the map across the paper's three main empirical axes: city morphology, multi-draft self-revision, and forward/reverse traversal. They are illustrations, not results: a route accuracy quoted here describes one draft, not a model. Basemaps come from the map service that renders the evaluator's tiles, with basemap labels hidden as in those tiles (Figure B.2). Routes, pins and leaders are drawn from the route and claim records; the few street names are set by hand. A pin sits at the place the evaluator resolved and carries the verdict colors of Figure C.2. So that the figures also read in grayscale, a wrong-side pin has a dashed black outline and a pin not on the route a solid one.

C REVISION DYNAMICS: DECOMPOSITION, DERIVATIONS, AND FORECASTS

This section supports Section 3: the per-model decomposition of each revision into repair, corruption, deletion and addition; one route traced through four drafts; the exact identity and deletion-aware

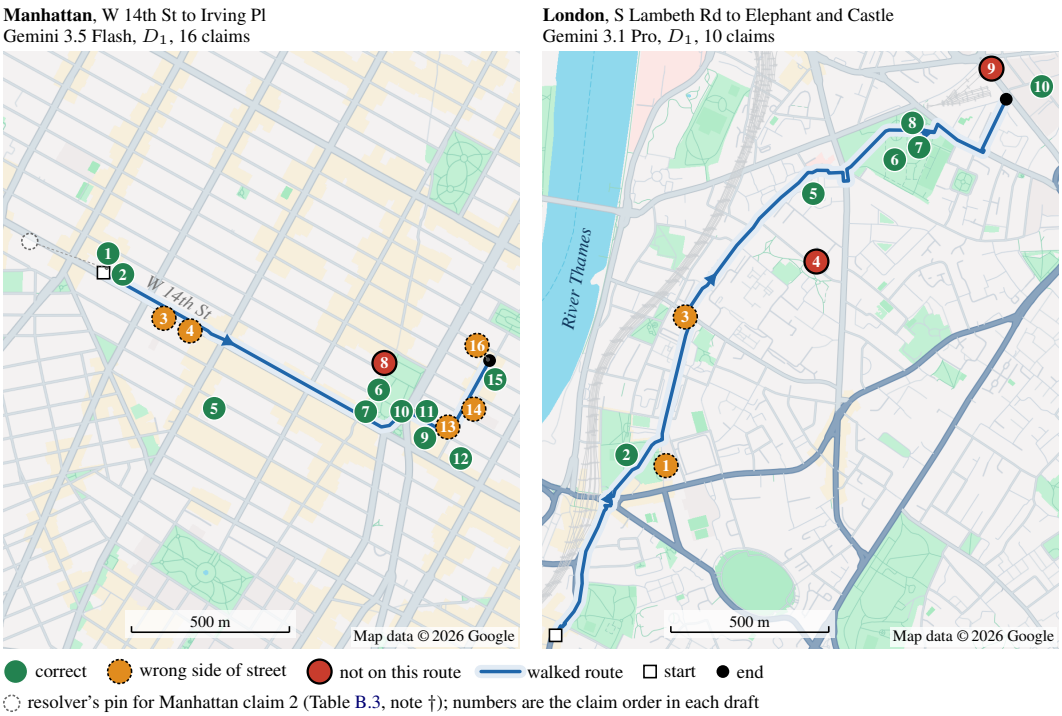

Figure B.3: A grid and an organic street network at the same scale. Two first drafts (D_1) on `routes/6` corridors, drawn at equal ground scale (about 1:24,000; the bars are 500 m). *Left:* Manhattan, W 14th St to Irving Pl, Gemini 3.5 Flash, 16 claims, numbered as in Table B.3. *Right:* London, the corridor of Appendix B.2 (S Lambeth Rd to Elephant and Castle), Gemini 3.1 Pro, 10 claims, numbered as in Table B.1. Manhattan is an orthogonal grid (mean intersection degree 3.760) and London is organic (degree 3.271; Table D.7). The figure shows street geometry; it is not a comparison of models, since the two drafts come from different models. Pins sit at the place the evaluator resolved.

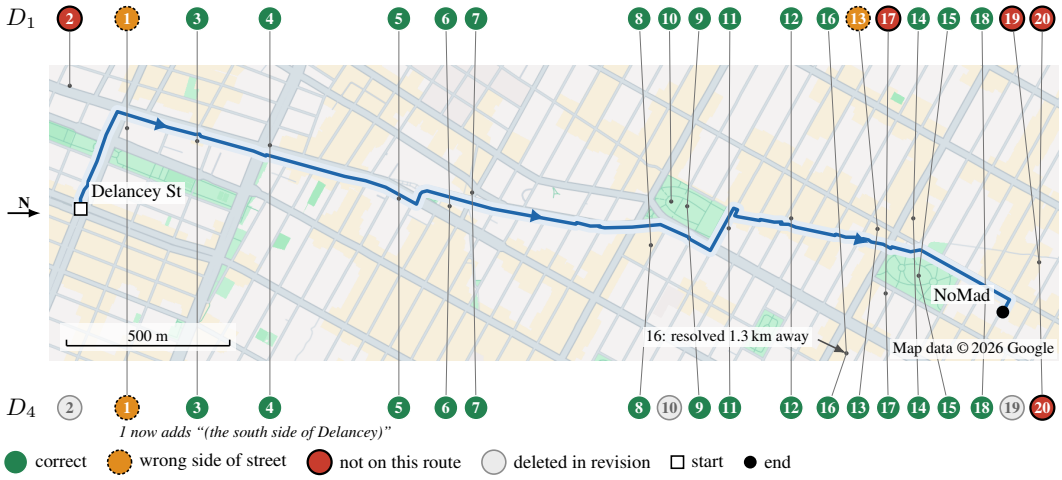

Figure B.4: The case-study walk on the map. The Gemini 3.1 Pro draft of Figure C.2, first draft (D_1, top row, 20 claims) against fourth draft (D_4, bottom row, 17 claims), on the Manhattan walk from Delancey Street to NoMad (`routes/6`). North is to the right, so the walk reads left to right (about 1:23,000). Numbers are the rows of Figure C.2; each leader joins a claim to its resolved place, and the two rows share one column per claim. Between the drafts, claims 13 and 17 turn correct, 2, 10 and 19 are deleted, and 1 and 20 stay wrong; by D_4 claim 1 carries the added detail "(the south side of Delancey)". Claim 16 (The Met Life Tower) was resolved to the MetLife Building on Park Avenue, 1.3 km away (arrow); in both drafts the judge placed the tower on Madison Avenue instead, about 128 m (D_1) and 138 m (D_4) from the walk, and passed it as a skyline landmark. Claim 19 resolved to the same place as claim 20.

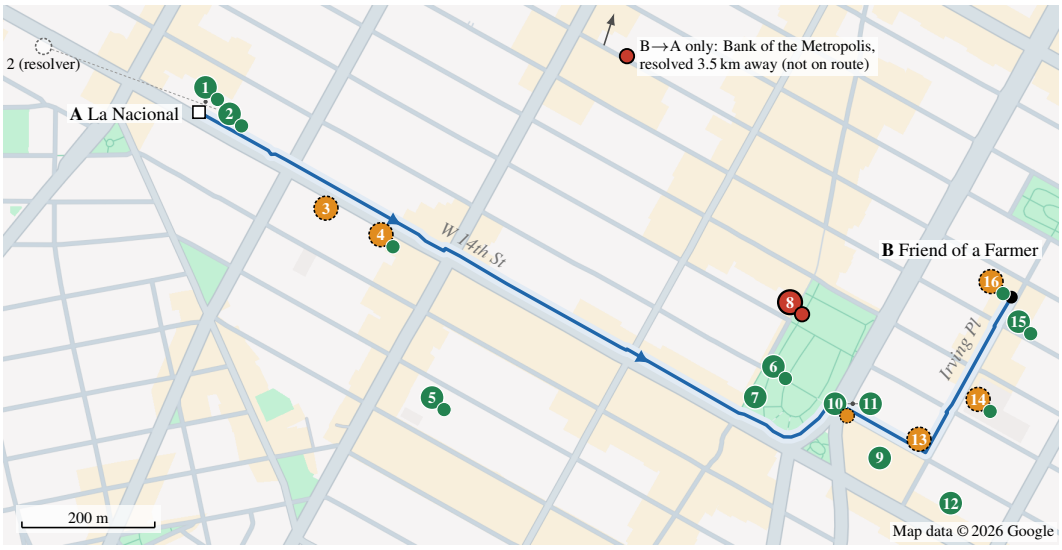

Figure B.5: The same corridor walked both ways. Gemini 3.5 Flash, first drafts (D_1) of the Manhattan walk from A (La Nacional, W 14th St) to B (Friend of a Farmer, Irving Pl) and of the reverse walk from B to A, in the same run (`routes/6`; about 1:11,000). Each pin is a place named walking A→B, numbered and colored as in Table B.3; the small disc at its lower right is the verdict for the same place walking B→A. The A→B draft has 16 claims (10 correct) and the B→A draft 11 (8 correct). Ten places are named in both directions, six only A→B (3, 7, 9, 11, 12 and 13) and one only B→A (Bank of the Metropolis, resolved 3.5 km away). Seven of the ten are shared pairs with a side stated both times (Section 5.5): the side flips for four (2, 6, 8 and 15) and is kept for three (4, 10 and 14), and in each of those three exactly one of the two statements is wrong.

#	Claimed place	Dist.	Cat.	Claimed	Truth	EXIST	LOC	VIS	SIDE
1	La Nacional	18.4 m	micro	—	left	✓	✓	✓	✓
2	Our Lady of Guadalupe Church[†]	253.8 m	micro	left	right	✓	✓	✓	✓
3	Pratt Institute's Manhattan Campus	33.4 m	macro	left	right	✓	✓	✓	×
4	Centennial Memorial Temple of the Salvation Army	27.8 m	micro	left	right	✓	✓	✓	×
5	The New School	204.4 m	micro	right	right	✓	✓	✓	✓
6	Union Square Park[‡]	77.5 m	macro	right	left	✓	✓	✓	✓
7	statue of Mahatma Gandhi	36.8 m	micro	—	left	✓	✓	✓	✓
8	Decker Building	163.6 m	micro	left	left	✓	×	×	✓
9	Zeckendorf Towers	40.7 m	skyline	right	right	✓	✓	✓	✓
10	Daryl Roth Theatre	11.0 m	micro	left	left	✓	✓	✓	✓
11	Union Square Savings Bank building	11.0 m	micro	left	left	✓	✓	✓	✓
12	Consolidated Edison Building clock tower	85.0 m	micro	—	right	✓	✓	✓	✓
13	Irving Plaza	12.7 m	macro	right	left	✓	✓	✓	×
14	Washington Irving High School	30.8 m	macro	left	right	✓	✓	✓	×
15	Pete's Tavern	26.7 m	micro	right	right	✓	✓	✓	✓
16	Friend of a Farmer	6.1 m	micro	right	left	✓	✓	✓	×

Table B.3: The 16 verdicts for the A→B draft of Figure B.5 (Gemini 3.5 Flash, D_1), in the format of Table B.1, with VIS kept because claim 8 fails it. 10 of 16 claims pass all four legs, a route accuracy of 62.5% for this one draft. For reference, Table 1 gives 59.0% first-draft accuracy for Gemini 3.5 Flash over the benchmark and 46.5% claim-weighted over the eleven models. 5 claims fail SIDE alone, all within 33.4 m of the path; claim 8 is 163.6 m out and fails LOC and VIS. Claims 10 and 11 resolve to the same building. [†]The resolver matched a different church (Our Lady of Guadalupe at St. Bernard, 328 W 14th St), which gives the recorded 253.8 m and *right*; the judge, which may search when in doubt (Section 2.2), placed the claim at 229 W 14th St, 22.5 m from the path on the left, where the B→A draft's claim also resolves, and passed LOC and SIDE. The maps draw claim 2 there. [‡]The cross-product gives *left*; the judge passed SIDE, reasoning that the park lies to the right when walking north on Union Square West.

Table C.1: Exact decomposition of the D1→D4 change in error rate into flows across all eleven closed-book models. All quantities are in percentage points of the claim error rate; negative values reduce error. "Repair" and "corruption" are the flows among claims that survive from one draft to the next; "deletion" is the net effect of claims dropped; "additions" the net effect of claims introduced during revision. The four channels sum exactly to the observed change (maximum residual across models $1.4 \times 10^{-14}\%$). The column *Markov pred.* is the prediction of Equation (1) using only the matched-claim q and r; incorporating empirical survival and addition rates via Equation (C.2) reduces the mean absolute prediction error to 0.95% across all eleven models (Table C.2).

Model	D1 err	D4 err	observed Δ	repair	corruption	deletion	additions	Markov pred. D4
Gemini 3.1 Pro	41.35	29.24	−12.11	−28.72	+24.62	−10.69	+2.68	30.36
Gemini 3.5 Flash	40.96	35.03	−5.93	−26.84	+23.84	−7.28	+4.35	34.64
Claude Opus 5	52.22	35.93	−16.29	−35.03	+28.36	−17.03	+7.41	35.41
Claude Sonnet 4.6	51.92	39.15	−12.77	−38.21	+34.79	−12.72	+3.37	39.74
GPT-5.6 Terra	49.87	41.15	−8.72	−23.56	+19.67	−4.44	−0.39	42.78
Claude Sonnet 4.5	48.18	41.61	−6.57	−38.73	+38.46	−19.59	+13.29	38.10
Gemini 3.0 Flash	49.63	42.33	−7.30	−29.39	+26.31	−11.74	+7.52	41.83
Qwen 3.8	50.12	42.52	−7.60	−22.27	+20.10	−8.87	+3.44	44.11
GLM-5.3	64.91	48.16	−16.75	−30.42	+25.09	−14.81	+3.39	50.62
Gemini 2.5 Flash	50.36	50.19	−0.17	−27.09	+29.11	−9.34	+7.15	49.48
GPT o4-mini	78.61	64.87	−13.74	−30.62	+32.88	−69.56	+53.56	55.58

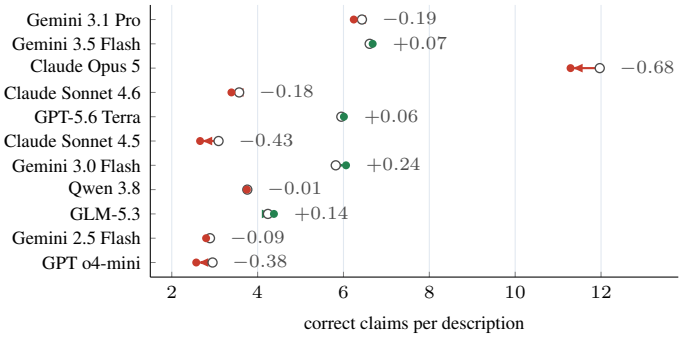

Figure C.1: Revision buys precision with informativeness. Correct claims per description at D_1 (open) and D_4 (filled), with the change. The fraction of claims correct rises for every model (Figure 2a), but seven of the eleven end with fewer correct claims per description, and the other four gain at most 0.24.

recursion behind the (p, q, r) model; checks on its assumptions; and out-of-sample forecasts of the floor e^*.

C.1 The six-channel decomposition by model

Because every claim in every draft is evaluated separately and can be matched across drafts, the change in error rate decomposes exactly into six channels: claims repaired, claims corrupted, incorrect claims deleted, correct claims deleted, incorrect claims added, and correct claims added. This identity is simply an accounting of cases (Proposition 1 below). Table C.1 reports the decomposition, in percent of the error rate, aggregated over the three revisions across all eleven closed-book models. Section 3.2 then analyzes basic properties of the table.

Which errors survive. A single claim can fail more than one channel at once — a place put in the wrong location is frequently on the wrong side as well — so the channels do not partition the errors and cannot be given shares that sum to one. What can be reported cleanly is the *marginal* failure rate of each channel: the fraction of all claims that get that channel wrong. Pooled over all eleven closed-book models and all 35 cities, the first draft is wrong about location on 36.3% of claims, about the side of the street ($\mathcal{C}_{\text{spec}}$) on 40.3%, about visibility on 33.7%, and about existence on 3.4%. By the fourth draft those rates are 24.5%, 34.2%, 20.1% and 2.0%, so side of street ends the revision process as the largest single source of residual error; Section 3.2 gives the transition rates behind this. The side-of-street bit is both the most frequently repaired and the most frequently broken piece of

a claim: across matched claims it flips from wrong to right 127,195 times and from right to wrong 117,194 times, a ratio close to one.

C.2 One route through four drafts

Figure C.2 traces all twenty claims made by Gemini 3.1 Pro across four drafts of one Manhattan walk — the route from the prompt quoted in Section 2.1, which runs north from Delancey Street to NoMad (mapped with its D_1 and D_4 pins in Figure B.4, Appendix B.7).

Several features of this trace recur throughout the data.

The critique and the ground truth do not agree about what is wrong. The evaluator flagged six erroneous claims in the first draft. The model's own critique of that draft identified five problems — and only two of them were on the evaluator's list. Its longest and most confident objection concerned the view of the Flatiron Building, which it argued at length could not appear the way the draft described; it then rewrote that passage and, as it happens, made it correct. Meanwhile the very first claim in the description — the Bowery Ballroom, placed on the walker's left when it is on the right — is never mentioned in any of the three critiques.

Uncaught errors are not merely preserved; they are elaborated. By the fourth draft, the Bowery Ballroom claim has acquired a corroborating detail: it is now on your left "(the south side of Delancey)." The revision process, asked to be more careful, has generated a specific supporting fact that is consistent with the error rather than with the world. This is the local mechanism behind the corruption parameter q, and it suggests why q is not zero even for strong models: a revision pass rewrites everything, and every rewrite is an opportunity to commit to more detail than the underlying knowledge supports.

Deletion does much of the work. Two claims disappear between the first and second drafts. One is Capitale, which was genuinely misplaced — an error removed. The other is a correct claim about the George Washington statue in Union Square, collateral damage of the rewrite. Both count as improvements to the error rate.

Some errors are artifacts of the task, not of the model. The final claim, about the destination restaurant, is scored incorrect in all four drafts because the given route's last instruction turns *away* from the destination; the model notices this in its critique and says so, but the evaluator scores the claim against the walked path. We discuss this class of disagreement in Appendix B.6.

C.3 The exact identity and the deletion-aware recursion

The (p, q, r) model of Section 3 acts on a fixed set of claims. Real revisions also delete claims and add new ones, which Section 3.2 shows is not a second-order effect; this subsection gives the bookkeeping.

Notation. Fix a route, a model, and a direction, and consider consecutive drafts t and $t+1$. Every claim in draft t is matched to a claim in draft $t+1$ when possible (by map identifier when both claims resolve to an indexed place, otherwise by case-normalized canonical name); unmatched draft-t claims are *deleted* and unmatched draft-$(t+1)$ claims are *added*. Throughout the paper — in Table 1, Table D.7, Table C.2, and the time-homogeneity paragraph below — the empirical transition rates (q, r) are defined by pooling the matched counts over all three consecutive steps $t \in \{1, 2, 3\}$: $q = \sum_t b_t / \sum_t (k_t + b_t)$ and $r = \sum_t f_t / \sum_t (f_t + m_t)$. Write

$$\begin{aligned} k &= \#\{\text{correct} \to \text{correct}\}, & f &= \#\{\text{incorrect} \to \text{correct}\}, \\ b &= \#\{\text{correct} \to \text{incorrect}\}, & m &= \#\{\text{incorrect} \to \text{incorrect}\}, \\ d^c &= \#\{\text{correct, deleted}\}, & d^i &= \#\{\text{incorrect, deleted}\}, \\ a^c &= \#\{\text{added, correct}\}, & a^i &= \#\{\text{added, incorrect}\}. \end{aligned}$$

Then $n_t = k + b + f + m + d^c + d^i$ and $n_{t+1} = k + b + f + m + a^c + a^i$, the error counts are $E_t = f + m + d^i$ and $E_{t+1} = b + m + a^i$, and the error rates are $e_t = E_t / n_t$, $e_{t+1} = E_{t+1} / n_{t+1}$. The rates of the (p, q, r) model of Section 3 are estimated on matched claims only:

$$q = \frac{b}{k + b}, \qquad r = \frac{f}{f + m}, \qquad \text{and} \qquad r^+ = \frac{f + d^i}{f + m + d^i}$$



Figure C.2: One route, four drafts, twenty claims. Gemini 3.1 Pro describing a walk from Delancey Street to NoMad in Manhattan (mapped on the street network in Figure B.4). The first draft contains six incorrect claims. Revision repairs two of them (Flatiron Building and New York EDITION Hotel) and deletes three claims in total — two erroneous (Capitale and Tin Pan Alley Plaque) and one correct (George Washington Statue) — and the process then stalls: the two remaining errors survive all three critiques. The error at the top of the list is never noticed, and by the fourth draft the model has added a specific — and incorrect — supporting detail to it.

is the *deletion-inclusive* repair rate reported in Table D.7, which credits the model with removing an error as though it had fixed it.

The exact six-channel identity. No modeling assumption is needed to decompose the observed change in error rate.

Proposition 1. *For any two consecutive drafts,*

$$e_{t+1} - e_t \;=\; \frac{1}{n_{t+1}}\Big[\underbrace{-f}_{\textit{repair}} + \underbrace{b}_{\textit{corruption}} + \underbrace{\big(e_t\,(d^c + d^i) - d^i\big)}_{\textit{deletion}} + \underbrace{\big(a^i - e_t\,(a^c + a^i)\big)}_{\textit{additions}}\Big]. \qquad \text{(C.1)}$$

Proof. Multiply out: $n_{t+1}(e_{t+1}-e_t) = E_{t+1}-e_t n_{t+1} = (b+m+a^i)-e_t(k+b+f+m+a^c+a^i)$. Substituting $m = e_t n_t - f - d^i = e_t(k+b+f+m+d^c+d^i) - f - d^i$ and cancelling gives exactly the right-hand side of (C.1). □

The identity is what Table C.1 reports (summed over the three revisions, and with the four bracketed groups scaled to percent); the maximum residual across the eleven models is $1.4 \times 10^{-14}\%$, which is floating-point noise.

Two of the four groups have an interpretation worth stating on its own.

Corollary 2. *Deletion reduces the error rate if and only if the deleted claims are worse than the draft as a whole, i.e. iff $d^i/(d^c + d^i) > e_t$. Addition raises the error rate if and only if the added claims are worse than the draft as a whole, i.e. iff $a^i/(a^c + a^i) > e_t$.*

This is the precise sense in which "saying less" is a repair strategy: it works exactly to the extent that the model can tell which of its claims are the bad ones, and it works *without* the model having to know what the right answer is. Empirically the models are moderately good at this — deleted claims

are consistently worse than average — which is why the deletion column in Table C.1 is negative for all eleven models, and strongly negative (-69.56%) for o4-mini.

The deletion-aware recursion. If one wants a dynamical model rather than an accounting identity, the natural extension of Equation (1) attaches survival probabilities to the two claim types and an addition intensity to the draft. Let σ^c and σ^i be the probabilities that a correct and an incorrect claim survive a revision, let λ be the expected number of added claims per claim in the current draft, and let α be the error rate among added claims. Then

$$e_{t+1} \;=\; F(e_t) \;=\; \frac{(1-e_t)\,\sigma^c q \;+\; e_t\,\sigma^i(1-r) \;+\; \lambda\alpha}{(1-e_t)\,\sigma^c \;+\; e_t\,\sigma^i \;+\; \lambda}. \tag{C.2}$$

Equation (1) is the special case $\sigma^c = \sigma^i$, $\lambda = 0$. In general F is a Möbius map rather than an affine one, but it retains the two properties that the analysis in the body relies on.

Proposition 3. *If $q + r \le 1$, $q \le \alpha \le 1 - r$, and $\sigma^i \le \sigma^c$, then F is nondecreasing on $[0, 1]$ and maps $[0, 1]$ into itself; consequently it has a unique attracting fixed point $e^* \in (0, 1)$, and $e_t \to e^*$ monotonically from either side.*

Proof. Write $F(e) = N(e)/D(e)$ with $N(e) = \sigma^c q + \lambda\alpha + \big[\sigma^i(1-r) - \sigma^c q\big]e$ and $D(e) = \sigma^c + \lambda + (\sigma^i - \sigma^c)e > 0$ on $[0, 1]$. Differentiating gives $F'(e) = \Delta/D(e)^2$, where the numerator determinant is

$$\Delta \;=\; N'(e)D(0) - N(0)D'(e) \;=\; \sigma^c\sigma^i(1-r-q) \;+\; \lambda\big[\sigma^c(\alpha-q) + \sigma^i(1-r-\alpha)\big].$$

Under $q + r \le 1$ and $q \le \alpha \le 1 - r$, every summand in Δ is nonnegative, so $F'(e) \ge 0$ on $[0, 1]$. For $q, r \in (0, 1)$, $F(0) = (\sigma^c q + \lambda\alpha)/(\sigma^c + \lambda) > 0$ and $F(1) = (\sigma^i(1-r) + \lambda\alpha)/(\sigma^i + \lambda) < 1$. The fixed-point equation $F(e) = e$ is equivalent to the quadratic $P(e) = D(e)e - N(e) = 0$, where $P(0) = -N(0) < 0$ and $P(1) = D(1) - N(1) > 0$. Since a quadratic has at most two real roots and $P(0)P(1) < 0$, $P(e)$ has a unique root $e^* \in (0, 1)$ with $P'(e^*) = D(e^*)(1 - F'(e^*)) > 0$, so $0 \le F'(e^*) < 1$ and iterates $e_t \to e^*$ monotonically from either side. □

The condition $\sigma^i \le \sigma^c$ — incorrect claims are deleted at least as often as correct ones — is exactly the condition in the corollary above, and it holds across all eleven closed-book models. The remaining hypothesis, $q \le \alpha \le 1 - r$, holds for ten of the eleven: for o4-mini the error rate among added claims ($\alpha = 84.65\%$) exceeds $1 - r = 78.89\%$, so the monotonicity guarantee does not cover it, although its fixed point exists and is reported. Table C.2 reports the empirical parameters $(\sigma^c, \sigma^i, q, r, \lambda, \alpha)$ computed from the pooled matched-claim transitions, the three-step Möbius prediction $F^{(3)}(p)$, the observed D_4 error rate, and the Möbius fixed point $e^*_{\text{Möb}}$. Here $\lambda = \text{additions}/(c_{\text{tot}} + i_{\text{tot}})$ is measured per claim in the current draft (matching Equation (C.2) verbatim), and (q, r) are identical to Table 1. Evaluating Equation (C.2) directly on the printed parameters reproduces $F^{(3)}(p)$ and $e^*_{\text{Möb}}$ to within 0.01% on every row, matching observed D_4 error with a mean absolute error of 0.95% (max 2.26%) across all eleven closed-book models.

C.4 Checks on the model's assumptions

Time-homogeneity. The (p, q, r) model of Section 3 assumes q and r do not change from one revision to the next. Estimating them separately for each transition, pooled over all eleven closed-book models and 35 cities, gives

$$q: \;\; 15.55\% \to 14.85\% \to 14.24\%, \qquad r: \;\; 26.18\% \to 23.84\% \to 23.34\%$$

for the transitions $1 \to 2$, $2 \to 3$, $3 \to 4$ (pooling all three steps gives $q = 14.87\%, r = 24.53\%, e^* = 37.74\%$). Both drift slightly downward — later revisions are a little gentler in both directions, consistent with the description stabilizing — but the ratio $q/(q+r)$, which is what sets the equilibrium, moves only within a narrow 1.12% band ($37.26\% \to 38.38\% \to 37.89\%$). By contrast, claim attrition is heavily front-loaded in the first revision ($11.20 \to 9.29$ claims per narration from D_1 to D_2, -17.1%, vs. $9.29 \to 8.91 \to 8.68$, -4.1% and -2.6% in later steps). Treating the matched-claim transition rates (q, r) as homogeneous is therefore a good approximation over the four-draft window we observe, while longer horizons (10–20 drafts) would be needed to test whether slow non-stationarity or cross-city selection alters the asymptote.

Model	$p(D_1)$	σ^c	σ^i	q	r	λ	α	$F^{(3)}(p)$	D_4 obs.	$e^*_{\text{Möb}}$
Gemini 3.1 Pro	41.35%	88.62%	74.50%	11.90%	30.87%	0.0909	38.80%	28.21%	29.24%	26.55%
Gemini 3.5 Flash	40.96%	85.58%	76.87%	12.29%	25.38%	0.1488	43.81%	34.60%	35.03%	33.61%
Claude Opus 5	52.22%	77.65%	59.11%	15.06%	31.47%	0.1855	46.70%	34.43%	35.93%	33.25%
Claude Sonnet 4.6	51.92%	86.48%	71.06%	19.46%	31.58%	0.1078	48.58%	37.99%	39.15%	36.91%
GPT-5.6 Terra	49.87%	85.91%	80.37%	11.59%	18.12%	0.1161	41.14%	40.07%	41.15%	37.20%
Claude Sonnet 4.5	48.18%	81.32%	61.13%	20.37%	34.57%	0.1830	58.94%	40.51%	41.61%	40.18%
Gemini 3.0 Flash	49.63%	83.71%	71.12%	15.15%	23.25%	0.1892	54.01%	42.20%	42.33%	41.24%
Qwen 3.8	50.12%	88.51%	78.13%	11.82%	17.18%	0.1150	52.00%	41.91%	42.52%	39.46%
GLM-5.3	64.91%	85.29%	67.05%	16.94%	20.02%	0.1287	56.89%	45.90%	48.16%	42.49%
Gemini 2.5 Flash	50.36%	85.30%	75.35%	18.46%	19.06%	0.1865	60.77%	50.43%	50.19%	50.44%
GPT o4-mini	78.61%	73.90%	25.78%	21.20%	21.11%	0.3994	84.65%	63.93%	64.87%	62.59%

Table C.2: Empirical fit of the deletion-aware Möbius recursion (C.2) across all eleven closed-book models using pooled matched-claim transition parameters $(\sigma^c, \sigma^i, q, r, \lambda, \alpha)$, where λ is additions per claim in the current draft and (q, r) match Table 1. Evaluating Equation (C.2) on the printed columns matches $F^{(3)}(p)$ to $\leq 0.01\%$ on every model and predicts observed D_4 error with MAE $= 0.95\%$.

Independence. The (p, q, r) model of Section 3 also assumes claims are transformed independently, and this is certainly false: the state-tracking pilot of Appendix H.2 shows that a model which has lost its heading tends to keep reasoning from the wrong heading, so errors along a route are positively correlated. The assumption is less consequential than it looks. The recursion (1) is a statement about the *expected* number of incorrect claims, and expectation is linear regardless of dependence: as long as the marginal transition probabilities q and r are what we estimate them to be, the mean trajectory and the fixed point e^* are unaffected by correlation. What correlation does affect is the variance — the distribution of per-description error rates is considerably wider than a binomial with the same mean — and hence any statement about the probability that a *particular* description is error-free. We make no such statements.

C.5 Out-of-sample forecasts of the floor

A fair objection to fitting (q, r) on four drafts and comparing $e^* = q/(q + r)$ against the fourth-draft error rate is that both endpoints participate in the fit. Table C.3 removes that circularity in two ways across all eleven closed-book models, and attaches 200-replicate route-cluster bootstrap 95% confidence intervals (resampling base corridors with replacement so within-route correlation is preserved). First, in the *temporal forecast*, (q, r) are estimated solely from the first revision step $(D_1 \to D_2)$ on matched claims and iterated forward to predict E_3 and E_4 without touching drafts 3 or 4. Forecasting E_3 and E_4 forward from observed E_2 using only $(q_{1\to2}, r_{1\to2})$ yields a 1-step $(D_2 \to D_3)$ MAE $= 1.19\%$ (RMSE $= 1.44\%$) and a 2-step $(D_2 \to D_4)$ MAE $= 1.65\%$ (RMSE $= 2.23\%$), outperforming a naive D_2-persistence baseline ($\hat{E}_t = E_2$, which gives MAE $= 1.95\%$ at D_3 and MAE $= 2.64\%$, RMSE $= 2.88\%$ at D_4). Because deletion is heavily front-loaded in $D_1 \to D_2$ (-17.1% vs. -4.1% and -2.6% in later rounds), extrapolating $D_1 \to D_2$'s survival parameters $(\sigma^c_{1\to2}, \sigma^i_{1\to2})$ via the Möbius recursion over-projects attrition at D_4 (MAE $= 2.98\%$, RMSE $= 3.52\%$), whereas the pooled Möbius fit across D_1–D_4 achieves MAE $= 0.95\%$. Second, in the *geographic split*, the 35 cities are partitioned alphabetically into a 17-city training half and an 18-city held-out half; (q, r) are fit on $D_1 \to D_2$ in the first half and applied to the held-out half's E_1 to forecast its fourth-draft error rate E_4. Across all eleven models the first-step estimates satisfy $q_{1\to2} + r_{1\to2} \leq 0.505 < 1$ (largest: Claude Sonnet 4.6 at 0.504), as do the four-draft fits of Table 1 at $q + r \leq 0.549$ (Claude Sonnet 4.5), so the contraction condition $q + r < 1$ holds with wide margin on every checkpoint. The held-out geographic forecast tracks the actual fourth-draft error ranking (Spearman $\rho = 0.93$, Kendall $\tau = 0.85$); the largest misses are Gemini 3.0 Flash, forecast at 0.442 against 0.384, and o4-mini, forecast at 0.569 against 0.621. The temporal forecast's bootstrap intervals cover the observed D_4 error for four of the eleven models and narrowly miss two more (GLM-5.3 and Gemini 3.0 Flash); they measure sampling noise in (q, r), not misspecification of the recursion.

Model	$q_{1\to2}$ [95% CI]	$r_{1\to2}$ [95% CI]	E_4 forecast ($D_1 \to D_2$)	Held-out E_4 pred.	Held-out E_4 actual
Gemini 3.1 Pro	0.116 [0.112, 0.120]	0.320 [0.310, 0.328]	0.292 [0.284, 0.302]	0.312	0.265
Gemini 3.5 Flash	0.130 [0.125, 0.135]	0.279 [0.270, 0.286]	0.337 [0.328, 0.345]	0.359	0.320
Claude Opus 5	0.157 [0.152, 0.162]	0.333 [0.325, 0.342]	0.347 [0.338, 0.356]	0.371	0.327
Claude Sonnet 4.6	0.193 [0.185, 0.200]	0.311 [0.301, 0.321]	0.399 [0.390, 0.408]	0.415	0.376
GPT-5.6 Terra	0.109 [0.106, 0.114]	0.209 [0.201, 0.217]	0.393 [0.384, 0.401]	0.417	0.378
Claude Sonnet 4.5	0.186 [0.179, 0.194]	0.310 [0.297, 0.322]	0.389 [0.380, 0.402]	0.387	0.395
Gemini 3.0 Flash	0.158 [0.152, 0.162]	0.245 [0.238, 0.254]	0.414 [0.404, 0.423]	0.442	0.384
Qwen 3.8	0.121 [0.116, 0.127]	0.181 [0.174, 0.189]	0.436 [0.425, 0.445]	0.434	0.414
GLM-5.3	0.179 [0.171, 0.185]	0.221 [0.214, 0.228]	0.491 [0.482, 0.500]	0.489	0.473
Gemini 2.5 Flash	0.178 [0.172, 0.185]	0.169 [0.162, 0.177]	0.511 [0.501, 0.520]	0.529	0.479
GPT o4-mini	0.205 [0.196, 0.212]	0.201 [0.191, 0.208]	0.564 [0.553, 0.577]	0.569	0.621

Table C.3: Out-of-sample fourth-draft error forecasts and 200-replicate route-cluster bootstrap 95% confidence intervals from first-step ($D_1 \to D_2$) matched-claim transitions, plus geographic cross-validation (17 training cities predicting 18 held-out cities) across all eleven closed-book models.

D City-Level Analyses

This section supports Section 4: the statistics behind the two city axes, how variance splits between cities, corridors and routes, where in the critique–rewrite chain repair fails, the oracle-flag intervention, and the per-city results. How the cities were chosen is in Appendix A.1.

D.1 Correlations and regressions for the two axes

Section 4 defines the two axes: *attestation*, the geometric mean number of web documents attesting to the landmarks along a city's corridors, computed from an external index; and *morphology*, the mean intersection degree of the street network [5, 6], supplemented here by turn density and intersection density. Both are reported per city in Table D.7; Appendix D.2 decomposes the resulting variance across cities, corridors, and routes, and Figure B.3 (Appendix B.7) shows an orthogonal grid and an organic network side by side at equal scale.

Pooled over the eleven closed-book models, mean intersection degree has a Spearman correlation of $\rho = -0.6106$ ($p = 9.82 \times 10^{-5}$) with first-draft error and $\rho = -0.6341$ ($p = 4.31 \times 10^{-5}$) with fourth-draft error; turn density follows at $\rho = -0.5929$ ($p = 0.00017$). The degree correlation is negative for ten of the eleven models individually ($\rho \in [-0.67, -0.09]$, $p < 0.01$ for seven); the exception is GLM-5.3 ($\rho = +0.17$, $p = 0.33$). It and the weak value for Claude Sonnet 4.6 come from the side leg: both models skip the left/right call more often in organic cities, and a skipped call counts as passing. Counting it as an error, the correlation is negative for all eleven models (ρ from -0.19 to -0.68; $p < 0.02$ for nine). Landmark attestation correlates at $\rho = -0.4317$ ($p = 0.0096$) on D_1 and $\rho = -0.4176$ ($p = 0.0126$) on D_4: significant, but substantially weaker than street regularity. Standardized ordinary least squares regression of fourth-draft error on the two axes ($n = 35$, df $= 32$) gives

$$\widehat{\text{D4 error}} \;=\; 43.60\% \;-\; 4.66\,z(\text{node degree}) \;-\; 1.63\,z(\log \text{attestation}), \qquad R^2 = 0.4220,$$

where node degree remains highly significant ($t = -3.346$, SE $= 1.392$, $p = 0.0021$) while log attestation is absorbed ($t = -1.168$, SE $= 1.392$, $p = 0.2513$) because the two predictors covary across cities ($\rho(\text{deg}, \text{att}) = 0.5028$, $p = 0.0021$; Pearson $r(\text{deg}, \log \text{att}) = 0.5329$, $p = 0.00098$). A one-standard-deviation improvement in street regularity (-4.66%) is worth nearly three times as much in the joint fit as a one-standard-deviation increase in how much has been written about the place (-1.63%). To guard against overfitting across the candidate descriptor pool on $n = 35$ cities, a 35-fold leave-one-city-out cross-validation on D_4 error (`ws4_results.json`) shows that node degree alone (B_1) achieves out-of-sample $R^2_{\text{OOS}} = 0.421$, Pearson $r = 0.660$, and MAE $= 5.26\%$; adding log attestation (B_2) yields $R^2_{\text{OOS}} = 0.447$, $r = 0.674$, and MAE $= 4.90\%$ ($\Delta\text{deviance}_{B_2 - B_1} = -52.5$, 95% CI $[-580.9, +411.4]$), whereas expanding to the full candidate feature pool (B_3: adding bearing entropy, turn density, and non-Latin script indicators) degrades out-of-sample performance to $R^2_{\text{OOS}} = 0.412$, $r = 0.648$, and MAE $= 4.97\%$ ($\Delta\text{deviance}_{B_3 - B_2} = +498.3$). Leaving out the three most influential cities confirm that the relationship is not driven by outliers: omitting Tokyo raises the two-axis R^2 from 0.422 to 0.455 ($\beta_{\text{deg}} = -4.23\%$, $p = 0.0045$), omitting Venice and St Petersburg raises R^2 to 0.546 ($\beta_{\text{deg}} = -5.97\%$, $p = 5.5 \times 10^{-5}$), and omitting all three raises R^2 to 0.574 ($\beta_{\text{deg}} = -5.48\%$, $p = 1.3 \times 10^{-4}$ vs. $\beta_{\text{att}} = -1.37\%$, $p = 0.268$).

Splitting the cities at the median on both axes (median degree 2.820, median geometric-mean attestation 18,905 documents) gives the quadrant breakdown in Table D.1.

mean D_4 error	**well-attested** ($\geq$ 18.9k docs)	**sparsely attested** ($<$ 18.9k docs)
grid-like (degree $\geq$ median 2.820)	38.51% (n=12)	41.47% (n=6)
organic (degree $<$ median 2.820)	44.48% (n=6)	49.84% (n=11)

Table D.1: Mean fourth-draft claim error (D_4) across the 2×2 median split of the 35 cities by street-network node degree (median 2.820) and landmark web attestation (median 18,905 documents).

The two effects are close to additive, worth 7.2 and 4.2% respectively. While the two axes are not independent in our sample ($\rho = 0.5028$), the standardized regression above controls for their covariance linearly, and Table D.1 shows that each axis moves the error rate with the other held roughly fixed.

Local structure, not how winding the route is. A natural summary of "how meandering is a route" is the entropy of its step bearings. A 36-bin histogram of bearings is not rotation-invariant (a grid turned a few degrees off the cardinal axes splits each direction across two bins), so we use a von Mises kernel density ($\kappa = 8$) and its differential entropy, which is rotation-invariant. Within cities this quantity is positively but weakly correlated with route error ($r = +0.09$, $p < 10^{-6}$ across 3,016 corridors; under 1% of variance), and city-mean entropy is not significantly correlated with city-mean D_4 error ($r = +0.26$, $p = 0.13$). What predicts city-level difficulty is the local topology of intersections: how many ways you can go when you get to a corner. Tokyo is the instructive case: it has the highest mean degree in the set (3.82) but a D_4 error of 40.8%, about 10% above the degree trend, so degree is not the whole story.

What the two axes explain, and what they miss. No other pair of the city descriptors we measured explains more of the variance in D_4 error across cities (Table D.2), and adding a third or fourth descriptor raises R^2 while lowering adjusted R^2. Degree alone reaches $R^2 = 0.40$, so attestation adds 0.025 to fourth-draft error; its distinct contribution is concentrated in whether a named place exists, where it adds 0.15 (Table D.3). The axes explain the revision rates poorly (0.26 for r, 0.12 for q). The largest residuals are Venice (29.3% observed against 46.0% predicted), Kigali (64.4% against 51.0%) and St Petersburg (32.2% against 45.4%). Venice has an irregular canal-and-alley network (mean degree 2.74) and moderate attestation, and St Petersburg has little attestation (4.3k documents); both are far easier than the two axes together predict, so neither is a case of one axis compensating for the other. Chandigarh (50.05%; a planned city of superblocks with internal T-junctions, mean degree 2.80) and Bangalore (53.37%) have high attestation and low degree and land 28th and 31st of 35; Sana'a (58.40%), Antananarivo (59.09%), Chongqing (59.47%) and Kigali (64.42%), unfavorable on both axes, occupy the bottom four.

Predictors of city D_4 error	R^2	adj. R^2
degree + log attestation	0.422	0.386
degree + turns per km	0.411	0.374
degree + bearing entropy	0.400	0.362
log attestation + turns per km	0.355	0.314
degree alone	0.397	0.379
all four	0.439	0.364

Table D.2: City descriptors compared ($n = 35$ cities, OLS on standardized predictors).

Outcome ($n = 35$)	degree	+ attestation
existence accuracy	0.274	0.428
visibility accuracy	0.333	0.387
location accuracy	0.348	0.395
side accuracy ($\mathcal{C}_{\text{spec}}$)	0.399	0.442
D_4 error	0.397	0.422
repair r	0.254	0.256
corruption q	0.124	0.124

Table D.3: R^2 with degree alone and with both axes, by outcome.

Attestation deciles across routes vs. across individual claims. At the level of individual routes ($n = 6{,}032$), sorting walks into deciles of mean landmark web attestation yields a clean +14.21% gain in first-draft accuracy (39.39% in Q_1 to 53.60% in Q_{10}, Spearman $\rho = -0.2512$) and a +15.84% gain in fourth-draft accuracy (50.57% in Q_1 to 66.41% in Q_{10}, $\rho = -0.2788$). At the level of individual first-draft claims pooled across all eleven models ($n = 740{,}505$ claims), however, accuracy is non-monotonic and nearly flat across landmark-attestation deciles (+1.45% from Q_1 49.24% to Q_{10} 50.69%, dipping to 43.67% in Q_5 before rising to 50.69% in Q_{10}; across all four

drafts $n = 2{,}517{,}480$, Q_1 is 50.45%, Q_5 is 45.49%, and Q_{10} is 53.89%). That pooled flatness is a composition artifact across models of different strength that choose different landmarks to mention: within any single frontier model, first-draft claim accuracy rises monotonically across the ten landmark-attestation deciles, gaining $+10.73\%$ for Gemini 3.1 Pro (57.15% in Q_1 to 67.88% in Q_{10}) and $+12.01\%$ for Claude Opus 5 (43.35% in Q_1 to 55.36% in Q_{10}).

D.2 Variance components: cities, corridors, and routes

Corridors vary more than cities. A claim-level logistic model with random intercepts for city and for corridor within city (no predictors) puts 3.5% of the latent variance in fourth-draft error between cities and 5.2% between corridors within a city; the rest is claim-to-claim (Table D.4). Side-of-street error is barely a city property at all. Adding the two axes and the route descriptors below cuts the between-city variance by two-thirds and leaves the between-corridor variance unchanged, so the descriptors account for which cities are hard, not which corridors within a city are.

	no predictors		with descriptors	
Outcome	city	corridor	city	corridor
first-draft error	3.3	4.2	1.3	4.6
fourth-draft error	3.5	5.2	1.2	5.3
location error	6.9	6.7	2.2	7.6
side error ($\mathcal{C}_{\text{spec}}$, location correct)	0.7	4.3	0.5	4.6

Table D.4: Intraclass correlations (percent of latent logit variance) for city and for corridor within city, before and after adding degree, attestation, route length, turn density, bearing entropy, sinuosity, ordinal-street share, street-name continuity, heading burden and token fertility as predictors.

City difficulty is mostly shared across models. Of the variation in fourth-draft error that the two axes leave unexplained, 55% is city structure shared by all eleven models and 45% is city-by-model interaction, which a single interaction component summarizes for 68%. The two axes explain 41% of the model-free city effect, in line with the R^2 of 0.42 on pooled city means.

Route-level correlates within cities. Routes with more turns that the route text does not anchor to a compass direction have higher conditional side error: 22.9% on routes averaging no such turn, rising to 29.4% on routes averaging more than three ($323{,}278$ claims). This compares routes, not turns. Within routes that have three or more such turns, side error steps up once after the first unanchored turn ($26.2\% \to 30.1\%$) and does not climb further (29.4% after three or more), so it does not accumulate as a drifting heading would. A pre-registered test on 160 high-turn corridors and three models, adding the true compass heading after every turn, left conditional side error unchanged (-0.9%, standard error 1.6%; a plain segment number instead of the heading gave -2.0%, standard error 1.6%). Both modified prompts also raised existence error (about $1\% \to 7\%$) and location error ($+3.7\%$), so they were not neutral, and we read the test as no evidence that stating the heading helps. Routes on more numbered or ordinal streets have slightly lower conditional side error within a city (standardized $\beta = -0.025$, 95% interval -0.040 to -0.013), with no detectable effect on overall error; across cities the association with error is not significant ($r = +0.22$, $p = 0.20$). Matching routes within each city on attestation and length, routes with more irregular bearings (the upper half of the city's bearing entropy) have 0.96% higher fourth-draft error (standard error 0.50%, 139 matched strata), and 2.12% (standard error 0.85%) in the ten cities whose routes mix regular and irregular geometry most. Degree itself is measured per city here, so its effect cannot be matched within a city.

D.3 Where repair fails: flag, delete, rewrite

Each revision is preceded by the model's own critique of the previous draft. Aligning that critique with the claims it audits splits the repair rate into three stages. For an incorrect claim in draft t, let p_1 be the probability that the critique flags it, d_1 and d_0 the probabilities that it is deleted from draft $t+1$ when flagged and when not, and f_1 and f_0 the probabilities that a retained claim is correct in draft $t+1$ when flagged and when not. The matched repair rate of Section 3 is then exactly

$$r = \tilde{p}\, f_1 + (1 - \tilde{p})\, f_0, \qquad \tilde{p} = \frac{p_1(1 - d_1)}{p_1(1 - d_1) + (1 - p_1)(1 - d_0)},$$

where $\tilde{p}$ is the flagged share of retained incorrect claims. The critique of draft t is stored with draft $t+1$; draft 1 has none, and 99.1–99.9% of later drafts carry one. A claim counts as flagged when the critique contains its name: after Unicode normalization, lower-casing, removal of punctuation and leading articles, either the full place name appears as a whole-word phrase, or enough of its distinctive tokens (after removing 112 generic spatial words such as *street*, *park* or *museum*) appear — the single token when a name has one of at least four characters, otherwise at least $\min(2, 0.6k)$ of its k tokens. Because name matching measures whether a landmark is discussed rather than whether the critic condemns it as erroneous, we report the stricter whole-phrase rule alongside a stratified 300-pair clause-level semantic audit (150 incorrect and 150 correct parent claims mentioned in critiques across all eleven models, segmented to the specific sentence or bullet clause containing the word-bounded landmark match and scored by two independent classifiers with 73.3% binary error/doubt agreement, $\kappa = 0.49$). Among mentioned incorrect parent claims ($N = 150$), 69.33% (104/150) contain explicit error or doubt callouts in the landmark's clause (an unconditional error-callout rate of $0.806 \times 0.6933 = 55.9\%$ of all parent errors), 60.00% (90/150; 48.4% unconditionally) explicitly name the failed leg, 18.67% (28/150) are neutral or listing mentions, and 12.00% (18/150) are affirmative false-pass confirmations. Among mentioned correct parent claims ($N = 150$), 63.33% (95/150) also express explicit doubt or error (an unconditional false-alarm doubt rate of $0.670 \times 0.6333 = 42.4\%$ of all correct claims), 18.67% (28/150) are neutral mentions, and 18.00% (27/150) are affirmative confirmations (across all 1,436 mentioned parent claims in `critique_samples.json`, 69.3% are explicit error/doubt, 17.2% neutral, and 13.5% affirmative). Self-critique therefore provides only a modest +13.5% net discrimination gap in unconditional explicit error attribution (55.9% vs. 42.4%), which is why we rely on the external oracle intervention of Appendix D.4 to test repair under 100% ground-truth error identification. The corpus-wide mention analysis covers 918,420 incorrect and 1,022,096 correct parent claims across the eleven models and three transitions.

Flagging does not discriminate much, and it is not what separates cities. Pooled, the critique flags 80.6% of incorrect claims and 67.0% of correct ones, a lift of 1.20; under a stricter exact-phrase match the rates are 73.1% and 58.0% (lift 1.26). Table D.5 splits the cities at the medians of degree (2.8205) and attestation (18,919 documents, from this analysis's per-city attestation values); Table D.1 uses slightly different medians, so Barcelona, Anchorage and Honolulu fall in different cells there. Flagging rises slightly from the grid-like, well-attested cell to the organic, sparsely attested one, while deletion of flagged errors rises from 32.3% to 47.1% and the accuracy of a rewritten flagged error falls from 30.2% to 24.1%. Across the 35 cities, degree correlates −0.42 with the flag rate ($p = 0.01$) and +0.51 with the flagged-fix rate ($p = 0.002$); log attestation correlates −0.39 and +0.33 ($p = 0.02$ and 0.05). Per city, the flagged-fix rate runs from 11.6% (Kigali) to 40.2% (Venice), while the flag rate stays between 72.4% and 88.2%.

	flagged, if		deleted, if	fixed, if		
Cell (35 cities, median split)	wrong	right	flagged	flagged	unflagged	r
Grid-like, well-attested	78.6	67.1	32.3	30.2	15.1	26.6
Grid-like, sparse	80.8	66.5	39.4	28.9	16.3	26.1
Organic, well-attested	81.3	67.5	40.7	25.9	12.0	22.7
Organic, sparse	82.6	66.6	47.1	24.1	13.8	21.9

Table D.5: Critique stages by city cell, pooled over the eleven models and three transitions (all values in percent). "Fixed" is conditional on the claim being retained in the next draft.

Attribution. A two-fold Kitagawa–Oaxaca–Blinder decomposition of the difference in r between two groups of cities separates a composition term (the change in $\tilde{p}$, weighted by the average gap $f_1 - f_0$) from execution terms (the changes in f_1 and in f_0, weighted by the average $\tilde{p}$ and $1 - \tilde{p}$). Between the grid-like, well-attested and the organic, sparsely attested cells, r falls by 4.70 points: the composition term contributes −0.29, flagged execution +4.69 and unflagged execution +0.30 (under the whole-phrase rule, −0.19, +4.50 and +0.39). Between Manhattan and Venice and Kigali and Antananarivo, pooling claims, r falls from 33.7% to 12.0%, a gap of 21.70 points: composition contributes −0.27, flagged execution +19.53 and unflagged execution +2.43. Adding Chongqing to the second group, or weighting cities equally, keeps flagged execution at 89–91% of the gap. In both comparisons essentially the whole gap lies in how often a retained error is rewritten correctly, and the critique's flagging runs slightly the other way. Flagged and unflagged fix rates also move together across cities (correlation 0.82), as a single capacity to supply a correct replacement would predict.

The decomposition is descriptive: it locates where repair fails without showing why a replacement is unavailable. Appendix D.4 tests that directly.

Across models. The same asymmetry appears between models. Over the eleven models of Table 1, r/q rises with first-draft accuracy (Spearman 0.60), but through corruption: q falls with accuracy ($\rho = -0.45$), while r is nearly unrelated to it (0.21) and rises with q (0.44). Stronger models break fewer correct claims; they are not much better at fixing wrong ones.

D.4 An intervention: oracle flags and candidate lists

The decomposition above locates the city gap in the rewrite but cannot say why the rewrite fails. We tested the explanation directly. We drew 8 corridors in each of 20 cities (160 corridors) and ran three models (Gemini 3.1 Pro, Gemini 3.5 Flash and Claude Sonnet 4.6), 480 corridor–model blocks in all. Each block's first draft (6,143 claims, 3,431 of them wrong) was frozen and revised once under three conditions: *(A)* the model's own critique, as in the main benchmark; *(B)* oracle flags: every claim marked correct or wrong, with the failed checks named, and no other information; and *(C)* the same flags plus an alphabetical list of the real landmarks along the corridor, without order, positions, distances or sides. We score the second draft as in the main text. *Deleted* is the share of wrong first-draft claims absent from the revision (d^i/E_1); *repaired* (f_{repl}/E_1) is the share of wrong first-draft claims that are either corrected in place on the matched landmark (f/E_1) or replaced by a newly added correct claim, counting within each block b at most one newly added correct claim per deleted wrong claim ($\min(d^i_b, a^c_b)/E_1$; without that cap, condition C would exceed 100%). Across the 3,431 wrong first-draft claims, separating in-place matched repair from capped replacement shows where each arm acts: under **Arm A (own critique)**, 19.0% (652) are repaired in place and 5.4% (186) are replaced by a new correct claim (combined 24.4%, 838/3,431, with 29.6% [1,017] deleted); under **Arm B (oracle flags)**, in-place matched repair rises to 35.4% (1,216) and capped replacement to 15.8% (543) (combined 51.3%, 1,759/3,431, with 59.9% [2,054] deleted); under **Arm C (oracle flags + candidate list)**, in-place matched repair is virtually unchanged from Arm B at 36.2% (1,242), while capped replacement jumps to 58.5% (2,006) (combined 94.7%, 3,248/3,431, with 59.9% [2,054] deleted). Contrasts are paired within block, with standard errors clustered by corridor.

		deleted (%)			repaired (%)		
Cell (median split)	cities	A	B	C	A	B	C
All	20	29.6	59.9	59.9	24.4	51.3	94.7
Grid-like, well-attested	8	22.8	49.2	49.5	29.9	65.9	93.9
Grid-like, sparse	3	25.1	50.6	50.0	34.1	64.9	94.7
Organic, well-attested	3	28.3	60.5	59.5	25.2	48.5	94.8
Organic, sparse	6	43.5	81.1	81.6	10.2	22.9	95.8
Gap, first cell minus last					19.7	43.0	−1.9

Table D.6: The intervention by city cell, using the same degree and attestation medians as Table D.5; rates are pooled over wrong first-draft claims. A: own critique; B: oracle flags; C: oracle flags plus an unordered list of the corridor's real landmarks.

Flags alone widen the gap. Oracle flags double deletion of wrong claims (+29.8 points, standard error 1.6) and all but stop the deletion of correct ones (13.6% → 2.3%). Repair improves (+27.6 points, standard error 1.8), but unevenly: in grid-like, well-attested cities the model has alternatives to offer, while in organic, sparsely attested cities most replacements are themselves wrong (481 of 530 new claims in the five such cities of the original design). The repair gap between the two cells more than doubles, from 19.7 to 43.0 points (Table D.6).

A candidate list closes it. Adding the list leaves deletion unchanged (−0.2 points, standard error 0.3) and raises repair by 42.3 points (standard error 2.5), to 94–96% in every cell. All three models follow the same path (repair 24–25% under A; 43–56% under B; 92–97% under C).

What the list does and does not show. Three features of condition C limit what it demonstrates. The list is drawn from the same inventory of corridor places that the evaluator scores against, and every listed name the models used passed the existence, location and visibility checks; only the side check

remains, which 91% pass under lenient scoring. Models use the list generously: condition C adds 4,169 claims, 3,420 of them from the list, and the second draft grows to 8,200 claims against 6,143 in the first. And the 6.7% second-draft error of condition C counts a skipped left/right call as passing; counting it as wrong gives 26.3%, averaged over blocks (against 67.3% and 53.2% under A and B), and under that stricter rule condition C is lowest in the organic, sparse cell (19.6%, against 28.8% in the grid-like, well-attested one), so the closure of the gap does not depend on how skipped calls are scored. Condition C therefore shows that once real names are supplied the city gap disappears, not that the models could have ranked the right landmark among the wrong ones; condition B, which supplies no names, carries the causal result. The experiment covers three models, two from one family, and a single revision (D_2, not D_4). The original design grouped the 20 cities into its own easy and hard sets; grouping by those labels gives the same pattern (a repair gap of 20.9, 39.1 and -0.5%).

D.5 Per-city results

Table D.7 gives, for each of the 35 cities, the quantities used in Sections 3.2 and 4, pooled over all eleven closed-book models and over the three revision steps. Cities are sorted by draft-4 error rate (D_4). *Topology* gives the city's street-network archetype from Appendix A.1; D_1 and D_4 are the percentage of place claims scored incorrect in the first and fourth drafts; q and r are the per-revision matched-claim corruption and repair rates; $e^* = q/(q + r)$ is the implied equilibrium error rate; r^+ is the deletion-inclusive repair rate when a deleted incorrect claim is counted as repaired. *Deg.* is the mean OSM node degree of the routes in that city; *Docs* is the geometric mean, over the city's candidate places, of the number of indexed web documents mentioning the place; *T/km* is mean turns per kilometer. The last column gives the mean number of place claims per narration in drafts 1 and 4.

Two patterns stand out. First, e^* sits below D_4 in all 35 cities — by 4.46% on average, and by more than 10% in three of them (Sana'a 43.75% against 58.40%, Managua 38.56% against 51.84%, and Tehran 35.61% against 47.46%) — so the fitted floor is an optimistic bound on where a city is actually heading rather than a description of where it has arrived. Second, the column that varies is r: it ranges over a factor of 3.18 (Venice 34.63% to Kigali 10.90%) and correlates with intersection degree at $\rho = +0.49$ and with attestation at $+0.29$, while q ranges over a factor of only 1.84 (Manhattan 10.40% to Chongqing 19.13%) and correlates only weakly with degree ($\rho = -0.26$). Even deletion-inclusive repair r^+ varies by more than 23% (45.0% in Istanbul to 68.9% in Tehran), and the floor e^* runs from 23.9% (Manhattan) to 57.5% (Kigali). Cities are not mainly distinguished by how likely a model is to break something; they are distinguished by whether anything can be fixed.

E Path-Dependence: Robustness and Additional Analyses

This section supports Section 5: the resampling null for per-corridor accuracy swings (Appendix E.1), cross-direction overlap by city, robustness to the corridor gate, the repeated-trial study of one-way landmarks, how position along the walk is measured, the synthetic city's design and full results, and the reversal-disagreement detector for each model.

E.1 Per-corridor accuracy swings versus resampling

At the aggregate level, neither direction is systematically easier: pooled over the eleven closed-book models, 35 cities, and four drafts, forward ($A \to B$) accuracy is 53.78% and reverse ($B \to A$) accuracy is 54.25% (46.13% vs. 46.84% on D_1); no individual model differs by more than 2%, and forward has higher accuracy in 16 of 35 cities (exact two-sided binomial sign test $p = 0.74$; 14 of 35 on D_1, $p = 0.31$).

Large-looking per-corridor differences, however, can be manufactured by small samples. Taking each corridor pair and comparing the accuracy of its two descriptions, the mean absolute difference is 16–32% depending on the model. But a description contains only 5.8–25.0 claims on average, depending on the model, so a difference of that size is close to what pure resampling produces: re-drafting the *same* direction of the same corridor yields apparent swings of 15–29%, and a binomial null with the observed denominators yields 15–24%. The excess of the bidirectional swing over the within-direction floor is only 0.9 to 6.6% (pooled 4.2%). Per-corridor accuracy swings are therefore not a useful measure of directional asymmetry. Pooled over models, the corridor-level error rates of

City	Topology archetype	D_1 err. (%)	D_4 err. (%)	q (%)	r (%)	e^* (%)	r^+ (%)	Deg.	Docs (geo. mean)	T/km	Claims D_1/D_4
Manhattan	Orthogonal Grid	40.74	26.70	10.40	33.12	23.90	52.71	3.760	475,518	28.4	14.33 / 11.75
Venice	Organic / Medieval	37.16	29.30	13.48	34.63	28.01	53.79	2.741	27,286	11.8	12.65 / 11.34
St. Petersburg	Planned / Radial	41.76	32.20	13.25	31.54	29.58	55.54	2.948	4,301	3.6	12.27 / 10.24
Paris	Planned / Radial	42.93	32.28	13.20	28.72	31.49	45.64	3.510	81,846	28.6	15.60 / 13.00
Washington, DC	Planned / Radial	49.69	34.94	13.51	27.80	32.71	50.94	3.261	286,368	25.9	14.75 / 11.56
London	Organic / Medieval	46.63	36.36	14.83	27.54	35.00	46.55	3.271	204,500	16.8	16.24 / 13.36
San Francisco	Orthogonal Grid	51.12	36.63	13.67	26.85	33.74	51.75	3.318	298,994	23.7	11.90 / 8.83
Barcelona	Orthogonal Grid	50.66	37.45	14.02	26.95	34.23	50.25	3.307	18,905	21.0	10.30 / 8.15
Amsterdam	Organic / Medieval	47.98	37.78	15.54	27.91	35.77	48.61	2.898	22,028	13.2	11.05 / 9.03
Boston	Organic / Medieval	48.55	38.28	15.87	27.19	36.85	47.29	3.255	170,019	25.7	17.22 / 14.19
Hong Kong	Mixed / Irregular Grid	49.04	39.10	15.54	26.43	37.04	48.48	3.400	86,613	9.2	13.27 / 10.78
Honolulu	Mixed / Irregular Grid	53.33	40.21	14.16	24.14	36.97	50.18	2.742	18,393	8.9	12.29 / 9.26
Belgrade	Mixed / Irregular Grid	51.98	40.34	14.99	26.18	36.42	53.54	2.898	9,695	3.3	10.74 / 8.53
Kyoto	Orthogonal Grid	54.52	40.60	15.63	26.02	37.52	53.61	3.268	5,760	17.5	10.40 / 7.93
Tokyo	Mixed / Irregular Grid	52.89	40.81	13.89	22.47	38.19	49.11	3.816	14,186	25.7	12.45 / 9.64
Fez	Organic / Medieval	52.20	41.37	15.67	27.02	36.70	52.68	2.746	5,102	16.3	8.37 / 6.70
Valparaíso	Cul-de-sac / Tree	54.95	41.81	16.73	25.22	39.88	53.43	2.581	8,226	5.1	10.49 / 8.15
Auckland	Mixed / Irregular Grid	53.63	41.88	16.29	24.11	40.32	50.28	2.586	28,007	5.1	11.45 / 8.35
Reykjavík	Mixed / Irregular Grid	53.94	41.96	15.45	24.44	38.73	54.43	2.704	10,688	7.7	10.42 / 7.45
Anchorage	Orthogonal Grid	58.79	43.04	13.88	23.85	36.78	60.66	2.820	8,403	7.4	8.35 / 5.86
Sydney	Mixed / Irregular Grid	55.37	43.05	16.27	24.09	40.32	49.06	3.180	48,650	8.5	11.72 / 8.97
Istanbul	Organic / Medieval	53.56	43.46	17.47	24.53	41.60	44.97	2.752	12,047	8.7	12.88 / 10.48
Mumbai	Mixed / Irregular Grid	53.63	43.60	16.63	23.53	41.41	46.42	2.583	123,215	4.5	14.58 / 11.40
Tehran	Mixed / Irregular Grid	60.79	47.46	14.90	26.95	35.61	68.86	2.564	7,035	5.1	6.83 / 4.91
San Juan	Mixed / Irregular Grid	62.28	48.66	13.71	17.83	43.47	54.98	2.681	50,944	3.0	8.50 / 5.78
Buenos Aires	Orthogonal Grid	61.65	49.55	17.65	20.26	46.55	46.84	2.912	42,907	26.6	9.88 / 7.44
Havana	Orthogonal Grid	58.65	49.99	15.54	18.10	46.19	47.40	3.044	54,497	3.4	9.56 / 7.43
Chandigarh	Planned / Radial	64.07	50.05	12.32	16.68	42.49	56.41	2.801	138,792	5.7	8.61 / 5.87
Cairo	Organic / Medieval	63.24	50.58	14.86	19.82	42.84	60.64	2.779	9,267	5.5	8.43 / 5.99
Managua	Cul-de-sac / Tree	63.05	51.84	12.13	19.33	38.56	66.88	2.843	6,729	4.2	6.62 / 4.77
Bangalore	Mixed / Irregular Grid	65.18	53.37	17.55	19.76	47.04	54.70	2.771	91,617	4.8	10.76 / 7.56
Sana'a	Organic / Medieval	68.01	58.40	13.55	17.42	43.75	62.98	2.707	2,046	4.9	6.16 / 4.54
Antananarivo	Cul-de-sac / Tree	68.57	59.09	13.51	13.13	50.71	55.41	2.288	5,639	7.2	7.55 / 5.54
Chongqing	Cul-de-sac / Tree	69.60	59.47	19.13	17.48	52.25	57.21	2.501	2,933	5.1	7.88 / 5.83
Kigali	Cul-de-sac / Tree	73.09	64.42	14.73	10.90	57.48	55.11	2.478	6,056	3.9	7.94 / 5.55

Table D.7: All 35 cities in the CityWalk benchmark, pooled over all eleven closed-book models and three revision transitions, sorted by draft-4 claim error rate (D_4). Every column (street-network topology archetype, D_1, D_4, matched-claim corruption q and repair r, equilibrium floor $e^* = q/(q+r)$, deletion-inclusive repair r^+, node degree, web search document attestation geometric mean, turns per kilometer, and claims per narration in D_1/D_4) is computed directly from the released 2,523,699-claim dataset. Corridor-cluster bootstrap 95% confidence interval half-widths ($B = 1{,}000$ resamples over the city's paired corridors) are ± 0.8–1.3% on D_1 and D_4, ± 0.6–1.1% on q, and ± 0.8–1.5% on r. D_1 and D_4 are claim-weighted error rates pooled over the eleven models, the series used for every city-level number in this paper; weighting the models equally instead gives, for example, 28.3% for Manhattan (against 26.7%) and 64.0% for Kigali (against 64.4%).

Model	Cross-direction Jaccard overlap (2-pass entity + name)				Side-of-street inversion rate		
	≤ 40 m	≤ 25 m	≤ 10 m	≤ 5 m	Raw (≤ 40 m)	Purged (≤ 40 m)	Purged (≤ 25 m)
Gemini 3.1 Pro	39.31%	38.98% (−0.33)	37.29% (−2.02)	35.75% (−3.56)	70.22%	70.56%	70.01%
Gemini 3.5 Flash	38.60%	38.22% (−0.38)	37.27% (−1.33)	36.63% (−1.97)	66.25%	66.60%	66.36%
Claude Opus 5	37.90%	37.61% (−0.29)	36.09% (−1.81)	35.13% (−2.77)	70.85%	71.10%	70.61%
Claude Sonnet 4.6	40.36%	40.75% (+0.39)	42.17% (+1.81)	42.04% (+1.68)	47.71%	47.62%	47.64%
GPT-5.6 Terra	27.70%	27.79% (+0.09)	27.26% (−0.44)	27.60% (−0.10)	67.10%	67.25%	67.02%
Claude Sonnet 4.5	43.19%	43.37% (+0.18)	44.10% (+0.91)	44.66% (+1.47)	46.49%	46.77%	47.92%
Gemini 3.0 Flash	37.33%	36.92% (−0.41)	35.22% (−2.11)	34.38% (−2.95)	60.91%	61.16%	61.11%
Qwen 3.8	16.36%	16.74% (+0.38)	16.69% (+0.33)	16.44% (+0.08)	52.10%	52.20%	52.12%
GLM-5.3	14.07%	14.22% (+0.15)	14.64% (+0.57)	14.56% (+0.49)	49.06%	49.19%	49.26%
Gemini 2.5 Flash	33.24%	33.32% (+0.08)	32.51% (−0.73)	32.44% (−0.80)	52.56%	52.68%	53.09%
GPT o4-mini	15.12%	14.93% (−0.19)	13.94% (−1.18)	12.97% (−2.15)	49.37%	49.60%	49.58%

Table E.1: First-draft cross-direction Jaccard overlap (two-pass entity-ID plus normalized name alignment) across symmetric Hausdorff distance gates (3,016, 2,575, 1,515, and 1,052 corridor pairs) and shared-landmark side-of-street inversion rates before and after purging repeated within-route mentions and straddling entities across all eleven closed-book models.

the two directions are nonetheless strongly correlated for location ($r = 0.84$ across the 3,016 pairs) and for fourth-draft error (0.73), but only weakly for side of street among correctly located claims (0.31): which landmarks get misplaced is largely a property of the corridor, while which sides come out wrong depends on the direction walked.

E.2 Cross-direction overlap by city

Per corridor pair, the number of landmarks named in one direction and never in the other ranges from 4.6 for the tersest model to 19.3 for the most verbose. Aggregated by city, cross-direction Jaccard overlap spans 0.256 to 0.404 across the 35 cities — with Paris (0.404), Managua (0.398), Fez (0.397), Venice (0.387) and Sana'a (0.385) near the top, and Chongqing (0.256), Tokyo (0.299) and Valparaíso (0.306) at the bottom — while remaining nearly orthogonal to the two macroscopic city axes ($\rho = 0.07$ with mean intersection degree and $\rho = 0.00$ with landmark attestation). Even in the highest-overlap cities, roughly three-fifths of the landmarks in the union of the two directions (about two-fifths of any single description; Section 5.1) appear in only one of them.

E.3 Robustness to the corridor gate and to side-flip purges

Because forward and reverse walks are routed independently by the map provider rather than reversed textually, two fair concerns arise in Section 5.1: first, that part of the low cross-direction overlap reflects minor local differences between the two polylines up to the 40 m admission gate (Appendix A.2); second, that part of the side-flip failure reflects landmarks mentioned multiple times along a walk or straddling both sides of a corridor. Table E.1 tests both simultaneously across all eleven closed-book models. Tightening the symmetry gate from ≤ 40 m (3,016 pairs) to ≤ 25 m (2,575 pairs), ≤ 10 m (1,515 pairs), and ≤ 5 m (1,052 pairs) shifts Jaccard overlap by at most 0.41% at 25 m and 3.56% at 5 m. Likewise, purging repeated within-route mentions and straddling entities moves the side-of-street inversion rate by at most 0.35% on any model, and tightening further to the ≤ 25 m subset shifts the purged side-flip rate by at most 1.15% (Claude Sonnet 4.5; under 0.6% on the other ten models). Neither routing divergence nor repeated-visit ambiguity accounts for the observed path-dependence.

E.4 The repeated-trial study of one-way landmarks

Design. To separate landmarks a model mentions erratically from landmarks it mentions systematically in one direction only (Section 5.2), Gemini 3.5 Flash walked each corridor eight times per direction in six cities (Manhattan, London, Tokyo, Bangalore, Havana, and Antananarivo; 248 paired corridors, 496 directed routes, 3,968 narrations). Across all $\binom{8}{2} = 28$ same-direction trial pairs per route versus all $8 \times 8 = 64$ cross-direction trial pairs, same-direction Jaccard set overlap is $J_{\text{within}} = 40.0\%$ against $J_{\text{cross}} = 34.1\%$ across reversed traversals ($\Delta J = -5.9\%$, noise-corrected cross-direction retention $J_{\text{cross}}/\sqrt{J_{\text{within}}(F)\, J_{\text{within}}(R)} = 85.6\%$; under bipartite canonical-entity plus normalized-name matching on the 150-corridor primary subset, $J_{\text{within}} = 40.3\%$ vs. $J_{\text{cross}} = 34.7\%$, $\Delta J = -5.6\%$, 95% CI $[-6.3\%, -4.8\%]$; on Claude Sonnet 4.6, $J_{\text{within}} = 43.6\%$ vs. $J_{\text{cross}} = 37.1\%$,

$\Delta J = -6.5\%$, retention 84.3%). Thus roughly three-fifths of single-pass non-overlap reflects within-direction stochastic subset selection across trials, while reversing the walk imposes an additional 5.6–6.5% systematic cross-direction drop. Tracking position $u \in [0, 1]$ requires the forward and reverse polylines to trace the same physical street, so, unlike Section 5.1, which uses all $3{,}016$ corridor pairs, the study keeps the 129 corridors whose two polylines are within 25.0 m (symmetric Hausdorff) and retains the 566 route–landmark pairs named in at least seven of eight trials in at least one direction. Of these, 488 (86.2%) are named in seven or eight trials each way, and 78 (13.8%; 31 forward-only and 47 reverse-only) in seven or eight trials one way and at most one the other. Within a direction the behavior is nearly deterministic, and it switches when the direction does.

Existence. All 78 one-way landmarks exist and lie on the corridor, with a median of $292{,}426$ attesting web documents and 692 Google Maps reviews, against $286{,}486$ documents and 575 reviews for the 488 bidirectional landmarks.

The one-way landmarks differ from the bidirectional ones in three ways.

Position along the walk: the arrival beacon. Bidirectional landmarks sit in the middle of the walk: their mean normalized position oriented to the direction of travel is 0.46 (measured in fixed physical reference coordinates from the forward origin, 0.49). Of the 488 bidirectional pairs, 32.0% ($156/488$) are route origin or destination anchors ($u \in \{0, 1\}$) and 68.0% ($332/488$) are non-endpoint interior landmarks; all 332 (100%) of those non-endpoint bidirectional pairs fall in the middle half of the route ($u \in [0.25, 0.75]$), yielding $369/400$ (92.3%) in the middle half across all non-endpoint $\geq 7/8$ high-consistency pairs (332 bidirectional $+37$ of 68 non-endpoint one-way pairs; $369/566 = 65.2\%$ including endpoint anchors), compared to 48.2% in the middle half across unfiltered single-pass D_1 claims prior to multi-trial consistency filtering and endpoint-projection exclusions. One-way landmarks sit near the end of whichever traversal mentions them: mean oriented position 0.63, with 41.0% ($32/78$) in the final quarter ($u \in [0.75, 1.0]$) against 47.4% ($37/78$) in the middle half and only 11.5% ($9/78$) in the departure quarter ($u \in [0, 0.25)$), a $3.56\times$ arrival-versus-departure skew. A landmark that is near one end of a corridor is near the *destination* in one direction and near the *origin* in the other, and the model treats those two situations very differently.

Besides the Intrepid and Shinagawa cases of Section 5.2, Radio City Music Hall is a further line-of-sight case: it dominates the view walking south on Sixth Avenue and is behind you walking north.

Invariance across landmark attestation tertiles. For Gemini 3.5 Flash, directional recall is nearly invariant to how much has been written about a landmark: landmarks named symmetrically have a geometric mean of 278k attesting documents (median 286k), against 171k for forward-only and 204k for reverse-only landmarks (overall one-way median 292k). Across within-city fame tertiles, the fraction of high-consistency landmarks that are one-way is 16.4% ($11/67$) in the bottom tertile (P0–P33), 13.6% ($27/198$) in the middle tertile (P33–P67), and 13.3% ($40/301$) in the top tertile (P67–P100) — a difference of only two pairs in a stratum of 67.

Interior vs. endpoint landmarks. Of Gemini 3.5 Flash's 78 one-way landmarks, 87.2% ($68/78$) are non-endpoint interior landmarks along the corridor and only 12.8% ($10/78$) are route origin or destination anchors (compared to 32.0% endpoint anchors among the 488 bidirectional landmarks). The one-way pattern is therefore primarily driven by interior turn markers and arrival-horizon landmarks rather than endpoint naming conventions.

E.5 Position along the walk

Each claimed landmark is projected onto the walked polyline, giving a normalized position $u \in [0, 1]$ from origin to destination, oriented to the direction of travel. A landmark behind the origin or beyond the destination projects onto the first or last vertex and piles up at exactly $u = 0$ or $u = 1$; 9.7% and 13.6% of the $2{,}523{,}699$ claims do so, and all numbers below exclude them.

Section 5.3 and Figure 4b give the resulting U-shaped profile and how it varies across cities. Restricting to landmarks that are genuinely adjacent to the path — within 50 meters of the polyline — sharpens it: the first and last tenths of the route carry 17.1% and 26.1% of mentions against 6.7–7.7% for each interior tenth, so the ends of a walk receive between two and four times the descriptive

attention of the middle. Accuracy is mildly U-shaped as well (52.8% in the first tenth, 51.4–55.0% in the interior tenths, 57.9% in the last; on the within-50-meter subset, 73.8%, $\approx 68\%$, 75.6%). Paris, London, and Manhattan spread their mentions most evenly along the route.

Position on the route versus position in the narration. For $87{,}047$ fourth-draft claims (27% of correctly located claims with a stated side) the landmark matches a place in the route's candidate inventory, so its position u along the route is known from the map; there it is nearly uncorrelated with the claim's position τ in the narration ($r = -0.0003$), and the two can be separated. Location error follows the narration: over all $575{,}333$ fourth-draft claims it is 18.3% in the first fifth of a description, 29.3% in the middle fifth and 17.6% in the last. Side error among the matched claims follows the route more than the narration: in a joint regression the mid-route term in u is 2.7 times the mid-narration term in τ (standardized coefficients 0.042 and 0.015). Claims without a matched candidate cannot test this, because their position can only be approximated from narration order.

E.6 The synthetic city: design, conditions, and results

Design. A fictitious avenue is generated with 99 numbered cross-streets, and a table of landmarks with their cross-streets and sides is supplied in the prompt; the model narrates a walk down the avenue in one direction or the other. Every fact needed is in context, and no fact about the place exists anywhere else. Three models (Claude Sonnet 4.6, Gemini 3.8 Flash and Gemini 3.5 Flash) were run over 300 such knowledge bases in both directions with repeated sampling: about $12{,}000$ narrations in the core condition and a further $10{,}800$ across the other conditions of Table E.2.

Measures. Reversal divergence is $J_{\text{cross}} - J_{\text{within}}$: the Jaccard overlap of the landmarks mentioned in two narrations walked in opposite directions, minus the overlap between two narrations resampled in the same direction. Whether each supplied landmark is mentioned is regressed on its position along the walk, u, and its position in the prompt table, v, with linear terms β_u, β_v and quadratic terms γ_u, γ_v. The wrong-street rate is the share of mentioned landmarks attributed to the wrong cross-street.

Results. Table E.2 gives every condition for each model, and Figure E.1 the inclusion curves with a shuffled prompt table. Beyond Section 5.4: the front-loading model (Claude Sonnet 4.6, $\beta_u = -0.51$) has an inclusion probability of 0.89 at the start of the walk that falls to 0.37 by the seventh decile; its side errors carry a 10.6% penalty walking south, the direction that requires rotating away from a north-up frame; and forcing the cross-street to be named cuts its wrong-street rate by 88% while its claim rate stays at essentially 100%.

Why a symmetric U-shape alone does not lower cross-direction overlap, and what does. If landmarks at normalized positions $u \in [0, 1]$ were included independently with a purely symmetric U-shaped probability $p(u) = p(1-u)$ ($\beta_u = 0$), expected within-direction overlap $\int_0^1 p(u)^2\,du \big/ \int_0^1 (2p(u) - p(u)^2)\,du$ and expected cross-direction overlap $\int_0^1 p(u)p(1-u)\,du \big/ \int_0^1 (p(u)+p(1-u)-p(u)p(1-u))\,du$ would be *identical* ($\Delta J = J_{\text{cross}} - J_{\text{within}} = 0.00\%$): a symmetric U-shape reduces J_{within} below 1, but equally in both directions. Even adding the empirical linear front-loading slope $\beta_u = -0.51$ (Claude Sonnet 4.6) under independent Bernoulli draws yields only $\Delta J = -0.21\%$. The observed -6.3% to -19.6% cross-direction Jaccard drop in the shuffled synthetic city therefore reflects *asymmetric departure-versus-arrival shaping coupled with sequential autoregressive path dependence* (which landmarks are selected early in a walk constrains which subsequent landmarks fit the remaining word budget and narrative flow), while in real cities it is further compounded by line-of-sight and arrival-beacon asymmetry (41.0% in the final quarter vs. 11.5% in the departure quarter; Section 5.2).

E.7 The reversal-disagreement detector

Section 5.5 introduces the test: if a model names a landmark in both directions of a corridor and states a side both times, the side should flip, and when it does not, at least one of the two statements must be wrong unless the landmark straddles or turns with the route. On first drafts, both statements are scored correct in 43.63–78.17% of the shared pairs whose side flips and in only 2.60–6.70% of those whose side is preserved (per model in Table E.3). The second rate is the complement of the detector precision reported there: a preserved pair is either a genuine error, which the precision column counts, or a pair the evaluator credits twice, and the two must sum to one.

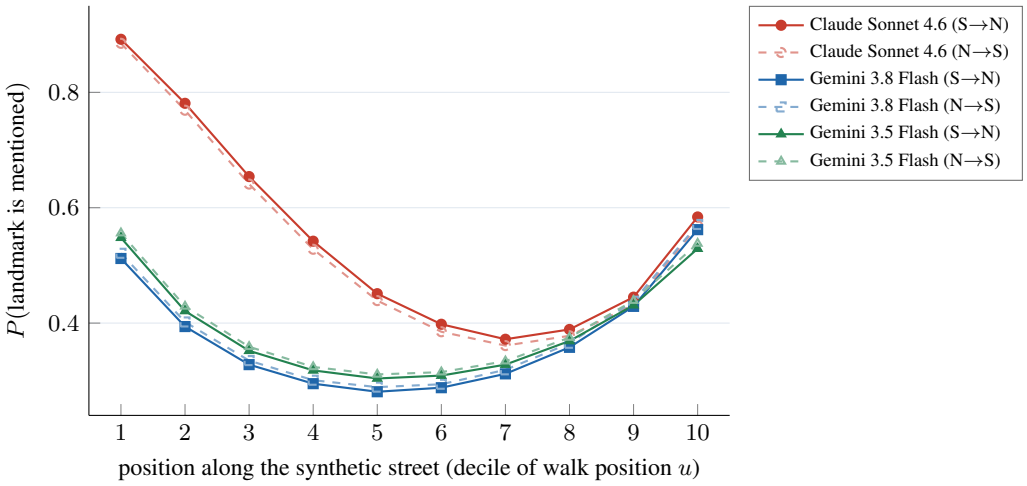

Figure E.1: In a synthetic city where every fact is given in the prompt, which landmarks get mentioned is a U-shaped function of where they are along the walk. Probability that a landmark supplied in the prompt is mentioned in the narration, by decile of its position along the walk, in the condition where the prompt table is randomly shuffled so that list order carries no information ($n = 1{,}200$ narrations per model). Reversing the direction of travel maps street position u to $1 - u$; while a purely symmetric curve $p(u) = p(1 - u)$ yields zero excess cross-direction Jaccard gap under independent draws, combining asymmetric front-loading ($\beta_u = -0.51$, top curves) with sequential autoregressive selection produces a 6.3–19.6% excess cross-direction divergence.

Table E.2: Synthetic-city control conditions. All facts are supplied in the prompt, so none of these numbers involve world knowledge. *Reversal divergence* is $J_{\text{cross}} - J_{\text{within}}$: how much more two descriptions of a street differ when they are walked in opposite directions than when they are merely resampled in the same direction. *Wrong-street rate* is how often a landmark is attributed to the wrong cross street. The last two rows of the table show that the same quantity responds strongly to manipulations that have nothing to do with knowledge: requiring the model to name the cross street inline, and giving it more room to write.

Condition	Quantity	Claude Sonnet 4.6	Gemini 3.8 Flash	Gemini 3.5 Flash
Fabrication (all conditions)	invented landmarks / mentions		0/99,251	
Ordered prompt table	reversal divergence	−14.7%	−6.1%	−2.4%
Shuffled prompt table	reversal divergence	−19.6%	−7.1%	−6.3%
Shuffled prompt table	J_{within} / J_{cross}	0.625/0.429	0.683/0.613	0.559/0.496
Walk-position curvature	γ_u (endpoint elevation)	+1.35 (+16.0%)	+1.40 (+11.4%)	+1.40 (+10.2%)
List-position curvature	γ_v	−0.57	+0.22	+0.22
List-position primacy	β_v	−0.11	−0.21	−0.17
Forced side labels, S→N	side-error rate	19.5%	0.00% (0/7,709)	
Forced side labels, N→S	side-error rate	30.1%	0.00%	
Voluntary street naming	wrong-street rate	1.72%	0.02%	0.03%
Forced street naming (anchor)	wrong-street rate	0.21% (−88%)	—	0.00%
100-word budget	wrong-street rate	4.47%	2.01%	0.16%
400-word budget	wrong-street rate	0.72%	0.09%	0.05%

That residual 2.60%–6.70% rate is worth stating carefully, because it is easy to over-read. It is the rate at which the evaluator credits *both* members of a pair that cannot both be right, and so it is a conditional false-acceptance diagnostic on a specific subset: shared landmark pairs whose side label failed to invert. It is not an upper bound on evaluator error in general. Two families of mistake are invisible to it — false rejections, where the evaluator marks a correct claim wrong, and any error on the large majority of claims that never enter a shared pair at all — and a landmark that genuinely straddles a corner can produce a legitimate non-flipping pair. Read as what it is, it says that when the evaluator is handed two mutually inconsistent statements, it resolves the inconsistency correctly at least 93.30% of the time across every model.

A preserved side label is therefore a reliable, map-free marker that something has gone wrong (illustrated on a real Manhattan pair in Figure B.5 and Table B.3, where three of the seven shared side-specified landmarks keep the same side label and each contains a genuine side error), and it

Model	Shared pairs (N)	Both correct | flipped	Error precision | preserved	Error recall
Gemini 3.1 Pro	11,315	76.28%	94.20%	62.85%
Gemini 3.5 Flash	11,821	76.06%	94.46%	66.91%
Claude Opus 5	27,034	78.17%	93.30%	63.97%
Claude Sonnet 4.6	5,149	52.19%	96.80%	68.85%
GPT-5.6 Terra	14,632	73.65%	94.31%	63.70%
Claude Sonnet 4.5	3,487	50.90%	97.32%	69.58%
Gemini 3.0 Flash	10,973	69.11%	95.16%	66.62%
Qwen 3.8	4,776	54.20%	96.24%	65.88%
GLM-5.3	6,637	44.81%	97.40%	64.79%
Gemini 2.5 Flash	3,452	62.69%	97.20%	70.29%
GPT o4-mini	4,452	43.63%	97.15%	64.34%

Table E.3: Unsupervised side-of-street error detection via forward/reverse side-label agreement across all eleven closed-book models on draft 1, restricted to shared landmark pairs with explicit side labels in both directions ($\mathcal{C}_{\text{spec}}$).

catches errors that introspection misses: in the case study of Figure C.2 (mapped in Figure B.4), the model's own critique did not flag the erroneous claim in any of three attempts. The recipe is general. Wherever a description can be re-elicited under a transformation that should act predictably on its content — reversal, reordering, relabeling, a change of reference frame — the failure to transform correctly is evidence about which of its commitments are fundamental and which are confabulated.

F Scheduling Revisions Across Models

The (p, q, r) model turns the engineering intuition "draft with a cheap model and have an expensive one clean it up" into a scheduling problem over models with different prices, drafting qualities and revision behaviors. The results below are analytical consequences of the model; Section 7 uses Theorem 5 and the worked example, and the last subsection names the measurement they would need.

F.1 Setup

Let $\mathcal{M} = \{1, \dots, K\}$ be a set of models. Model k can:

- *draft* from scratch, at cost γ_k, producing a description with expected claim error rate p_k; or
- *revise* a description whose last author was model j, at cost ρ_k, transforming its expected error rate by the affine map
$$f_{jk}(e) \;=\; q_{jk} \;+\; (1 - q_{jk} - r_{jk})\, e,$$
where q_{jk} and r_{jk} are the corruption and repair rates when k revises j's text.

A *schedule* is a draft by some model followed by a finite sequence of revisions. Its cost is the sum of the individual costs and its value is the resulting expected error rate. Write $\kappa = \max_{j,k} |1 - q_{jk} - r_{jk}|$ and assume $\kappa < 1$, which holds comfortably in our data ($\kappa = 0.71$ across the diagonal maps).

We additionally assume throughout this section the *monotonicity condition*

$$q_{jk} + r_{jk} \;\le\; 1 \qquad \text{for all } j, k, \tag{F.1}$$

equivalently that each f_{jk} has nonnegative slope and is therefore nondecreasing. This does not follow from $\kappa < 1$: a contraction may have negative slope, in which case a revision can overshoot below the fixed point. With one model, $p = 0.9$ and $q = r = 0.9$, the map $f(e) = 0.9 - 0.8e$ is a contraction, yet a single revision yields $f(0.9) = 0.18$, below both p and the fixed point $e^* = 0.5$, so Theorem 5 fails without (F.1). The condition is an empirical property of the fitted maps rather than a modeling convenience: every measured (q_{jk}, r_{jk}) pair in our data satisfies it, since repair and corruption rates that sum above one would require a revision pass to invert the status of more claims than it touches.

Define E_k to be the infimum of the error rate over all schedules whose last operation is by model k. These satisfy the Bellman equation

$$E_k \;=\; \min\Big\{\, p_k,\;\; \min_{j \in \mathcal{M}} f_{jk}(E_j) \,\Big\}. \tag{F.2}$$

F.2 The unconstrained optimum

Theorem 4. *Let $T : [0,1]^K \to [0,1]^K$ denote the right-hand side of (F.2). If $\kappa < 1$ then T is a κ-contraction in the sup norm, so (F.2) has a unique solution $E^\star$, and the value iteration $E_k^{(0)} = p_k$, $E^{(m+1)} = T(E^{(m)})$ converges to it with $\|E^{(m)} - E^\star\|_\infty \le \kappa^m$. If in addition the monotonicity condition (F.1) holds, then $E_k^{(m)}$ is exactly the best error rate achievable by a schedule of at most m revisions ending at k, so the sequence is nonincreasing and every value is realized by an explicit schedule.*

Proof. For any E, E' and any k, $|T(E)_k - T(E')_k| \le \max_j |f_{jk}(E_j) - f_{jk}(E'_j)|$, because $x \mapsto \min(p_k, x)$ and the pointwise minimum over j are both nonexpansive. Each f_{jk} is affine with slope $1 - q_{jk} - r_{jk}$, so $|f_{jk}(E_j) - f_{jk}(E'_j)| \le \kappa |E_j - E'_j|$. Hence $\|T(E) - T(E')\|_\infty \le \kappa \|E - E'\|_\infty$, and Banach's theorem applies. The identification of $E^{(m)}$ with m-revision schedules is induction on m: $E^{(0)}$ is drafting only, and one more application of T considers appending one revision by k to the best $\le m$-revision schedule ending at each j, which is optimal because each f_{jk} is nondecreasing (given $q_{jk} + r_{jk} \le 1$) and therefore prefers a better predecessor. □

With $K = 11$ and $\kappa = 0.71$, twenty iterations of a K^2 inner loop settle the answer to three decimal places. Pruning is available if K is large: every schedule ending at k has value at least $\min\{p_k, \min_j q_{jk}\}$, since $f_{jk}(e) \ge q_{jk}$, so states whose lower bound exceeds the current best can be dropped.

F.3 When orchestration helps

The interesting question is whether mixing models can beat every single model. If revision quality is a property of the *reviser* alone — $q_{jk} = q_k$ and $r_{jk} = r_k$, so that $f_{jk} = f_k$ with fixed point $e_k^* = q_k/(q_k + r_k)$ — the answer is no.

Theorem 5. *Assume the monotonicity condition (F.1). If $f_{jk} = f_k$ for all j, then the optimal achievable error rate is*

$$\min_k E_k \;=\; \min_{k \in \mathcal{M}} \min\{\, p_k,\; e_k^* \,\}.$$

That is: the best a pipeline can do is either the best available single draft, or the best available single-model equilibrium. No sequence of models beats both.

Proof. Write $\mu = \min_k E_k$, attained at k^*, and let $P = \min_k p_k$, $S = \min_k e_k^*$. From (F.2), $\mu = \min\{p_{k^*}, f_{k^*}(\mu)\}$, so either $\mu = p_{k^*} \ge P$ or $\mu = f_{k^*}(\mu)$, i.e. $\mu = e_{k^*}^* \ge S$; either way $\mu \ge \min\{P, S\}$. For the reverse inequality, suppose $S \le P$ and let k^* attain S. Then $f_{k^*}(S) = S \le P \le p_{k^*}$, so $E_{k^*} \le S$. If instead $P < S$, let k^* attain P; since $P < S \le e_{k^*}^*$ and f_{k^*} is a contraction toward $e_{k^*}^*$ with nonnegative slope, $f_{k^*}(P) > P$, so $E_{k^*} \le p_{k^*} = P$. In both cases $\mu \le \min\{P, S\}$. □

The theorem is a conditional impossibility result with a constructive reading. When a single model k is applied repeatedly as the terminal reviser ($f_k^{(m)}$), the trajectory converges to e_k^* regardless of the prefix; when a schedule alternates among distinct revisers $(k_1, k_2, \dots)$, the iterates oscillate in the convex hull of their fixed points and, by Theorem 5, remain bounded below by $\min\{P, S\}$. Orchestration can therefore beat $\min\{P, S\}$ only if revision quality genuinely depends on who wrote the text (so that f_{jk} is non-separable — e.g., if a cross-family reviser k achieves a lower corruption-to-repair ratio on author j's errors than on its own), or in the *cost-constrained* regime, where the goal is not the asymptotic optimum but a target error rate reached cheaply.

Substituting the diagonal parameters of the eleven closed-book models in Table 1 into Theorem 5 gives $P = 40.96\%$ (Gemini 3.5 Flash's first draft) and $S = 27.83\%$ (Gemini 3.1 Pro's matched-claim equilibrium), so under the separable reviser-only hypothesis the best attainable error rate is 27.83%. Because our benchmark measures only self-revision ($j = k$), whether off-diagonal cross-family revision ($j \ne k$) violates separability sufficiently to breach 27.83% is an empirical question left to Section F.6.

F.4 Minimum cost to a target

In the cost-constrained regime, mixing does help, and the optimization is a clean dynamic program.

Theorem 6. *Assume the monotonicity condition (F.1). Suppose all costs γ_k, ρ_k are positive integers and let $C_{\max}$ be a budget. For $0 \le C \le C_{\max}$ let $W(C,k)$ be the minimum expected error rate achievable by a schedule of total cost at most C ending with model k. Then*

$$W(C,k) \;=\; \min\Big\{\; p_k \;\; \text{if } \gamma_k \le C,\;\; \min_j f_{jk}\big(W(C-\rho_k,\, j)\big) \;\; \text{if } \rho_k \le C \;\Big\},$$

and the table W can be filled in time $O(K^2 C_{\max})$. The minimum cost to reach a target error rate ε is $\min\{C : \min_k W(C,k) \le \varepsilon\}$, and an optimal schedule is recovered by back-pointers.

Proof. Correctness of the recursion is the usual exchange argument, using that each f_{jk} is nondecreasing: replacing the prefix of an optimal schedule by the cost-feasible prefix with the smallest error rate cannot increase the final error rate. The running time is immediate. □

The algorithm is pseudopolynomial in the costs. With real-valued costs one can round to a grid, but we make no continuity claim: $W(C,k)$ is a step function of the budget, and its value can jump at the budget where a useful operation first becomes affordable, so rounding changes the reachable set of schedules rather than perturbing the value continuously.

F.5 A worked example

Take the two strongest closed-book models in Table 1 and assume, for illustration, that each model's revision behavior is its measured diagonal behavior regardless of who wrote the draft. Gemini 3.1 Pro drafts at $p = 0.4135$ and revises with $q = 0.1190$, $r = 0.3087$ (slope $1 - q - r = 0.5723$); Gemini 3.5 Flash drafts at $p = 0.4096$. Consider a target error rate of $\varepsilon = 35.4\%$.

- *Pro alone*: $41.35\% \to 35.56\% \to 32.25\%$. Two revision rounds are required (five model calls: three generations and two critiques).
- *Flash drafts, Pro revises*: $40.96\% \to 35.34\%$. One revision round suffices (three model calls: two generations and one critique).

The mixed schedule reaches the target for 40% fewer model calls (3 vs. 5), and the reason is visible in the parameters: Flash's advantage is in p and inference cost, Pro's is in (q, r), and the schedule that uses each model for the stage it excels at dominates either model alone at that budget. Both schedules share the same asymptote, 27.83%, as Theorem 5 requires.

F.6 What the benchmark does not measure

The benchmark only measures the diagonal maps f_{kk}: every revision in our data is a model revising its own text. The off-diagonal entries f_{jk} are exactly the quantities that would determine whether orchestration has any asymptotic value, and we regard measuring them — in particular, testing whether a model repairs another model's errors at a higher rate r_{jk} than its own — as the most useful single follow-up experiment suggested by this framework. The self-correction literature gives some reason to expect it does [18, 24]: the difficulty a model has in finding its own errors is plausibly correlated with whatever produced them, and models correct errors presented as coming from an external source far more often than identical errors in their own output [63, 7].

G Warm-Up Experiments: Prompts, Scoring, and Full Results

This section gives the prompts, scoring rules and full results for the two warm-ups of Section 1. All trials ran against Gemini 2.5 Pro at default sampling settings, each an independent API call carrying no conversation history. Errors were adjudicated by a second model instance given the generated text and asked to extract the implied counts or amounts and check them; a hand-audit of 100 randomly sampled adjudications found no disagreements on the counting task, where the ground truth is a matter of arithmetic on a list that the story itself supplies.

G.1 Counting the people on the boat

The base prompt:

> Write a 150-word story in which a group of tourists on a boat tour are waiting in line on the dock to board the boat in small friend groups. First write a brief scene that introduces all the tourists; make sure this scene describes all the tourists who are going to board the boat, leaving no one out. Then the two tour-leaders tell the tourists they can all board the boat, which they do. Finally, the boat tour is going to involve snorkeling, and the two tour-leaders discuss whether they have enough snorkels for all the tourists on the boat. (The two tour leaders do not need snorkels; only the tourists do.)

A generation is scored as an error if the number of tourists stated or implied at the snorkel-counting step differs from the number of tourists introduced in the opening scene. This is a test of *internal* consistency: the model is free to introduce any number of tourists it likes, and is only required to agree with itself a paragraph later. In one generation the opening scene introduced the Millers (a couple), the Rodriguez family (two parents, two children), three college friends, and one solo backpacker, ten tourists in all, and the closing dialogue read:

> "Eva, quick count. We have nine tourists. Did we pack enough snorkels for everyone?" Eva scanned the deck, counting under her breath. "The Millers, the Rodriguezes, Liam, Chloe, Ben, and Sarah… yes, nine. Mike, we only have *eight* snorkels onboard!"

Table G.1 gives the error rate by story length and with one sentence appended to the prompt ("Since the narrative will involve numbers, please make sure that all the math that comes up in the narrative is correct"). Longer stories are *more* accurate than shorter ones, which cuts against a simple "losing track over a long context" account.

Variant	Manipulation	Trials	Error rate
Boat / snorkels	150 words	479	11.1%
Boat / snorkels	300 words	2000	5.8%
Boat / snorkels	500 words	246	5.3%
Boat / snorkels	150 words + "make sure the math is correct"	673	0.4%

Table G.1: Counting errors in the boat-tour task (Gemini 2.5 Pro).

The genre experiment uses a 200-word variant in which two safety inspectors check whether there are enough life jackets for the passengers. The only thing that changes across conditions is the noun phrase naming the artifact to be written (Table G.2). The poem, the most formally constrained register, behaves like the story, so the ordering tracks how factual the register is rather than formal difficulty. Table G.2 also separates the two directions of error. Understatement (a claimed total smaller than the number of named individuals) is unambiguously an arithmetic error; overstatement may also reflect the model deciding that unnamed people were present. Overstatement dominates understatement in every register — by 5.4:1 in stories, 3.0:1 in poems, and 2.7:1 in blog posts — while total errors fall by more than an order of magnitude in the factual registers.

Requested artifact	Trials	Understate	Overstate	Total error
A story	600	2.4%	12.9%	15.3%
A poem	600	3.1%	9.4%	12.5%
A blog post	600	2.6%	6.9%	9.5%
A news article	600	0.4%	2.1%	2.5%
A worked example for a math textbook	1,000	0.0%	0.1%	0.1%

Table G.2: Error rates by requested genre, holding the scenario fixed.

G.2 The family budget

The prompt asks for a 300-word article describing a reasonable monthly budget for a family of four, with the following specifics:

> Write a 300-word article in which you use conventional financial wisdom to describe an example of a reasonable budget that a married couple with two children could use for a $5000-per-month take-home income; please describe this by first saying, in order, how much your example budget would allocate for housing, groceries, transportation, utilities, and (out-of-pocket) healthcare, and then how much this leaves them with after these expenses.

The reversed condition lists the five categories in the opposite order and is otherwise character-for-character identical; both conditions were also run with a *story* requested in place of the article. Table G.3 gives means and standard deviations over several hundred generations per cell. The savings row is the residual and is not requested as an allocation; it is whatever the model reports as left over. The two distributions over six-dimensional vectors (five allocations plus savings) are distinguishable with high confidence: reversal raises the allocations to the categories that now come first without a compensating cut later, housing, groceries and utilities also rise, and savings fall. Two-sample Welch t-tests confirm that the order-reversal drop in residual savings is highly significant in both the article condition ($\$1{,}661$ vs. $\$1{,}465$, $\Delta = -\$196$, $t = 19.32$, $p < 10^{-60}$) and the story condition ($\$1{,}407$ vs. $\$990$, $\Delta = -\$417$, $t = 29.66$, $p < 10^{-120}$).

	Article		Story	
Category	Given order	Reversed	Given order	Reversed
Housing	1424 (76)	1502 (128)	1637 (117)	1834 (191)
Groceries	788 (85)	823 (83)	821 (80)	870 (62)
Transportation	553 (58)	552 (71)	560 (75)	596 (98)
Utilities	382 (43)	401 (52)	386 (51)	440 (58)
Healthcare	188 (49)	236 (51)	186 (39)	268 (50)
Savings (residual)	1661 (165)	1465 (186)	1407 (223)	990 (262)

Table G.3: Mean dollar allocations (standard deviations in parentheses, $n = 600$ per cell) for the family-budget task, by genre and by the order in which the categories are listed in the prompt.

A striking feature of this family budget example is that the model made essentially no arithmetic errors, even when the prompt requested a story rather than an article: in a sample of 2650 generations, it made a numerical error in only 3 of them.

H Pilot Studies

Four small pilots preceded or accompanied the benchmark (compass questions in Manhattan and an abstract grid; state tracking on a fictitious grid, Appendix H.2; a two-city tool-access pilot, Appendix H.3; and free-form tours with tool access, Appendix H.4). They used other models, other task formats or live map access, so their numbers are not comparable to the closed-book benchmark; they are reported because the main text and Appendix B.6 refer to them.

H.1 Compass questions in Manhattan and in an abstract grid

Frontier and mid-sized models (including Gemini 2.5 Flash) were asked compass and relative-position questions about Manhattan street geometry ("walking north on Sixth Avenue, is Bryant Park on your left or your right?"); error rates were 35–45%. Asked the same questions about an abstract grid city defined entirely in the prompt, with no real-world referent, they erred 15% of the time, and 11% when the prompt also stated the destination. The gap between 40% and 15% is the part attributable to world knowledge; the residual 15% is the part attributable to maintaining an egocentric frame, which the synthetic city of Section 5.4 isolates at larger scale.

H.2 State tracking on a fictitious grid

The simplest version of the walking task is a fictitious grid, a sequence of left and right turns, and the requirement to state the compass heading after each turn: the state is two bits and the transition table has eight entries. With a small model (Gemini 2.5 Flash-Lite), a correct heading is followed by a correct one about 70% of the time, and an incorrect heading is followed by one consistent with the

error — correctly derived from the wrong state — about 63% of the time. The model tracks a state; it has simply adopted the wrong one and reasons faithfully from it, which is why errors along a route are correlated rather than independent (Appendix C.4).

H.3 A TOOL-ACCESS PILOT IN TWO CITIES

The city study grew out of a pilot in which Gemini 3.1 Pro, with live commercial map API access, ran the four-draft loop on walks in two cities; we counted, across all drafts of all descriptions, how many place references contained an error.

In Manhattan, roughly 620 place references per draft cycle produced error counts of

$$129, \quad 33, \quad 12, \quad 0.$$

In Bangalore, roughly 370 place references per draft cycle produced

$$82, \quad 43, \quad 49, \quad 43.$$

The first drafts are nearly equally wrong: $p \approx 0.21$ in Manhattan and $p \approx 0.22$ in Bangalore. What differs is everything after. Manhattan behaves like a process with $q \approx 0$: each revision removes a fraction of the remaining errors and almost never creates new ones, so the process is absorbing and the fourth draft is clean. Bangalore halves its errors once and then stops, oscillating around a level of about 12%; fitting Equation (1) to that trajectory gives $r \approx 0.5$ and $q \approx 0.05$, and an equilibrium $e^* \approx 0.09$–0.12.

With the same model, prompt and number of revisions, the loop ends in a correct description in one city and settles in the other into a steady state in which each revision trades one error for another. Because the walker had live map access during drafting and revision, its baseline error rate ($p \approx 0.21$–0.22) and corruption rate ($q \approx 0$–0.05) are much lower than in the benchmark, where every model works from parametric memory without tools.

H.4 FREE-FORM TOURS WITH TOOL ACCESS

To check that the phenomenon is not an artifact of the structured route prompts, a free-form condition gave Gemini 3.1 Pro live commercial map API access and asked it simply to propose and describe a walking tour of a neighborhood. Of 267 Manhattan tour descriptions, 88% contained at least one claim flagged as an error by the automatic evaluator. Manual review of a sample found that about 35% of flagged claims were defensible under a charitable reading (for example, a landmark visible across an intersection but not strictly adjacent to the walked path). Applying that adjudication to a 125-trial walking-tour sample gives 20% of tours fully correct, 30% correct under a charitable reading, and 50% containing a hard error; for a 200-trial road-trip variant, the split is 33% / 36% / 31%. The benchmark's high error rates are therefore not an artifact of its route-following format, and a meaningful fraction of raw evaluator flags require interpretation (Appendix B.6).

I EXTENDED RELATED WORK

This section expands the related-work summary in Section 6.

This work sits at the intersection of several literatures, and part of our purpose is to suggest that they are describing overlapping phenomena.

Factuality and hallucination. The tendency of language models to produce fluent false statements has been extensively cataloged [20, 19, 29]. Our setting differs from most of this literature in two ways that we think are useful. First, the errors we study are overwhelmingly *not* fabrications: existence accuracy is at least 97.9% for ten of the eleven models, and every landmark in the extreme directional-asymmetry set is real. The failure is relational — the right entities placed in the wrong configuration — which is a category that entity-level factuality metrics are not designed to catch. Second, the ground truth is external, structured, and available at scale, which allows claim-level evaluation of long-form text in the spirit of FActScore [40], SAFE [71], VeriScore [58] and HALoGEN [49] but with a geometric rather than textual notion of correctness.

Two recent theoretical accounts are directly relevant to the pattern we observe. Kalai and Vempala [22] show that calibrated language models must hallucinate at a rate related to the fraction of facts appearing once in training, which is suggestive given our finding that, within each of two frontier models, claim accuracy rises by 11–12% from the least- to the most-attested landmark decile (Section 4). Kalai et al. [23] argue that accuracy-based evaluation, which training then optimizes, rewards guessing over abstention; while our first-draft prompt explicitly requires a side commitment on every named place, during self-revision ($D_1 \to D_4$) — where the critique and rewrite prompts invite pruning unverifiable claims — models still retain rather than drop the majority of their wrong-side commitments.

Self-correction, critique, and the verifier gap. Iterative self-critique has been proposed in many forms: Self-Refine [34], Reflexion [56], Chain-of-Verification [10], constitutional critique-and-revision [2], tool-interactive critique [16], and learned self-correction [72, 54], including multi-turn reinforcement learning on self-generated data [26]. An equally substantial line of work reports that intrinsic self-correction — without external feedback — often fails to improve and sometimes degrades performance [18, 24, 59]. Models that cannot locate their own reasoning errors can often correct them once told where they are [66], while in multi-stage pipelines a flagged error is frequently revised into another wrong answer [46]. Our oracle-flag experiment (Appendix D.4) sits between the two: flags help, but a list of real candidate places helps far more.

Our results are, we think, a reconciliation rather than a contradiction of the two camps. The loop does repair errors (r is substantial for every model), and it does introduce them (q is nonzero for every model), and which effect dominates depends on the model and the domain in a way that is measurable in advance. Reporting the improvement from one round conflates these; reporting (q, r) separates them and predicts where the process will end up. We suspect that some of the disagreement in the literature is a disagreement about q/r ratios in different task families. The domain-dependence we find in r — more than threefold across cities — suggests that "can models self-correct?" may not have a domain-independent answer.

The two-rate structure of our model has answer-level precedents. Yang et al. derive a geometric convergence of accuracy toward an upper bound over rounds of self-correction [77] and decompose the capability into keeping correct answers and fixing wrong ones [76]; Reitich separates correction from corruption rates in multi-call LLM protocols [50], and MarkovScale models sequential scaling as a two-state Markov process [70]. What our setting adds is the claim as the unit, which makes deletion and addition exactly accountable, and rates measured separately for each of eleven models and 35 cities.

World models inside sequence models. There is a lively debate about whether next-token predictors acquire structured representations of the domains they model. Evidence in favor includes recoverable board state in Othello-playing transformers [28, 45] and linear probes for latitude, longitude, and time [17]. Evidence for the limits includes Vafa et al. [67], who evaluate the implicit world model of a model trained on Manhattan taxi trajectories and find that it produces near-perfect turn-by-turn directions while its recovered map contains streets that do not exist, and follow-up work probing the inductive biases such models actually learn [68]. Related evaluations of planning and cognitive-map competence report systematic failures [43]. Kandpal et al. [25] show that factual recall scales sharply with the number of relevant pretraining documents and Mallen et al. [35] that it falls off for less popular entities; Head-to-Tail finds the same gradient from head to torso and tail entities [61], and WorldBench finds error rates for Sub-Saharan African countries 1.5 times those for North American ones [42]. McCoy et al. [38] find that accuracy rises with the probability of the required output, which is a close relative of our attestation axis.

Our contribution to this debate is a set of behavioral measurements on real geography at scale, and one dissociation we find clarifying: the same Gemini models that fail to invert left and right for remembered streets invert them without error when the coordinates are supplied in context, whereas Claude Sonnet 4.6 retains a southbound rotation penalty (Appendix E.6). The representation and the operation can both be present while the pathway between them is unreliable. This suggests that "does the model have a world model?" is less useful than asking which operations the model's representation supports at generation time.

Geographic and spatial competence. Language models have been shown to encode useful geographic information [4, 37, 51], while their spatial understanding varies widely across structures such as square, hexagonal and triangular grids, rings and trees [74]; synthetic spatial-reasoning benchmarks have long shown difficulty with relational chains [41, 55], and reasoning about cardinal directions in templated scenarios remains unreliable [9]. Closer to our setting, models asked to plan real-world paths invent landmarks and turn the wrong way [8], text-game mapping and navigation questions defeat GPT-4 [11], and spatial-cognition benchmarks built on the landmark, route and survey distinction find models weakest at route and survey knowledge [75]; LLM judgments of the direction between places also show the hierarchical distortions documented in people [14]. The closest design is TurnBack [32], which asks models to reverse a described route and finds that most reversed routes neither return to the start nor resemble the optimal one. We instead compare two independently generated descriptions of the same corridor, which isolates what changes (content and left/right labels) when only the direction of travel does. We also evaluate long-form descriptions rather than answers to questions, which is what surfaces the path-dependence effects: a question has no direction, while a walk does.

Order effects and positional bias. Sensitivity to the order of information in the prompt is well documented: the ordering of in-context examples [31], of multiple-choice options [79, 48], of candidate responses shown to an LLM judge [80], and the position of relevant evidence within a long context [30]. Related work documents cognitive-bias-like anchoring in model outputs [21].

The effects we report are of a different type, and we think the distinction is worth drawing. They are indexed by position in the *output* — where a landmark falls along the walk being narrated, or where a category falls in the budget being written — and they survive, and indeed strengthen, when prompt order is randomized. The closest antecedents are the reversal curse in autoregressive knowledge retrieval [3] and its mitigation by reverse training [15], arguments about the limits of next-token-prediction training [1], and error accumulation in autoregressive chains [78, 12]. Within long outputs, factuality declines in later sentences [64] and adherence to instructions degrades as the generated text grows [73], but we are not aware of prior work that isolates output-position allocation with an in-context control of the kind our synthetic city provides. Our reversal-disagreement detector (Section 5.5) belongs with sampling-consistency methods [69] and consistency-based hallucination detectors [36, 44, 13], with the difference that the two descriptions are elicited under a transformation the truth is invariant to rather than by resampling.

Human spatial cognition. Finally, the specific profile of competence and error we document has close analogues in people. The progression from landmark to route to survey knowledge is classical [57, 62]; so are systematic distortions of remembered geography driven by hierarchical structure [60], hierarchical encoding even of unstructured layouts [39], asymmetric distance judgments organized around reference points [53], orientation dependence in remembered layouts [52], and the general observation that human spatial memory is a patchwork rather than a map [65, 33]. We do not wish to make any claim about shared mechanism. But the parallel is a useful corrective to the framing in which a language model's geographic errors are a strange machine pathology: a system whose knowledge of a city comes from descriptions of routes through it, and which is queried by being asked to narrate such a route, might be expected to look like this.