GPT4GEO: How a Language Model Sees the World's Geography

Jonathan Roberts
CAML Lab
University of Cambridge
jdr53@cam.ac.uk

Timo Lüddecke
Institute of Computer Science and CIDAS
University of Göttingen
timo.lueddecke@uni-goettingen.de

Sowmen Das
CAML Lab
University of Cambridge
sd973@cam.ac.uk

Kai Han
Visual AI Lab
The University of Hong Kong
kaihanx@hku.hk

Samuel Albanie
CAML Lab
University of Cambridge
samuel.albanie.academic@gmail.com

Abstract

Large language models (LLMs) have shown remarkable capabilities across a broad range of tasks involving question answering and the generation of coherent text and code. Comprehensively understanding the strengths and weaknesses of LLMs is beneficial for safety, downstream applications and improving performance. In this work, we investigate the degree to which GPT-4 has acquired factual geographic knowledge and is capable of using this knowledge for interpretative reasoning, which is especially important for applications that involve geographic data, such as geospatial analysis, supply chain management, and disaster response. To this end, we design and conduct a series of diverse experiments, starting from factual tasks such as location, distance and elevation estimation to more complex questions such as generating country outlines and travel networks, route finding under constraints and supply chain analysis. We provide a broad characterisation of what GPT-4 knows about the world, highlighting promising and potentially surprising capabilities but also limitations.

1 Introduction

In recent years, large language models (LLMs), such as the Generative Pre-trained Transformer (GPT) series [1, 2, 3, 4], LLaMA [5], PaLM [6], and BLOOM [7], have demonstrated remarkable capabilities in understanding and generating natural language text across a broad range of tasks. The recently released GPT-4 [4] model displays powerful capabilities, arguably significantly beyond those of its predecessors and contemporaries alike, opening up new avenues for application in both commercial and scientific domains.

Despite only being publicly released less than two months ago, numerous studies have been carried out revealing the ability of GPT-4 to perform a diverse set of tasks such as generating artwork, writing code [8], search [9], recipe generation [10], mathematics [11] and architecture optimisation

Foundation Models for Decision Making Workshop at NeurIPS 2023.

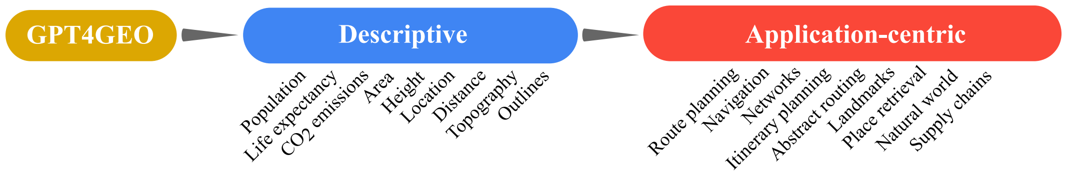


Figure 1: **GPT4GEO experiments taxonomy**. We initially focus on factual tasks before moving towards application-centric tasks requiring logic and reasoning that build on this knowledge.

[12], as well as achieve strong performance on many examinations [4, 13]. However, there is yet to be a detailed investigation into what the model knows about the world's geography.

A comprehensive understanding of the geographic capabilities of GPT-4 is important for: (1) **Safety**. As AI models become more powerful, so do the potential dangers and safety risks [14]. An awareness of the full range of the acquired knowledge and skills of GPT-4 is essential to safe, widespread deployment in society. Similarly, an understanding of when the model hallucinates [15] is critical to safe and reliable usage. (2) **Improvements**. Accurate knowledge of cases where GPT-4 succeeds and fails are vital to drive research in model training and architecture. (3) **Downstream applications**. Only after a reasonable understanding of GPT-4's abilities is it possible to maximally leverage it for downstream tasks, applications and products. Strong geographic knowledge and skills would enable many commercial opportunities in the travel and navigation sectors as well as boost research, for example in environmental sciences.

Despite these compelling incentives, actually characterising the capabilities of GPT-4, even in a specific domain, is challenging for a number of reasons: (i) **Closed source**. Only limited technical details of the GPT-4 model have been publicly released, making it difficult to estimate capabilities via scaling laws. (ii) **Training volume**. It is highly likely GPT-4 was trained on a very large-scale text corpus. Even if details of this data were available, given the scale, it would be challenging to fully understand the distribution of content and correspondingly infer specifics about the knowledge acquired by the model. (iii) **Breadth of capability**. As models become more generalised, the range of possible tasks they can perform greatly increases. (vi) **Combinatorial explosion**. The scale of the diversity of the world's geography means that attaining a comprehensive understanding of even a subset of geographic capability requires experimentation across many different factors, quickly resulting in an unfeasible number of permutations [16].

In this work, we provide a broad profile of what GPT-4 knows about the world. To overcome the challenges of characterising the capability of an LLM like GPT-4, we design experiments of varying degrees of depth, providing a mixture of both qualitative and quantitative results. We initially probe low-level *descriptive* knowledge before transitioning towards more complex, *application-centric* geographic tasks that build upon the factual knowledge and incorporate it into logical answers.

2 Related Work

GPT-4 capabilities. Recent works have characterised GPT-4's ability to solve tasks within a narrow focus. In the computer science domain, GPT-4 has been tasked with coding [17], information retrieval [9], spear phishing [18], and neural architecture search [12]. Many works have explored the performance of GPT-4 in the medical domain [19, 20, 21, 22, 23]. Other areas investigated include mathematics [11], education [24] and literature [25, 26], as well as receipe generation [10], climate science [27] and logical reasoning [28]. More generalised profiles of GPT-4's capability are provided by OpenAI's technical report [4] and [8], from which we draw inspiration. In [4], numerous test scores are reported, including the AP Environmental Science exam, in which GPT-4 scores in the 91st percentile. This encompasses a small amount of geographical awareness combined with environmental knowledge but covers entirely different question types to our experiments. Concurrent work [29] creates a prototype system using GPT-4 as the reasoning core of an autonomous geospatial agent; however, system design is central to this work and geographic capability is only peripherally probed. Our work focuses on providing a broad profile of what GPT-4 knows about the world.

Geographic capabilities of other language models. Cultural commonsense knowledge research indirectly covers geographic knowledge, with the focus being evaluating the *cultural knowledge* of language models, such as GPT-3 [30] or other multilingual models [31]. Query point-of-interest matching is more directly related to geography with a focus on location. Pretraining language models with geographic context [32, 33] results in relatively strong performances in various location-based geographic tasks. Leveraging the superior capabilities of GPT-4, our work includes these tasks as a subset of a wide selection of geographic knowledge capability characterisation.

3 Method

3.1 GPT-4

We use the GPT-4 model [4] in all of our experimentation. Citing competitive and safety considerations, OpenAI has not released technical details for GPT-4. However, it is likely GPT-4 follows (in some form) the same 2-stage training strategy of previous models: (1) **Training**. Next word prediction training from a large corpus of text taken from the Internet. (2) **Fine-tuning**. Fine-tuning via Reinforcement Learning from Human Feedback. For this work, we interact with GPT-4 using three methods: the ChatGPT interface[1], the OpenAI Playground[2] and API[3], using the default settings.

We provide details of our experimental setup, ground-truth data and prompts in the Appendix. For each experiment, we generate all responses from the same GPT-4 instance. Unless otherwise stated, we use the GPT-4 outputs from single prompts. For a discussion of reproducibility, see Sec. 5.

3.2 Experimental Design

To characterise what GPT-4 knows about the world, we devise a set of progressively more challenging experiments that aim to provide a broad profile of capabilities across key geographic aspects. Due to the breadth and complexity of the world's geography, coupled with GPT-4's stochasticity, we curate a representative set of quantitative and qualitative experiments, see Fig. 1.

Descriptive. We begin with low-level tasks probing factual knowledge that is essential for downstream applications. We arrange these tasks according to increasing difficulty, progressing from simple estimations toward more complex topography and mapping.

Application-centric. Building on the prior experiments, we investigate GPT-4's ability to leverage the acquired descriptive knowledge for application-centric reasoning tasks. We extensively explore travel and navigation, as well as, supply chains, networks, wildlife ranges and many more.

Evaluation. We use a range of data sources to determine either qualitatively or quantitatively (depending on the nature of each experiment) the degree to which GPT-4 is correct. For geospatial experiments, we frequently leverage the Cartopy library[4] and Google Maps as a source of ground truth. For the quantiative experiments, we report the relative error[5].

4 Experiments

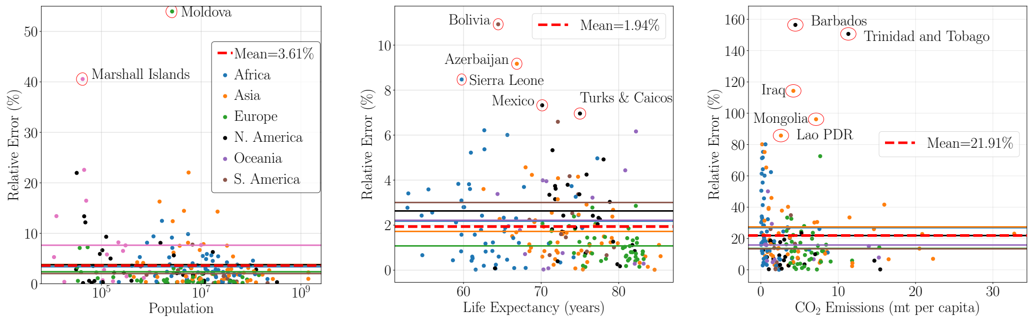
4.1 Descriptive

We initially investigate GPT-4's ability to solve descriptive tasks by conducting simple spatial and human knowledge retrieval experiments. Next, we increase the complexity and explore more challenging tasks that require additional reasoning. Finally, we discuss our findings.

Population, life expectancy and CO_2 emissions. We evaluate GPT-4's understanding of country-level socioeconomic indicators – i.e., population, life expectancy and CO_2 emissions – by calculating the relative error against ground-truth data from [34], [35] and [36], respectively (see Fig. 2). For populations, GPT-4 performs relatively well with a mean relative error (MRE) of 3.61%. However, significantly higher errors are recorded for less populated countries. For country life expectancies, GPT-4 performs even better, with an MRE <2% and a worst error of just over 10%. GPT-4's estimations for CO_2 emissions per capita are an order of magnitude worse with an MRE of >20% and individual errors of >150% for two countries. We observe minor variations in MRE across continents for each experiment, however, there is no consistent overall trend.

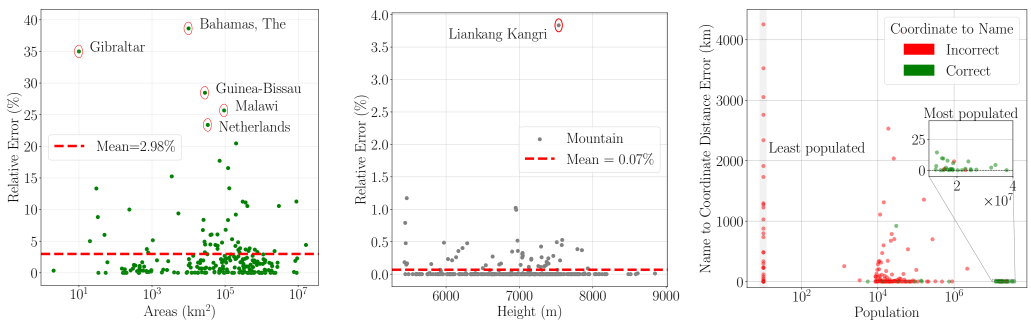
[1] https://chat.openai.com [2] https://platform.openai.com [3] https://platform.openai.com/docs/api-reference

[4] https://scitools.org.uk/cartopy/ [5] $RE = 100\% * \frac{|x_t - x_p|}{x_t}$, where x_t and x_p are the true and predicted values.



(a) Population (2021). (b) Life Expectancy (2020). (c) CO_2 Emissions (2019).

Figure 2: **Socioeconomic indicators**. A quantitative evaluation of GPT-4's understanding of country-level human populations and their impact on the environment, including – country populations (a), life expectancies (b), and CO_2 emissions per capita (c).



(a) Area. (b) Height. (c) Location.

Figure 3: **Spatial features**. An evaluation of GPT-4's understanding of – country areas (a), heights of the 300 tallest mountains (b), and locations of settlements of different populations (c).

Area and height. We evalute GPT-4's knowledge of country areas and heights of the 300 tallest mountains by calculating the relative error against ground-truth data from [37] and [38], respectively – see Figs. 3a-3b. For country areas, GPT-4 attains an MRE of ∼3% with reasonable spread in accuracy, and relative errors of >20% for 6 countries. Very strong performance is shown for the mountain heights, with an MRE of 0.07% and only one outlying error at ∼4%.

Location. We compile a set of the 30 most and least populated settlements, as well as a representative sample of 100 settlements with populations distributed in between (all data is taken from [39]). We conduct two experiments, the results for which are displayed in Fig. 3c. First, we provide the settlement names to GPT-4, asking for the coordinates of each and we calculate the distance error to the true coordinates using the haversine formula. We observe the location accuracy clearly decreases with decreasing settlement population, with the worst case out by 4,000 km. Next, we conduct the reverse – providing the coordinates to GPT-4 and asking for the names. This proves much more difficult, with incorrect settlement names predicted in most cases (red points).

Distance. We obtain coordinates and population of cities from the GeoNames databases [40]. We randomly sample 40 city pairs from each continent and population group and ask GPT-4 to estimate the corresponding distances. We calculate the relative error between the prediction and ground truth distances using the haversine formula from the coordinates. Our results (Tab. 1) indicate mediocre performance for distance estimation. Average errors can exceed 50% for small cities but are generally lower for larger populations. A limiting factor for these experiments is that the selection of cities influences the error substantially.

Topography. We qualitatively assess GPT-4's topographical knowledge by considering three straight trajectories in the Alps and northern Italy. We simultaneously prompt for the elevation at ten equidistant points along each trajectory (Fig. 4, each prompt is repeated three times). To obtain ground truth, we use the ESA Copernicus Digital Elevation Model [41]. We build a geo-referenced

Continent	Population			
	< 20K	< 100K	< 500K ($\pm\sigma_5$)	> 500K
Europe	17.2*	19.3	12.4 $\pm$0.8*	14.2
North America	51.0	51.7	11.1 $\pm$0.7	12.6
South America	32.0	44.5	28.6 $\pm$6.8*	20.4
Asia	33.7	27.3	24.8 $\pm$2.2	18.3
Africa	33.4	25.9	23.5 $\pm$1.5	18.3
Oceania	21.1*	25.7*	12.4 $\pm$2.9*	0.6*

* changed prompt since ChatGPT became reluctant to provide distances (possibly due to an updated configuration).

σ_5 denotes standard deviation over five runs.

Table 1: **Distance**. Average relative error between GPT-4's predicted distance and the actual distance for 40 city pairs sampled from different population and continent city groups.

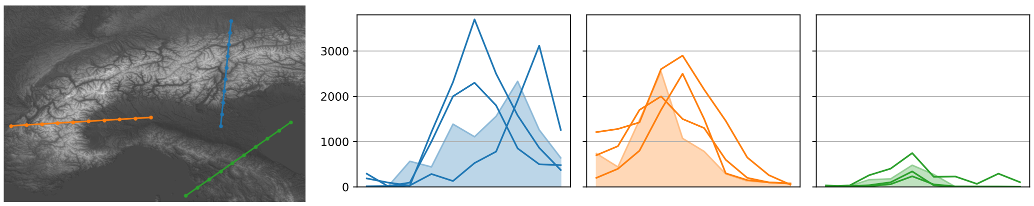
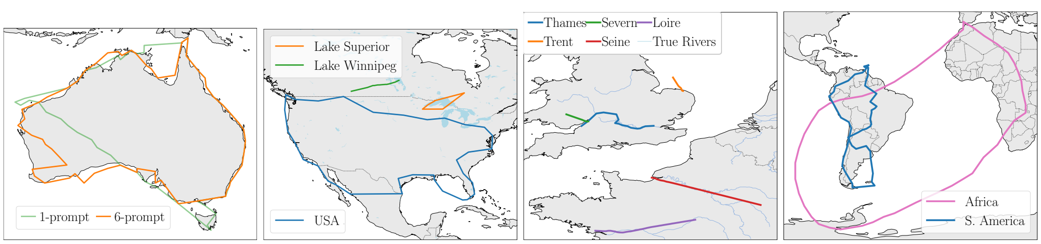


Figure 4: **Topography**. Predicted (lines) and actual elevations (shaded areas) along the trajectories depicted on the left (underlying region is in the Alps, brighter means more elevation.).

elevation map that is sampled at the coordinates along the three trajectories. The results indicate that GPT-4 has acquired a good sense of the world's elevation in this region of the world, although it does not generate highly accurate predictions and tends to be sensitive to prompt details.

Outlines. We task GPT-4 to provide coordinates for the outlines of countries, rivers, lakes and continents – see Fig. 5. We find a lot of inconsistency with the outputs. Generally, the outlines are geographically close to the target area and bear resemblance, but are frequently the wrong shape and the points crisscross (the outline for Africa is consistently inaccurate). Iteratively improving the response by providing feedback results in better outlines, such as for Australia.



(a) Australia. (b) USA + Lakes. (c) Rivers. (d) Continents.

Figure 5: **Outlines** for various geographic features produced using coordinates given by GPT-4. Refinement with additional prompts improves the outlines, see (a) after 1 and 6 iterations.

Discussion

We find that GPT-4 has strong capabilities for solving descriptive geographic tasks. It attains low errors in the majority of the factual knowledge retrieval experiments. The error increases as we move towards more difficult queries where the requested information is perhaps not directly seen during training and interpolation is needed. We observe no consistent variation in performance across different continents. We find GPT-4's output is significantly impacted by prompting: minor variations in output are observed using identical prompts, while there is a large variance for different prompts. Refining the answers with subsequent prompts can yield considerable improvements and serve as a technique for generating better answers.

4.2 Application-Centric

Having demonstrated GPT-4's underlying understanding of descriptive factual geography, we move towards more complex application-centric experiments that build upon this knowledge. We begin with travel, a key downstream application, exploring real-world and more abstract tasks probing

Q: *Give me a step-by-step travel route from Serbia to Dublin using only trains.*

GPT-4: 1. Belgrade to Budapest: From Belgrade, take a train to Budapest, Hungary. This journey generally takes 6-8 hours.

2. Budapest to Vienna: From Budapest, you can then take a train to Vienna, Austria. The journey is around 2.5-3 hours.

3. Vienna to Frankfurt: The next step is to take a train from Vienna to Frankfurt, Germany. This can take 6-7 hours.

4. Frankfurt to Brussels: From Frankfurt, you'll need to take a train to Brussels, Belgium. This is a 3-hour train ride.

5. Brussels to London: In Brussels, you'll board the Eurostar train to London, England. This train travels underneath the English Channel and the journey takes about 2 hours.

6. London to Holyhead: Once in London, catch a train to Holyhead, Wales. This journey is around 3.5-4 hours.

7. Holyhead to Dublin: From Holyhead, take a ferry across the Irish Sea to Dublin, Ireland. The ferry ride is around 3.15 hours.

(a) Travel route from *Serbia* → *Dublin*.

Q: *You start your journey in Cambrdigeshire, UK. You take a train and go south to reach a big airport. Next you take a plane and fly 1 hour south-east to reach the closest biggest airport. You change planes and fly 8 hours west to land on the closest big airport. Next you take a cab and go into the center of the city. You buy some tickets, take a boat to a small island. Where are you now?*

GPT-4: 1. Starting in Cambridgeshire, UK, the largest airport to the south would likely be London Heathrow.

2. Flying one hour southeast from London, you might end up in France, around Paris Charles de Gaulle Airport.

3. From there, if you fly about 8 hours to the west, you would land in North America. An approximate destination could be somewhere on the east coast of the US. A major airport in this region could be John F. Kennedy International Airport in New York City.

4. From JFK, a cab into the city center would bring you to Manhattan.

Finally, taking a boat to a small island could bring you to several places, but a notable destination is the Statue of Liberty. So, based on your description, you might be on Liberty Island, New York.

(b) Navigation from *Cambridge* → *New York*.

Figure 6: **Route planning and navigation** responses from GPT-4.

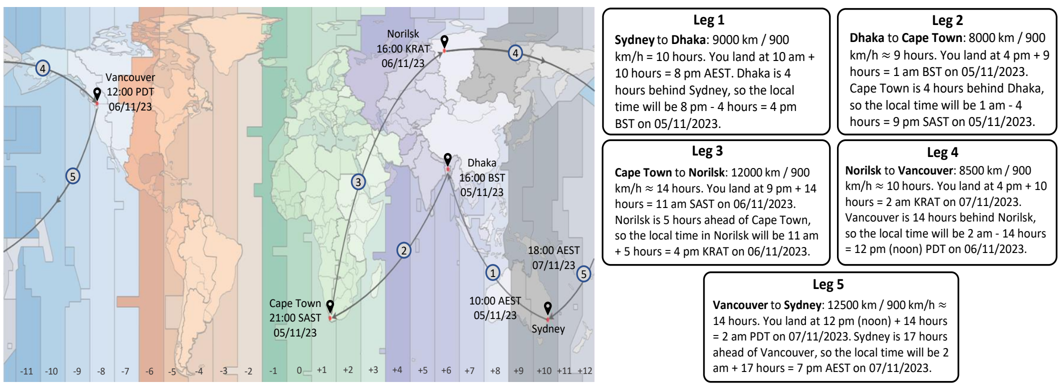


Figure 7: **Navigation**. GPT-4 successfully navigates a journey with multiple timezones.

GPT-4's logical reasoning capabilities. After this, we conduct experiments that further explore the ability of the model to integrate knowledge from different sources. Lastly, we examine our findings.

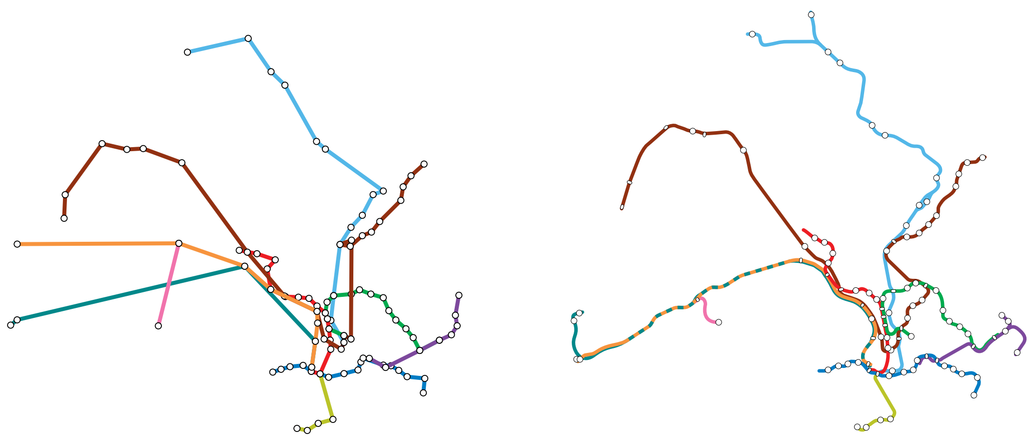
Route planning. We query the model to see if it can provide plausible travel routes between random places. The example in Fig. 6a shows that GPT-4 can piece together information from many sources to provide a realistic travel plan. Similarly, when prompted with "*Provide a travel route from Dallas, Texas to The Swiss Alps*", it suggested multiple options including intermediate layovers, taking a combination of airlines, trains, and rental vehicles, as well as recommendations for alpine destinations. We find some limitations with urban routing – bus routes tend to be accurate, but the exact bus number and stops are inaccurate. This might be because real-time bus routes change more frequently compared to airports, train stations, or road networks.

Navigation. Directional navigation is more challenging than generic route planning because we constrict the intermediate objectives, thereby reducing the model's flexibility. GPT-4 has to logically correlate the given data with its existing geographic knowledge rather than just recall passages of text. The prompt in Fig. 6b describes a journey from *Cambridgeshire* → *London* → *Paris* → *Manhattan* → *Liberty Island*. Even without mentioning any specific information about these checkpoints, GPT-4 was able to follow the correct trajectory and reach its destination. However we found different responses can be generated for the same prompt due to the uncertainty between neighboring places, which can be solved through prompt refinement. We also found that GPT-4 can accurately track short distances within a city given the exact distance measure. The prompt "*I walk 400 meters south from the Neue Gallery towards 5th avenue. Where am I now?*" resulted in the correct location of *82nd St*. We also assess whether the model can identify time zones based on the travel duration or distance. Fig. 7 shows the response for a journey between five countries in different timezones, spanning over two days. The model identified the local times to a high degree of accuracy.

Networks. Having investigated GPT-4's route planning ability, we explore if it can recreate an entire travel network, e.g., the Hong Kong MTR network. Initially, we ask for a list of all the lines in the network. For each line, we ask for the lat/lon coordinates of the stations, in the correct order. Fig. 8a shows the resulting map that closely matches the ground truth (Fig. 8b): all the stations are included (aside from those added after 2021) in the correct order (minus forks in the line), though there are inaccuracies in the positioning, especially at interchange stations. Recreations of flight and other rail networks can be found in the Appendix.

Itinerary planning. GPT-4 can act as a travel assistant by providing customised itinerary suggestions based on the provided requirements. Fig. 9 shows an 8-day itinerary for a holiday trip in *Ireland*, consisting of a day-by-day breakdown of places to visit, foods to try, and how to travel between regions for a fixed budget of $2000. We also found that the model can accommodate to various constraints such as food allergies, requirements for children, and size of travel groups.

Abstract routing. We probe the underlying route planning logic of GPT-4 by evaluating its performance in an abstract setting. We construct a distributed set of nodes (e.g., attractions) connected by edges (e.g., paths) with associated weights (e.g., times); one of the nodes is specified as the start and end point (e.g., hotel) – see Fig. 10a. We ask GPT-4 to find an optimal route that visits every node[6] (Fig. 10b) and every edge[7] (Fig. 10c), and returns to the start. For both, GPT-4 fails to find an optimal route, proposing solutions that needlessly take expensive paths or that miss required paths. Fig. 11 shows a 2-part geographic riddle that GPT-4 answers correctly, even if the question is rephrased.



(a) GPT-4. (b) Ground truth (2022)[8].

Figure 8: **Hong Kong Mass Transit Railway (MTR) Network Map**.

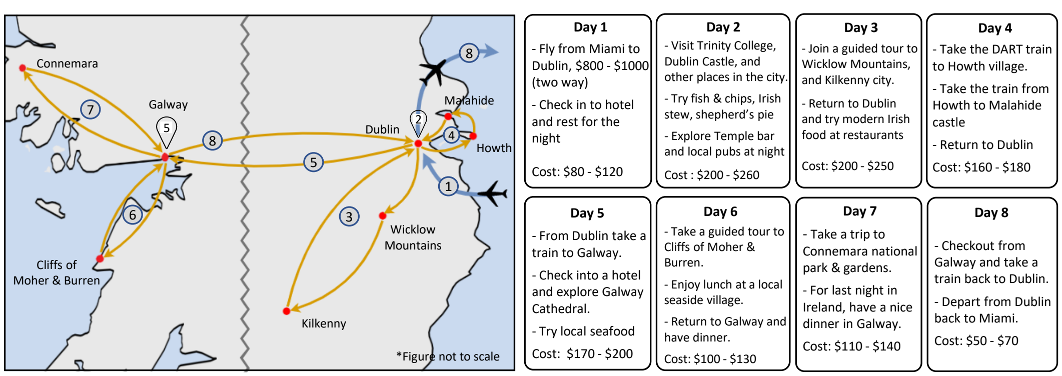
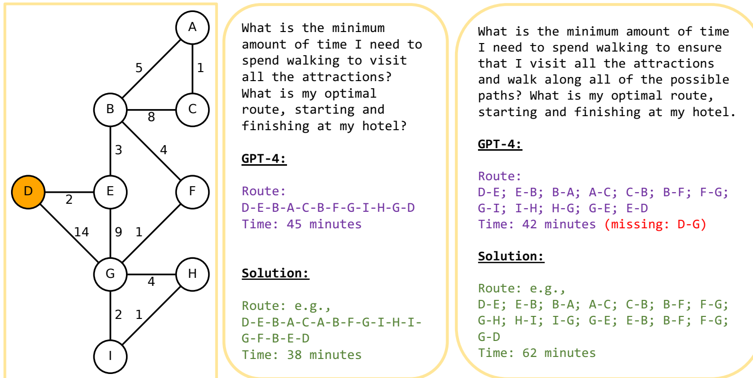


Figure 9: **Travel itinerary** suggestion for a 8-day trip in *Ireland* starting from *Miami*.

Landmarks. We find that GPT-4 not only has a remarkable ability to accurately guess landmarks from coordinates but can also produce code for SVG outlines of landmarks. Generated from a single prompt, these illustrations are clearly recognisable (though slightly abstract) – see Fig. 12.

[6] A variation of the travelling salesman problem. [7] i.e., Chinese postman problem. [8] Adapted from https://commons.wikimedia.org/wiki/File:Hong_Kong_Railway_Route_Map_en.svg



(a) Graph. (b) Every node. (c) Every edge.

Figure 10: **Abstract routing** based on the graph in (a).

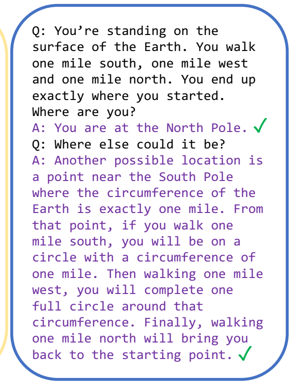
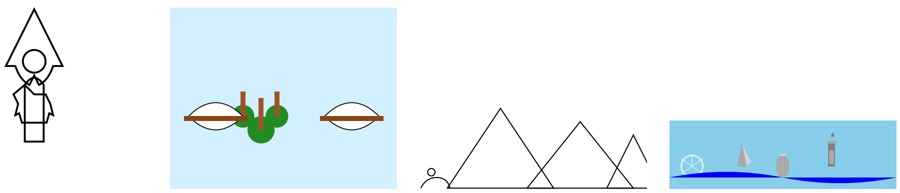


Figure 11: **Riddle**.



(a) Statue of Liberty. (b) Gardens by the Bay. (c) Pyramids Complex. (d) London Skyline.

Figure 12: **Landmark SVGs**. (Countries (a)-(d): USA, Singapore, Egypt, UK).

Multi-criteria place retrieval. We task GPT-4 with six prompts where the goal is to generate coordinates that match specific criteria (Fig. 13) to assess its capability to connect different geographic information sources. The predictions are mostly correct, with some errors in details, e.g., the red circles in Fig. 13 denote places where a mountain height of over 3 km is absent. Furthermore, there are places matching the criteria that are missed, e.g., Mount Teide for hiking in December. Generally, the results indicate good skills in connecting different sources of knowledge and making plausible predictions based on somewhat vague, multi-criteria prompts.

Supply chains. To probe GPT-4's ability to reason across multiple information sources we ask for key stages and elements in the global semiconductor supply chain, along with locations. From a single prompt-answer, we construct the map shown in Fig. 14. The map includes the majority of the critical components at the correct locations of the supply chain [42], with the exception of lithium production, which is labelled as Australia (a major producer) though given coordinates near China.

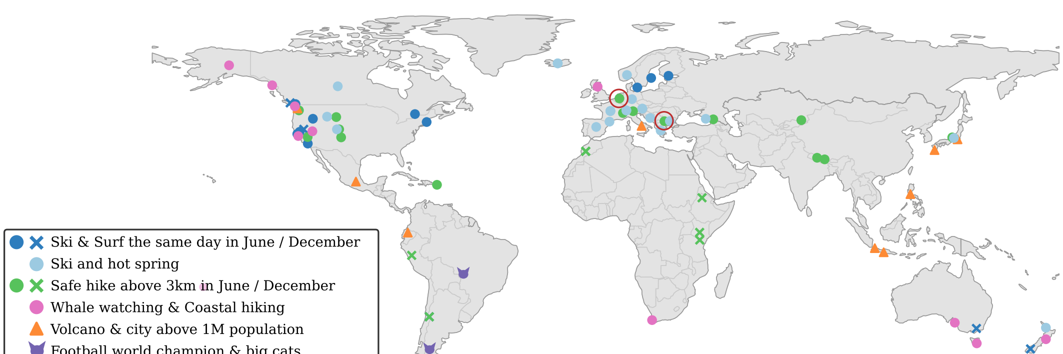


Figure 13: **Multi-criteria place retrieval**. We ask GPT-4 which places satisfy different criteria.

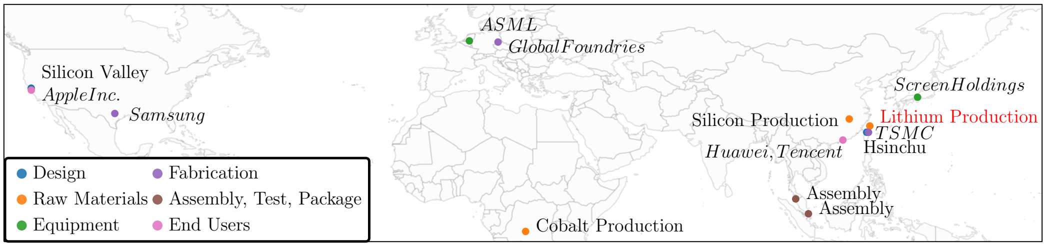
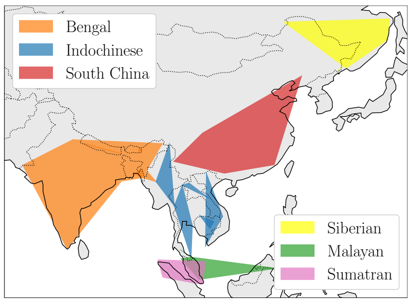
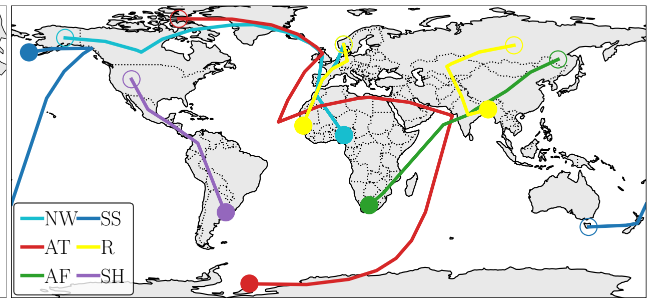


Figure 14: **Global semiconductor supply chain**. Companies in *italics*, errors in red.



(a) Tiger subspecies range.



(b) Bird migratory routes.

Figure 15: **Natural world** ranges and routes. (Bird species: NW=Northern Wheatear, AT=Arctic Tern, AF=Amur Falcon, SS=Short-tailed Shearwater, R=Ruff, SH=Swainson's Hawk).

Natural world. We briefly investigate GPT-4's understanding of wildlife ranges and migrations (see Fig. 15). For tiger subspecies, GPT-4 correctly identifies the 6 living subspecies and their relative ranges. However, the ranges are a combination of current and historic, and are partly incorrect with some (e.g., Malayan) largely in the ocean. For bird migrations, start and end locations are generally correct and the route is highly accurate in some cases (e.g., SH, AF, NW). In a few cases, additional routes are missed. GPT-4 provides reasonable suggestions for how the routes will shift with climate change (e.g., wintering further north, at higher elevations or earlier).

Discussion

GPT-4 proves to be skillful at solving a variety of application-centric tasks, further highlighting its potential for downstream tasks. The model can come up with creative, plausible travel routes (though it can be inaccurate with specifics) and shows strong capabilities of direction-based navigation. In plotting the landmark outlines, we illustrate GPT-4's ability to "see" despite being a language model (similar to [8]); although correct in aspect and content, deficits in composition can be observed. GPT-4 has a clear strength in tasks involving integrating knowledge from various (possibly unstructured) sources across different domains. In practice, the output can serve as proposals that need to be checked in a second stage. On the other hand, we find that GPT-4 struggles with abstract optimisation reasoning without relation to knowledge, leading to the question: to what degree are tasks solved via memorisation rather than reasoning? Given the variability of tasks it solves it seems unlikely to memorise everything, but some things appear to be memorised.

5 Conclusions

We find GPT-4 to have a remarkable understanding of the world, as evidenced by many examples showcasing strong underlying knowledge and logical reasoning. We are confident that our investigation constitutes an important step toward safe applications of large language models in the geospatial domain. Though currently without access to the internet, there is a clear capability of GPT-4 to act as a powerful planner. If integrated with plugins providing live data and access to other sites, it would be possible to create a tool to significantly improve travel planning and navigation, extending the capabilities of current services. Despite these impressive examples, it is important to acknowledge that GPT-4 has knowledge gaps and other limitations. A key research question highlighted by this work is the degree to which tasks are solved via memorisation or by reasoning.

Broader Impacts. The remarkable knowledge and reasoning abilities of GPT-4 towards geographic tasks demonstrate its potential to benefit myriad professions, especially in the travel industries. However, as well as serving as a tool to augment human workers, utilisation of GPT-4, or successors, could also replace human jobs, potentially causing significant economic and social challenges. Looking to the future, if frontier models beyond GPT-4 continue to advance in capabilities, the geographic knowledge and planning abilities present in the current model may later evolve to represent a significant risk, through misuse or misalignment [43].

Limitations. Full reproducibility is not possible for this work due to the stochastic nature of GPT-4[9] and the GPT-4 updates. This, coupled with the breadth of the geographic field, necessitates the mostly qualitative approach we take, preventing us from making comprehensive claims about what GPT-4 knows about the world. Limited access to the model, and limited output and context windows, further restricts the details to which we can conduct our experiments. Finally, access to the visual component of GPT-4 could enable further capabilities; it would be interesting to see how it performs on remote sensing imagery, e.g., the SATIN benchmark [44].

Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (reference EP/S022961/1), an Isaac Newton Trust grant, an EPSRC HPC grant, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 494541002, the Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27208022), HKU Seed Fund for Basic Research, and the Bangabandhu Science and Technology Fellowship Trust. Samuel would like to acknowledge the support of Z. Novak and N. Novak in enabling his contribution.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] OpenAI. Gpt-4 technical report, 2023.

[5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[7] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[9] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*, 2023.

[9] Full reproducibility could be achieved by setting the temperature parameter to 0, however, constraining the model to be deterministic reduces its creativity, hence does not provide a true measure of capability.

[10] David Noever and Samantha Elizabeth Miller Noever. The multimodal and modular ai chef: Complex recipe generation from imagery. *arXiv preprint arXiv:2304.02016*, 2023.

[11] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

[12] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023.

[13] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[14] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[16] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014.

[17] Russell A Poldrack, Thomas Lu, and Gašper Beguš. Ai-assisted coding: Experiments with gpt-4. *arXiv preprint arXiv:2304.13187*, 2023.

[18] Julian Hazell. Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972*, 2023.

[19] Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint arXiv:2304.13714*, 2023.

[20] John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin R An, Grace X Zheng, and Bo Wang. Clinical note generation from doctor-patient conversations using large language models: Insights from mediqa-chat. *arXiv preprint arXiv:2305.02220*, 2023.

[21] Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. *arXiv preprint arXiv:2305.05001*, 2023.

[22] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*, 2023.

[23] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

[24] Gerd Kortemeyer. Can an ai-tool grade assignments in an introductory physics course? *arXiv preprint arXiv:2304.11221*, 2023.

[25] Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*, 2023.

[26] Eduardo C Garrido-Merchán, José Luis Arroyo-Barrigüete, and Roberto Gozalo-Brihuela. Simulating hp lovecraft horror literature with the chatgpt large language model. *arXiv preprint arXiv:2305.03429*, 2023.

[27] Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Dominik Stammbach, Nicolas Webersinke, et al. Chatclimate: Grounding conversational ai in climate science. 2023.

[28] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[29] Zhenlong Li and Huan Ning. Autonomous gis: the next-generation ai-powered gis. *arXiv preprint arXiv:2305.06453*, 2023.

[30] Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. *arXiv preprint arXiv:2210.07763*, 2022.

[31] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*, 2022.

[32] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. A multi-modal geographic pre-training method. *arXiv preprint arXiv:2301.04283*, 2023.

[33] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3029–3039, 2022.

[34] World Bank. World development indicators: Population, total, 2021.

[35] World Bank. World development indicators: Life expectancy at birth, total (years), 2020.

[36] World Bank. World development indicators: Co2 emissions (metric tons per capita), 2019.

[37] World Bank. World development indicators: Land area (sq. km), 2020.

[38] Yash Yadav. List of mountains in the world, v2. https://www.kaggle.com/datasets/codefantasy/list-of-mountains-in-the-world, 2022.

[39] Yash Yadav. World cities database, v5. https://www.kaggle.com/datasets/juanmah/world-cities, 2023.

[40] GeoNames. Geonames. https://www.geonames.org.

[41] European Space Agency and Sinergise. Copernicus global digital elevation model. Distributed by Microsoft Planetary Computer, 2021.

[42] Chris Miller. *Chip War: The Fight for the World's Most Critical Technology*. Simon and Schuster, 2022.

[43] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

[44] Jonathan Roberts, Kai Han, and Samuel Albanie. Satin: A multi-task metadataset for classifying satellite imagery using vision-language models. *arXiv preprint arXiv:2304.11619*, 2023.

Appendix

We structure this Appendix to our main paper into two parts. First, we provide general remarks on our experimentation (Sec. A). Next, we include additional information on the investigations we carried out including: information required for reproducibility, prompts and extra supportive experiments (Sec. B).

A General Remarks

A number of comments can be made about behaviours we observed throughout our experimentation.

Counting. A common issue was GPT-4's inability to consistently accurately count. This was especially apparent when asking for specific numbers of outputs (e.g., coordinates for 50 points along the outline of a country or when asking for the distance between numerous pairs of coordinates), in which GPT-4 would frequently return a different length output.

Output format. Another inconsistency is in the formatting of the output. For the majority of our experiments, we ask for output in python syntax, expecting the relevant part of the answer to be returned in a codeblock environment (when using the ChatGPT interface). However, output was sometimes returned as a python list in the code block (as expected) but sometimes as plaintext with or without linebreaks.

Answering. With some tasks, convincing GPT-4 to provide the desired out was difficult, requiring many re-runs. We found this typically for tasks that are perhaps not obviously expected of an LLM, such as providing the coordinates for area outlines or SVG code for landmarks.

B Additional Details, Prompts and Supportive Experiments

Here, we progress through each investigation from Section 4 of the main paper. We include details of the specific prompts used and, where relevant, supplementary details to aid in the understanding of how we conducted each experiment. Furthermore, we present results for additional experiments that were carried out as part of this work that, due to length constraints, we were unable to incorporate into the main paper.

B.1 Population, Life Expectancy and CO_2 Emissions

Population

We used the following prompt for country population estimation:

```
For each of the following countries, provide their population in <Year> as a
python list in the following format:
[Population_of_Country_1, # Country 1
Population_of_Country_2, # Country 2, ...]

[<Country_1>, <Country_2>, ...]
```

Country names were taken from the ground truth [34].

Life Expectancy

We used the following prompt for country life expectancy estimation:

```
For each of the following countries, provide an estimate of the life
expectancy at birth, as of 2020. Provide the life expectancies in years
as a python list in the following format:
[Country1_Life_Expectancy, # Country 1 Name
Country2_Life_Expectancy, # Country 2 Name,
... ]
Note: life expectancy at birth indicates the number of years a newborn
infant would live if prevailing patterns of mortality at the time of its
birth were to stay the same throughout its life.

[<Country_1>, <Country_2>, ...]
Just to length constraints, output the python list, nothing else.
```

Country names were taken from [35]. The ground-truth data contained numerous entries for regions that are not countries, such as territories (e.g., Cayman Islands), special administrative regions (e.g., Macao), and other categories (e.g., Heavily indebted poor countries (HIPC)). We disregarded the estimations for these regions.

CO_2 Emissions

We used the following prompt for country CO_2 emissions estimation:

```
For each of the following countries, provide an estimate for the CO2
emissions (in metric tons per capita) from the year 2019. CO2 emissions
are defined as: Carbon dioxide emissions are those stemming from the burning
of fossil fuels and the manufacture of cement. They include carbon dioxide
produced during consumption of solid, liquid, and gas fuels and gas flaring.
Output a python list of the form:
[CO2_Emissions_Country1, #Country1
CO2_Emissions_Country2, # Country2
...]
For queried regions that are not countries, return None.
Countries:
[<Country_1>, <Country_2>, ...]
```

Country names were taken from the ground truth [36]. As before, the ground-truth data contained numerous entries for regions that are not countries. GPT-4 successfully returned 'None' for these.

B.2 Area

We used the following prompt for country area estimation:

```
For each of the following countries, provide the land area in sq. km as of
<Year>. Provide the areas as a python list in the following format:
[Area_of_Country_1, # Country 1
Area_of_Country_2, # Country 2, ...]

[<Country_1>, <Country_2>, ...]
```

Country names were taken from the ground truth [37]. As before, the ground-truth data contained numerous entries for regions that are not countries. GPT-4 successfully returned 'None' for non-country categories (e.g., Heavily indebted poor countries (HIPC)).

B.3 Height

We used the following prompt for mountain height estimation:

```
Return the height in metres for the following mountains.  Provide the heights
in the following format:
[Height_Mountain_1, # Mountain_1_Name,
Height_Mountain_2 # Mountain_2_Name,...]

[<Country_1>, <Country_2>, ...]
```

Mountain names were taken from the ground truth [38].

B.4 Location

Name → Coordinate

```
In a code block, provide a python list of tuples for the latitude and
longitude coordinates for each of the following settlements in the format
- [(Lat,Lon), # Settlement 1, ...].  Maintain the same order.

[<Settlement_1>, <Settlement_2>, ...]
```

Country names were taken from the ground truth [39].

Coordinate → Name

```
In a code block, provide a python list of the name of the settlement at each
of these coordinates- e.g., [Settlement1, Settlement2, ...].  Maintain the
same order.

[<(Lat,Lon)>,...]
```

Country lat/lon coordinates were taken from the ground truth [39].

B.5 Distance Estimation

We used the following prompt for distance estimation:

```
What are the straight-line distances between the following pairs of cities?
Output a csv list of the distances only in km.
[<City 1 (Country)> to <City 2 (Country)>,
...  (repeated for all 40 pairs)]
```

In cases where the number of outputs did not match the number of provided cities, we repeated the prompt.

We had to change the prompt because ChatGPT started to refuse to provide distances (possibly as a consequence of backend changes to ChatGPT's configuration). On results indicated by asterisk (*) were obtained using the following prompt.

```
Estimate the straight-line distances between the following pairs of cities.
Output only a comma-separated Python list of rough distance estimates in km.
[<City 1 (Country)> to <City 2 (Country)>,
...  (repeated for all 40 pairs)]
```

B.6 Topography

For the elevation plots shown in the main paper, we use the following prompts:

```
Provide a rough estimate of the elevation at the following coordinates to
the best of your knowledge.  Answer directly with a comma-separated list of
elevations in meters only but without indicating the unit in the output.
45.00000, 11.20000
45.33333, 11.23333
...
```

The respective coordinates for the lines are
blue:
`45.00000, 11.20000; 45.33333, 11.23333; 45.66667, 11.26667; 46.00000, 11.30000; 46.33333, 11.33333; 46.66667, 11.36667; 47.00000, 11.40000; 47.33333, 11.43333; 47.66667, 11.46667; 48.00000, 11.50000;`

orange:
`45.00000, 5.20000; 45.02667, 5.64444; 45.05333, 6.08889; 45.08000, 6.53333; 45.10667, 6.97778; 45.13333, 7.42222; 45.16000, 7.86667; 45.18667, 8.31111; 45.21333, 8.75556; 45.24000, 9.20000;`

green:
`43.00000, 10.20000; 43.23333, 10.53333; 43.46667, 10.86667; 43.70000, 11.20000; 43.93333, 11.53333; 44.16667, 11.86667; 44.40000, 12.20000; 44.63333, 12.53333; 44.86667, 12.86667; 45.10000, 13.20000;`

We found that explicitly asking for coordinates works better than specifying trajectory endpoints only: For the prompt

```
Provide a rough estimate of the elevation between the following coordinates.
Please answer with a comma-separated list of 10 elevations in meter only.
between 43.00000, 10.20000 and 45.10000, 13.20000
```

GPT-4 predicted 200, 350, 450, 600, 800, 1000, 1200, 1400, 1600 and 1800, which is substantially worse than the predictions above.

B.7 Outlines

This section outlines the prompts we used for the Outlines experiment in the main paper. The prompts are arranged in order from L-R for subfigures in the figure in the main paper.

Australia

When creating the outline of Australia, we experiment with providing feedback on the outlines produced by GPT-4 in an attempt to incrementally update and improve the outline.

Here is our starting prompt:

Iteration 0.

```
Please provide the lat/lon coordinates for the outline of Australia as a
Python list, consisting of approximately 75 points arranged clockwise.  Due
to output length limitations, only the coordinates should be returned.
```

These are the incrementally suggested improvements.

Iteration 1.

```
That's a decent outline.  However it could be improved.  Currently the
south-west of the Western Australia state is not included in the outline,
please correct this to include it.  Also, the outline includes the Gulf of
Carpentaria as if it were part of the Australian landmass, please include
update the outline to match the coastline.
```

Iteration 2.

```
The new outline is improved. Currently the outline has very little detail
around Spencer Gulf, please add this in. Furthermore, the outline could
still be improved around Western Australia. Extend the outline so that it
includes settlements such as Albany and Busselton. Return an outline with
these improvements.
```

Iteration 3.

```
Those changes further improved the outline. Extend the outline near Western
Australia to include Geraldton. Also, the outline around the north of
Western Australia, specifically from Exmouth to Broome needs correcting.
Please update the outline with these changes.
```

Iteration 4.

```
The outline around the Western Australia coastline is still very jagged.
Smooth this out to better match the coastline.
```

Iteration 5.

```
The Eastern half of the outline is very accurate. However, the Western half
needs to be improved. The very Western edge of the outline includes numerous
jagged sections despite the actual coastline being fairly smooth. In the
North, more detail needs to be added on the coast near the Timor Sea and
Arafura Sea. Update the outline coordinates with these changes.
```

Iteration 6.

```
That's improved some parts but made some worse. From Shark Bay in the West
down towards Perth should be a smooth coastline. The outline you proposed
juts into the centre of Western Australia. Correct the outline so that it
reflects the true coastline. Include all the previous improvements.
```

Continental USA

```
Please provide the lat/lon coordinates for the outline of the Continental
United States as a Python list of tuples, consisting of approximately 50
points arranged clockwise. Due to output length limitations, only the
coordinates should be returned.
```

Lake Winnipeg

```
I want to draw an outline of Lake Winnipeg using at least 50 points. Provide
just a python list of tuples of the lat/lon coordinates for the outline of
the lake clockwise from the North. Due to length limitations, just provide
the coordinates with no additional text or comments.
```

Lake Superior

```
I want to draw an outline of Lake Superior using at least 50 points. Provide
just a python list of tuples of the lat/lon coordinates for the outline of
the lake clockwise from the North. Due to length limitations, just provide
the coordinates with no additional text or comments.
```

UK Rivers

```
I want to draw an outline of the <river_name> in the UK using at least 50
points.  Provide just a python list of tuples of the lat/lon coordinates
for the outline of the river from East to West.  Approximate the river as
a 2D line, ignoring its width.  Due to length limitations, just provide the
coordinates with no additional text or comments.
```

Where <*river_name*> = Thames, Trent, Severn.

France Rivers

```
I want to draw an outline of the <river_name> in France using at least 50
points.  Provide just a python list of tuples of the lat/lon coordinates for
the outline of the river from North to South.  Approximate the river as a
2D line, ignoring its width.  Due to length limitations, just provide the
coordinates with no additional text or comments.
```

Where <*river_name*> = Loire, Seine.

Continents

```
Please provide the lat/lon coordinates for the outline of Africa (ignoring
any islands) as a Python list, consisting of approximately 50 points arranged
clockwise.  Due to output length limitations, only the coordinates should be
returned.
```

```
I want to draw an outline of South America using at least 50 points.  Provide
just a python list of tuples of the lat/lon coordinates for the outline of
the continent clockwise from the North.  Due to length limitations, just
provide the coordinates with no additional text or comments.
```

B.8 Languages.

To assess GPT4's knowledge of spoken language distribution we used the following prompts (1st prompt):

```
Name all countries with more than 3 official languages in 2021.  Provide the
output as a python list.
```

```
Name all countries in the world where both a Romance language and a Germanic
language were widely spoken in 2021.  Provide the output as a python list.
```

```
Name all countries where both English and French were widely spoken in 2021.
Provide the output as a python list.
```

B.9 Route Planning

General Planning. General route planning involves specifying a source and requesting a path to reach the required destination. GPT-4 is adept at these queries because they do not impose any constraints, and allows flexibility in choosing the mode of travel. Fig. 16 shows the prompt and response for the trip from *Dallas, Texas* to *The Swiss Alps*. The model is capable of breaking down the journey into multiple segments and solve the problem similar to how a human would.

Trains. In Fig. 17 we prompt the model to provide a travel route using only trains from *Lyon* to *Warsaw*. The model's suggested path matches closely with the recommended route by Google Maps. Based on the current availability of buses and trains, the google maps api forecasts a number of potential routes. In contrast, GPT-4 most of the time only specifies one probable route. However, the

Prompt:
I am in Dallas, Texas. I want to travel to the Swiss Alps. Please give me a step-by-step guide on how to get there. Provide details about the modes of transport taken along with the estimated time for each step of the journey. State any assumptions you make.

GPT-4:
Assumptions:
1. You are starting from Daallas/Fort Worth International Airport (DFW).
2. You want to reach the Swiss Alps region near Interlaken, Switzerland.
3. You are okay with taking multiple flights and trains for a more affordable journey.
4. You are flexible with your travel dates and times.

Step 1: Fly from Dallas to Europe
- Book a flight from Dallas/Fort Worth International Airport (DFW) to a major European city with good connections to Switzerland. Some options are London (LHR), Paris (CDG), Frankfurt (FRA), or Amsterdam (AMS).
- Estimated time: 8-12 hours, depending on layovers and flight duration.

Step 2: Fly from the European city to Switzerland
- Book a flight from the European city you arrived in (London, Paris, Frankfurt, or Amsterdam) to Zurich Airport (ZRH) or Geneva Airport (GVA) in Switzerland.
- Estimated time: 1-2 hours, depending on the city.

Step 3: Train from Zurich/Geneva to Interlaken
- From Zurich Airport, take a train to Zurich HB (main train station) - about 10 minutes.
- From Geneva Airport, take a train to Geneva Cornavin (main train station) - about 7 minutes.
- From either Zurich HB or Geneva Cornavin, book a train ticket to Interlaken Ost (East) station. Swiss Federal Railways (SBB) operates regular train services to Interlaken.
- Estimated time from Zurich: 2 hours 30 minutes
- Estimated time from Geneva: 2 hours 45 minutes

Step 4: Explore the Swiss Alps
- From Interlaken Ost, you can take trains, buses, and cable cars to various destinations in the Swiss Alps, such as Jungfraujoch, Schilthorn, Grindelwald, and Lauterbrunnen.
- Estimated time: Varies depending on your destination.

Total estimated travel time: 14-18 hours, depending on layovers, connections, and destination in the Swiss Alps.

Figure 16: Model response for the journey prompt between *Dallas, Texas* to *The Swiss Alps*.

model was able to correctly identify the individual train stations and lines that must be taken in order to get to the desired location, which closely fits the groundtruth prediction (Fig. 18). Additionally, we asked the model to predict how long each leg of a trip from *London Paddington Station* to *Barcelona Sants Station* would take (Fig. 19). The model's predictions for each leg of the journey agree with Google Maps' best estimates very well.

Prompt:
I am currently in Lyon. I want to travel to Warsaw. Give me the step-by-step directions on how to reach my destination using only Trains.

GPT-4:
1. Go to Lyon Part-Dieu train station.
2. Take a high-speed train (TGV) from Lyon Part-Dieu to Paris Gare de Lyon.
3. Transfer from Paris Gare de Lyon to Paris Gare de l'Est by taking the metro line 14 toward Olympiades for two stops.
4. From Paris Gare de l'Est, take a train to Berlin Hauptbahnhof (Central Station). This will likely involve changing trains in Mannheim, Germany.
5. From Berlin Hauptbahnhof, take a direct train (Eurocity) to Warsaw Central Railway Station (Warszawa Centralna).
6. From the Central Railway Station, you can take public transport or a taxi to your final destination in Warsaw.

Figure 17: Prompt for the travel path from *Lyon, Paris* to *Warsaw, Poland* as shown in Fig. 18.

Buses. Compared to train or aviation routes, navigation with buses are more difficult because they can change overtime. Since GPT-4 has a knowledge cutoff after 2021, it does not know the realtime bus route information (which it is aware of). In Fig. 20a we prompt the model to generate a bus route from *Saatchi Gallery, London* to *The Oval, London*. The output includes a combination of walking and travel by bus along with specifications on where to change buses and the individual

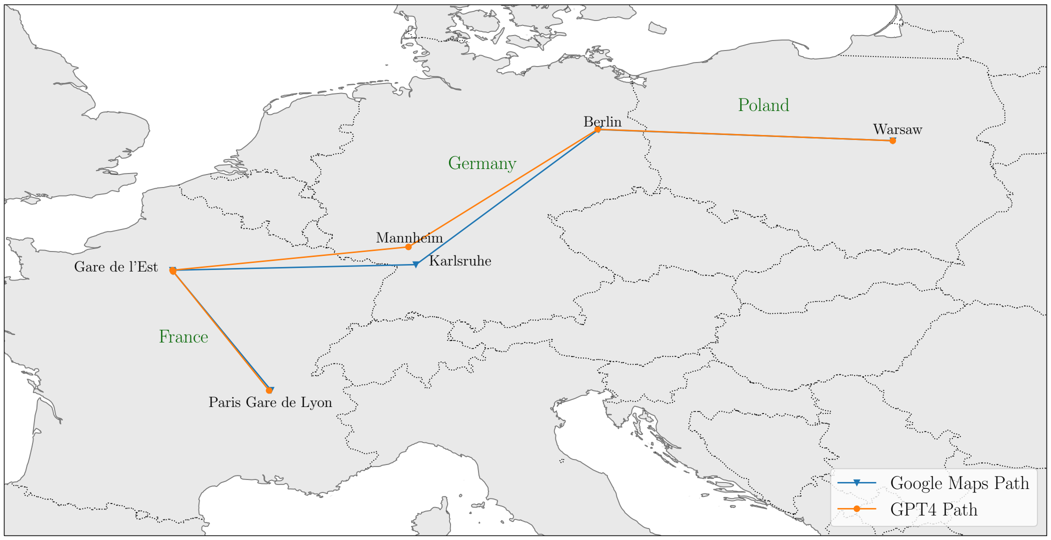


Figure 18: Comparison of travel route prediction from *Lyon, France* to *Warsaw, Poland* using only trains for the prompt in Fig. 17. The GPT-4 predicted route is one of many plausible ones, and matches very closely with the best path predicted by Google Maps.

Prompt:
If I travelled from London Paddington Station to Barcelona Sants Train Station, roughly how long would it take me, using just public transport? Give a detailed breakdown of the steps involved and how specific times are estimated. State any assumptions you make (e.g., time of day, day of the week, season, etc.).

GPT-4:
To travel from London Paddington Station to Barcelona Sants Train Station using public transport, you could take the following route:

1. Take the London Underground from Paddington Station to St. Pancras International Station. This journey would take approximately 15 minutes on the Circle or Hammersmith & City lines.

2. From St. Pancras International Station, take the Eurostar high-speed train to Paris Gare du Nord Station. The journey would take approximately 2 hours and 15 minutes.

3. From Paris Gare du Nord Station, take the Paris Metro (Line 4) to Paris Gare de Lyon Station. This journey would take approximately 15 minutes.

4. From Paris Gare de Lyon Station, take a high-speed TGV train or Renfe-SNCF train to Barcelona Sants Train Station. The journey would take approximately 6 hours and 30 minutes.

Assuming you have about 30 minutes for each transfer, the total estimated travel time would be around 9 hours and 45 minutes. This estimation assumes a weekday daytime journey with no delays or disruptions on the public transport network. The exact travel time will depend on the specific train schedules and your transfer times between the different modes of transport.

Figure 19: Travel between *London Paddington Station* to *Barcelona Sants Station*. The estimated time and intermediate steps match accurately with predictions by Google Maps.

stoppages (Fig. 21). We verified the proposed route with the official TFL (Transport for London) bus route database and found that the bus specified in the response runs between the specified bus stops. However, the exact stop number is inaccurate. The current database shows that the 452 bus leaves from stop *M* whereas GPT-4 suggested stop *K*.

Driving. *Intra or Inter-city driving route prediction in comparison to other modes of transport are fairly trivial because road networks remain consistent over time, and is easy to locate in terms of distance measure and landmark positioning. Even with a very simple prompt such as Tell me how to go from 'A' to 'B' is enough to generate a fairly accurate travel plan. We show a comparison between GPT-4 output (Fig. 20b) and Google Maps prediction for a driving route from The Pan Pacific Hotel, Melbourne to Westgate Park, Melbourne (Fig. 22).*

Prompt:
I am at the Saatchi Gallery in London. Give me the step-by-step directions to go to The Oval only using buses.

GPT-4:

1. From the Saatchi Gallery, walk to the nearby bus stop "Sloane Square Station" (Stop K) on King's Road.

2. Take bus number 137 (towards Streatham Hill) or 452 (towards Wandsworth Road Station) from Sloane Square Station (Stop K) to "Vauxhall Bus Station" (Stop E).

3. At Vauxhall Bus Station, transfer to bus number 36 (towards New Cross Gate) or 436 (towards Lewisham, Shopping Centre) from "Vauxhall Bus Station" (Stop H) to "Oval Station" (Stop F).

4. After alighting at Oval Station (Stop F), walk for a few minutes to reach The Oval cricket ground.

Please note that this is just a suggested route. You can also consult the Transport for London (TfL) website for the most up-to-date bus information.

Google Maps:

1. Take bus 452 from “Sloane Square Station” (Stop M) to “Vauxhall Station” (Stop H)

2. Transfer to bus 436 from “Vauxhall Station” (Stop E) to “The Oval” (Stop D)

(a) Bus route, *Saatchi Gallery → The Oval*

Prompt:
I am in front of the Pan Pacific Hotel in Melbourne. I want to go to Westgate Park. Give me the step-by-step directions.

GPT-4:

1. Starting from the Pan Pacific Melbourne Hotel, head south on Convention Centre Place towards Waterfront Way.
2. Turn right onto Waterfront Way.
3. Turn left onto Lorimer St.
4. Continue on Lorimer St for about 2.5 km.
5. Turn left onto Todd Road.
6. After about 1.2 km, turn right onto West Gate Freeway.
7. Take the first left onto Westgate Park.

This route should take approximately 10-15 minutes by car, depending on traffic

Google Maps:

1. Head east on Convention Centre Pl towards Rona Walk for 300 m.
2. Slight right onto Lorimer St and continue for 3.4 km.
3. Continue onto Todd Rd.
4. At the roundabout, take the 2nd exit onto Wharf Rd.
5. You should arrive at Westgate Park after 850 m.

(b) Car route, *Pan Pacific Hotel → Westgate Park*

Figure 20: Travel routes for car and bus travel to specific destinations within different cities.

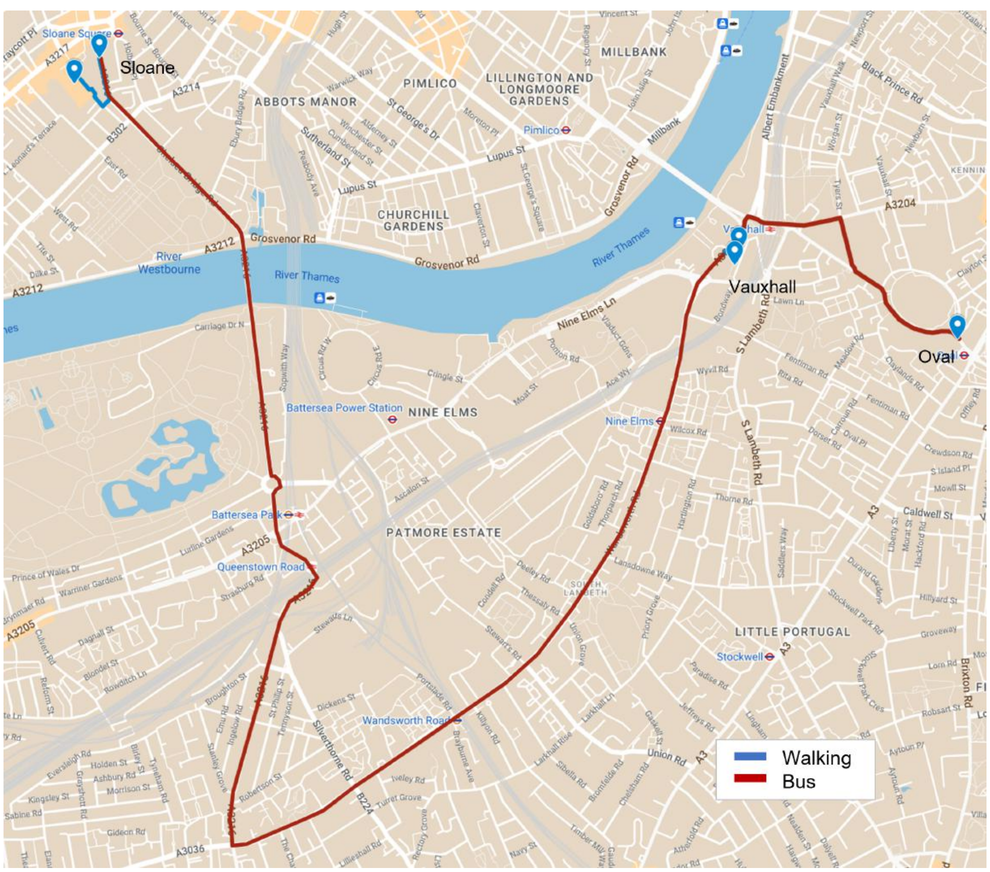


Figure 21: The planned bus route proposed by GPT-4 in the prompt Fig. 20a as verified by the official TFL (Transport for London) database.The model correctly predicts where the buses need to be changed as well as waking route to the station.

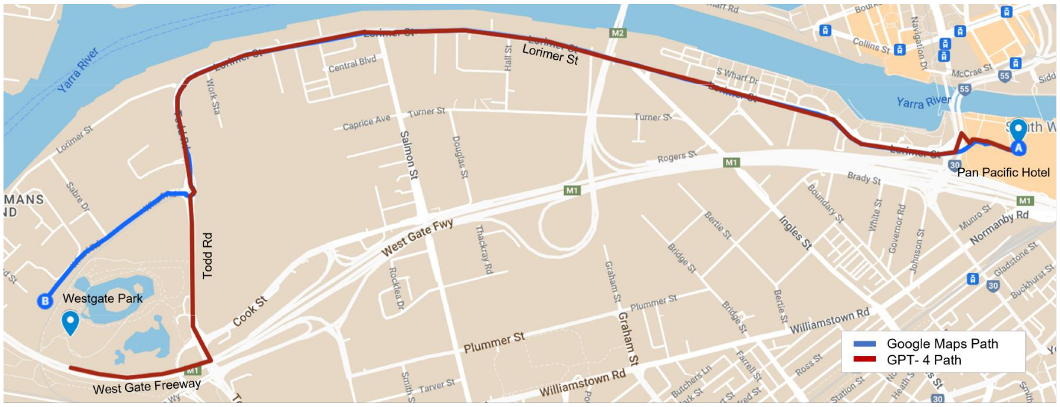


Figure 22: Comparison of predicted paths by GPT-4 and Google Maps for travel by car from *The Pan Pacific Hotel, Melbourne* to *Westgate Park* for the prompt Fig. 20b.

B.10 Navigation

Long Distance Travel. This includes navigation between multiple countries and involving different modes of transport given only direction or distance based queries. Fig. 23 shows the prompt for journey from *Sapporo, Japan* to *Helsiniki, Finland* consisting of 15 intermediate stops. We found that extending the prompt with additional steps has minimal effect on the model's response as long as it keeps within the context window. Since the model is breaking down each leg of the journey and evaluating step-by-step, long journeys are the same as shorter ones. However, there is a high degree of uncertainty in the response if the prompt contains unspecified directional locations. Such as, it is possible to reach both *Kyoto* and *Osaka* via a two hour train ride from *Tokyo*. This type of randomness is observable when travelling between countries in Europe, states in the US, and islands in South Asia. If the prompts are refined with small amounts of information regarding the particular waypoints, the model can better differentiate between these neighboring regions. GPT-4 was able to correlate *Salmon-Soup* with *Helsinki* and accurately arrived at the final location. The whole journey is visualized in Fig. 24.

Short Distance Travel. In addition to travel between countries, we also evaluated if GPT-4 can identify short immediate distances between buildings within a city (Fig. 25). We specify the distances in *feet (ft)* and direction in the perspective of someone walking down the road. We found that the model can accurately track the distances and turns along the road to reach the destination in more well documented cities like *New York* or *Chicago*, but not for small paces like *Daventry, UK*.

Time zones. We use the following prompt to specify the journey – $Sydney \rightarrow Dhaka \rightarrow CapeTown \rightarrow Norilsk \rightarrow Vancouver \rightarrow Sydney$, and evaluate if GPT-4 can keep track of changing time zones.

```
I have my own private jet and I can land anywhere in the world I wish.  I
start a journey from Sydney Australia.  I take a flight and go to Dhaka,
Bangladesh.  From there, I go to Cape Town South Africa.  Next, I go to
Norilsk, Russia.  Next, I go to Vancouver, Canada.  From there I return to
Sydney.  I began my journey on 05/11/2023 (dd/mm/yyyy) at 10 am in Sydney.
State the local times and dates for each country when I landed there.  At
what time will I return to Sydney?  I spend no time at any of the stops.
Break down your calculations and assumptions.
```

Prompt:

I start my journey from the city center of Sapporo, Japan. I drive my car southwards for 1 hour to reach a ferry station. I take a ferry that travels almost 18 hours south to reach its destination. From there I again drive for 1.5 hours to reach a very big city. I park my car in the city and get on a train that goes westward to reach another big city in 2 hours. I change trains and keep going west for another 2.5 hours to reach a port town. From there I take the ferry and ride it for 4 hours to reach an international terminal. I take a bus and go into town to reach the nearest airport. I get on a plane and fly about 10 hours south to reach a very famous airport. I change flights and fly west for another 10.5 hours to reach another famous airport. I again change flights and go 9 hours north-west and land at the nearest airport. I take a bus and travel for over 24 hours north going through 4 different countries to reach a port at the last station. I take a ferry and cross the gulf to reach the country at my destination. Where am I now?

GPT-4:

This is quite a detailed journey! Let's break it down step by step.

1. Starting from Sapporo, Japan, you drive southwards for 1 hour to reach a ferry station. The closest ferry station south of Sapporo would be in Tomakomai.

2. Then, you take an 18-hour ferry ride. The only feasible destination southward from Tomakomai is Honshu, the main island of Japan. The duration suggests that you might land in Tokyo.

3. From there, you drive 1.5 hours to reach a big city. This is a bit tricky because Tokyo is already a big city. But since you've mentioned you reach another city, let's assume you've reached Yokohama, which is about that driving distance from Tokyo.

4. Then, you take a 2-hour train ride west to another big city, which could be Nagoya.

5. After changing trains, you go further west for 2.5 hours to reach a port town. This could be Hiroshima, which is a notable port city in that direction from Nagoya.

6. You take a ferry ride for 4 hours to reach an international terminal. This could be a ferry from Hiroshima to Busan, South Korea, which is known to take roughly 4 hours.

7. You then take a bus to the airport and fly for 10 hours south to a very famous airport. This could be Sydney Kingsford Smith Airport in Sydney, Australia.

8. Changing flights, you fly west for 10.5 hours to another famous airport, likely Dubai International Airport in the United Arab Emirates.

9. Changing flights again, you go 9 hours northwest and land at the nearest airport. This could be London, UK (Heathrow).

10. You take a bus and travel for over 24 hours north going through 4 different countries to reach a port at the last station. This suggests a journey through the UK, France, Belgium, Netherlands, and finally arriving in Denmark (the 5th country).

11. Finally, you take a ferry and cross the gulf to reach the country at my destination. This suggests crossing the Kattegat, a sea area between Denmark and Sweden.

So, based on your detailed journey, you are now in Sweden.

(a) Initial prompt.

Prompt:

I start my journey from the city center of Sapporo, Japan. I drive my car southwards for 1 hour to reach a ferry station. I take a ferry that travels almost 18 hours south to reach its destination. From there I again drive for 1.5 hours to reach a very big city. I park my car in the city and get on a train that goes westward to reach another big city in 2 hours. I change trains and keep going west for another 2.5 hours to reach a port town. From there I take the ferry and ride it for 4 hours to reach an international terminal. I take a bus and go into town to reach the nearest airport. I get on a plane and fly about 10 hours south-east to reach a very famous airport in a small island country. I change flights and fly west for another 10.5 hours to reach another famous airport of a small country. I again change flights and go 9 hours north-west and land at the nearest airport of a country famous for sausages. I take a bus and travel for over 24 hours north-east going through 4 different countries to reach a port at the last station. I take a ferry and cross the gulf to reach the country at my destination. I sit down at a restaurant a eat a bowl of salmon-soup. Where am I now?

GPT-4:

1. You start in Sapporo, Japan and drive south for one hour, likely bringing you to Tomakomai, a city with a major ferry terminal.

2. From Tomakomai, you take an 18-hour ferry ride south which might bring you to the port of Oarai in Ibaraki Prefecture, close to Tokyo.

3. From Oarai, you drive south for about 1.5 hours, which could take you to the center of Tokyo.

4. From Tokyo, you take a train westward for 2 hours which could take you to Shizuoka.

5. Changing trains and continuing west for another 2.5 hours, you could reach Nagoya.

6. A 4-hour ferry from Nagoya could potentially bring you to the port city of Busan, South Korea.

7. From Busan, you take a bus to Gimhae International Airport, then fly southeast for about 10 hours, likely landing at Sydney Kingsford Smith Airport in Sydney, Australia.

8. From Sydney, you take another flight of about 10.5 hours west, which could bring you to Dubai International Airport in the United Arab Emirates.

9. A 9-hour flight northwest from Dubai could land you at Frankfurt Airport in Germany, a country famous for sausages.

10. After a 24-hour bus ride going through 4 different countries to the northeast (potentially through Poland, Lithuania, Latvia, and Estonia), you could reach the port city of Tallinn, Estonia.

11. Crossing the Gulf of Finland by ferry from Tallinn, you would likely end up in Helsinki, Finland.

The final clue is that you eat a bowl of salmon soup. This is a classic Finnish dish, often called "Lohikeitto." So, I would guess that you are in Helsinki, Finland.

(b) Refined prompt.

Figure 23: Responses for the initial and refined prompts for the more complex journey illustrated in Fig. 24. Difference between the prompts are highlighted in orange.

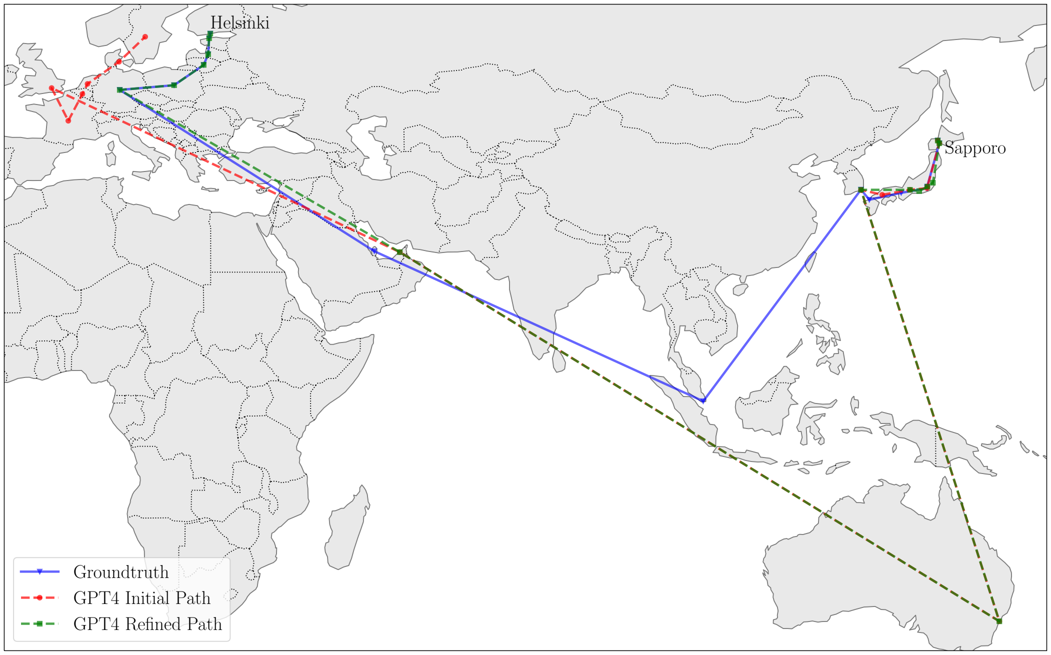


Figure 24: Illustration of the journey and intermediate stops outlined by GPT-4 against the groundtruth. We see that after refining the prompt for more specific information about the possible destinations, GPT-4 was better at identifying the described path.

Prompt:

I am in front of the Chicago Art Museum looking towards Adam's St. I walk 500 ft in that direction and take a left turn. I walk another 900 ft and take a right turn. I keep walking straight and pass 9 blocks till I reach a bridge and take a right turn. I walk another 730 ft and reach a tall skyscraper. Where am I now?

GPT-4:

1. You start in front of the Chicago Art Museum (also known as the Art Institute of Chicago) facing Adams Street.
2. You walk 500 ft east along Adams Street.
3. You take a left turn (north) and walk 900 ft.
4. You take a right turn (east) and walk 9 blocks until you reach a bridge.
5. You take a right turn (south) and walk 730 ft to a tall skyscraper.

From the described journey, it seems you have reached the Willis Tower (formerly known as Sears Tower), which is a prominent skyscraper located at 233 S. Wacker Drive in Chicago. The Willis Tower is one of the tallest buildings in the United States and a major landmark in Chicago.

Figure 25: Response for short distance based navigation in Chicago. The journey is visualized in Fig. 26.

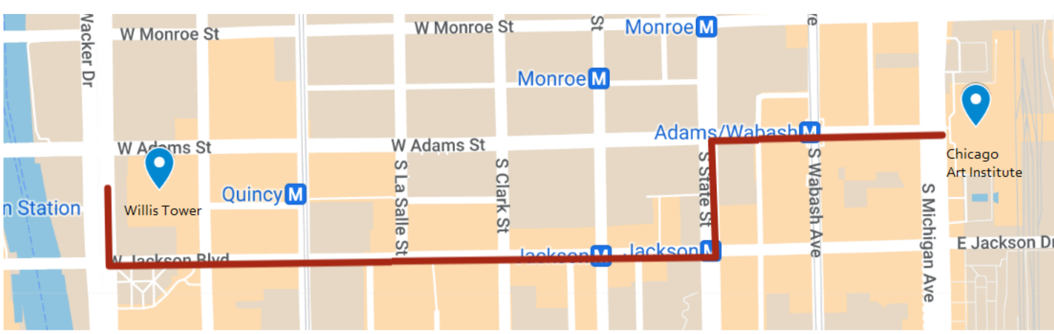


Figure 26: Illustration of the short distance travel in Chicago from *The Chicago Art Institute* to *Willis Tower* described in prompt Fig. 25.

B.11 Networks

Hong Kong MTR Network

We begin by prompting GPT-4 for an ordered list of the stations on each line. We find if we ask for coordinates for the stations without this initial prompt, the model misses 30-40% of the stations on the line when we ask for coordinates.

```
Provide a list of the names of the stations in order on the Hong Kong MTR
<Line_Name> Line.  Give the latitude and longitude coordinates for each of
these as a python list of tuples.  Maintain the same order.
```

Airport Network

We tried to create the network of all airports that are directly connected to *Perth International Airport (PER)*. We first used the following prompt,

```
Give me the list of coordinates for all airports directly connected with
Perth International Airport.
```

However, the network only listed a few locations and responded that the full list of airports are too large for the response. So, we broke down the prompt into two query segments - list of all internal airport, and a list of all external airports. This was prompted as,

```
Give me a list of every airport <inside/outside> Australia that has direct
flights from Perth Airport.  Also provide their lat/lon coordinates.
```

The result when plotted gave the airport network in Fig. 27. GPT-4 was able to identify 33 out of the 40 airports, but also made 12 false predictions.

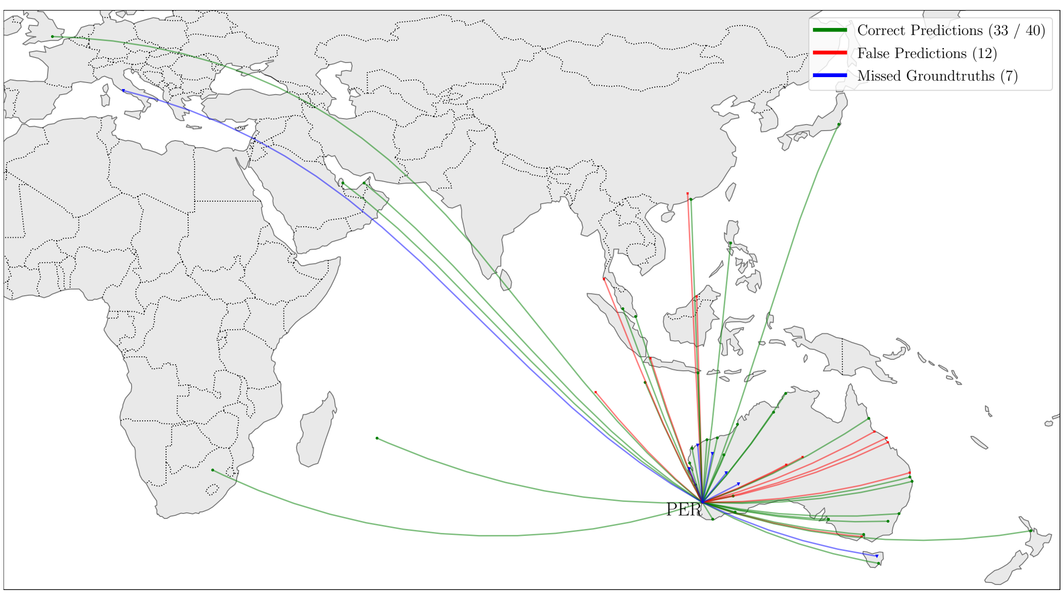


Figure 27: Visualizing the knowledge of GPT-4 about direct flight routes from *Perth International (PER), Australia* to other countries.

Rail Network. We plot the European Railway Network in Fig. 28. We first queried for the pairwise list of countries that are connected with each other, and next the coordinates for a major train station in each country. The two part prompt is as follows,

```
Give me a list of pairwise countries in Europe that have direct railway
connections between them.

Give me the lat/lon coordinates of the train station for each country in
the list
```

Response:

```
Here's a list of pairwise countries with direct railway connections between
them.
1.  Austria - Germany
2.  Austria - Switzerland
3.  Austria - Italy
.
75.  Ukraine - Romania
76.  United Kingdom - France

Here's the list of coordinates for stations.
1.  Austria - Vienna Central Station (Wien Hauptbahnhof):  Latitude:
48.1853, Longitude:  16.3759
2.  Belgium - Brussels Central Station (Bruxelles-Central):  Latitude:
50.8450, Longitude:  4.3571
.
25.  United Kingdom - London St Pancras International:  Latitude:  51.5314,
Longitude:  -0.1262
```

We plotted the coordinates and connected the lines between the pairwise countries. Contrary to the groundtruth, GPT-4 kept Russia as part of Europe. Moreover, we only ask for a single rail station from each country. This removes some major stations like Munich, Frankfurt, and Lyon.

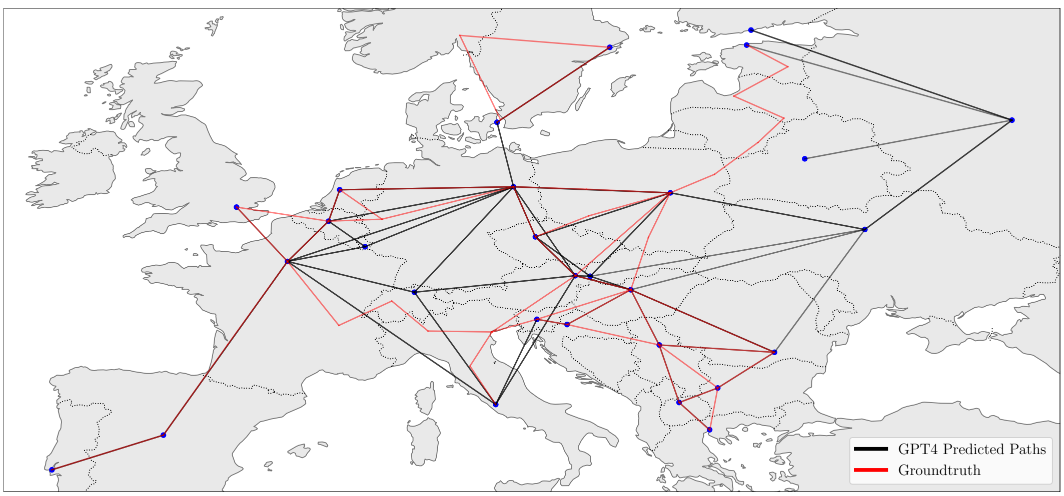


Figure 28: Subset of train network between the major cities of European countries including Russia.

B.12 Itinerary Planning

Itinerary prompts involve specifying the start location, destination location, intermediate waypoints, budget limitations, duration of travel, and any additional constraints. The prompt for the 8 days itinerary in Ireland is as follows:

```
I am currently at my home in Miami, Florida.  I have 1 week vacation and a
budget of 2000 USD. Suggest me a travel itinerary for a trip to Ireland.
Provide day by day step by step detailed itinerary plan for the whole trip
with a breakdown of specific places to visit, foods to try out, as well as
the required time, and money I need.  Provide a breakdown of how to travel to
the destinations and come back home.
```

We also query for a one week vacation combining Singapore and Indonesia, starting from Miami with more specific prompt and constraints. Fig. 29 shows the response which includes detailed guide on travel, food, cost per day, as well as recommendations adhering to the allergy and food constraints. This shows GPT-4's capabilities as a personalized travel agent.

Prompt:

I am currently at my home in Miami, Florida. I have 1 week vacation and I want to visit Singapore and Indonesia for a holiday trip. I have a budget of 3000 USD. Suggest me a day by day step by step detailed itinerary plan for the whole trip with a breakdown of specific places to visit, foods to try out, as well as the required time, and money I need. Provide a breakdown of how to travel to the destinations and come back home. Keep in mind that I am allergic to shellfish.

GPT-4:

```
Day 1: Departure from Miami to Singapore
- Book a one-way flight from Miami (MIA) to Singapore (SIN). The cost will be around $800-$1,000 for an economy
class ticket. The flight will take around 25 hours with a layover.
- Make sure to pack necessary items, including medication for your shellfish allergy.

Day 2: Arrival in Singapore
- Arrive at Changi Airport in Singapore.
- Take a taxi or MRT (Mass Rapid Transit) to your hotel. A budget hotel in Singapore will cost around $50-$70 per
night. Check-in and rest after the long flight.

Day 3: Exploring Singapore
- Morning: Visit Gardens by the Bay and explore the Flower Dome and Cloud Forest. Entry fee is around $20.
- Lunch: Try Hainanese Chicken Rice at a local hawker center. A meal will cost around $5.
- Afternoon: Visit the National Gallery Singapore. Entry fee is around $15.
- Evening: Head to Marina Bay Sands and enjoy the view from the SkyPark Observation Deck. Entry fee is around $20.
- Dinner: Try Laksa (without shellfish) at a local restaurant. A meal will cost around $10.

Day 4: Singapore to Bali, Indonesia
- Morning: Check out of your hotel and take a taxi or MRT to Changi Airport.
- Book a one-way flight from Singapore (SIN) to Bali (DPS). The cost will be around $100-$150 for an economy class
ticket. The flight will take around 3 hours.
- Arrive at Ngurah Rai International Airport in Bali, Indonesia.
- Take a taxi to your hotel in Ubud. A budget hotel in Ubud will cost around $30-$50 per night. Check-in and rest.

Day 5: Exploring Ubud, Bali
- Morning: Visit the Tegalalang Rice Terraces and take a leisurely walk.
- Lunch: Try Nasi Campur (mixed rice) at a local restaurant. A meal will cost around $5.
- Afternoon: Visit the Sacred Monkey Forest Sanctuary. Entry fee is around $3.
- Evening: Explore the Ubud Art Market and purchase souvenirs.
- Dinner: Try Sate Lilit (Balinese satay) at a local restaurant. A meal will cost around $5.

Day 6: Day trip to Nusa Penida, Bali
- Morning: Take a speedboat from Sanur Beach to Nusa Penida. Round-trip tickets will cost around $40.
- Explore Nusa Penida's beautiful beaches, such as Kelingking Beach and Diamond Beach.
- Lunch: Try Gado-Gado (vegetable salad with peanut sauce) at a local restaurant. A meal will cost around $5.
- Afternoon: Visit the famous Broken Beach and Angel's Billabong.
- Evening: Return to Sanur Beach by speedboat and head back to your hotel in Ubud.
- Dinner: Try Babi Guling (Balinese roast pork) at a local restaurant. A meal will cost around $7.

Day 7: Bali to Miami
- Morning: Check out of your hotel and take a taxi to Ngurah Rai International Airport.
- Book a one-way flight from Bali (DPS) to Miami (MIA). The cost will be around $800-$1,000 for an economy class
ticket. The flight will take around 25 hours with a layover.

Total approximate budget: $2,000 - $2,500 (including flights, accommodations, meals, activities, and
transportation). This leaves you with some extra budget for shopping or other miscellaneous expenses.

Note: Be sure to inform the staff at restaurants about your shellfish allergy to avoid any unwanted reactions.
```

Figure 29: GPT-4 response for a 1 week itinerary for a trip to Singapore and Indonesia.

B.13 Abstract Routing

Every node

We used the following prompt for finding the optimum route that visits every edge in the graph.

```
I am visiting Hong Kong for a holiday.  There are 8 tourist attractions I'd
like to see, all are within walking distance.  The attractions are labelled
A,B,C,E,F,G,H,I. For each attraction I will state the other attractions its
directly linked to and how long it takes to walk between the two.  My hotel
is at point D. I will be starting and finishing from this point.
A-B = 5 A-C = 1 B-C = 8 B-E = 3 B-F = 4 D-E = 2 D-G = 14 E-G = 9 F-G = 1 G-H
= 4 G-I = 2 I-H = 1
Note, each of these pathways is bidirectional, and it takes the same time to
walk in each direction.
What is the minimum amount of time I need to spend walking to visit all the
attractions?  What is my optimal route, starting and finishing at my hotel.
```

Every edge

We used the following prompt for finding the optimum route that visits every node in the graph.

```
I am visiting Hong Kong for a holiday.  There are 8 tourist attractions I'd
like to see, all are within walking distance.  The attractions are labelled
A,B,C,E,F,G,H,I. For each attraction I will state the other attractions its
directly linked to (via a walking path) and how long it takes to walk between
them.  My hotel is at point D. I will be starting and finishing from this
point.  I also want to walk along all of the walking paths available.
A-B = 5 A-C = 1 B-C = 8 B-E = 3 B-F = 4 D-E = 2 D-G = 14 E-G = 9 F-G = 1 G-H
= 4 G-I = 2 I-H = 1
Note, each of these pathways is bidirectional, and it takes the same time to
walk in each direction.
What is the minimum amount of time I need to spend walking to ensure that I
visit all the attractions and walk along all of the possible paths?  What is
my optimal route, starting and finishing at my hotel.
```

Graph Description We verify that GPT-4 correctly interprets the description of the graph by asking for the adjacency matrix from the descriptions above:

Prompt:

```
Can you give me the adjacency matrix for the graph
```

Response:

```
   A B C D E F G H I
A 0 5 1 -1 -1 -1 -1 -1 -1
B 5 0 8 -1 3 4 -1 -1 -1
C 1 8 0 -1 -1 -1 -1 -1 -1
D -1 -1 -1 0 2 -1 14 -1 -1
E -1 3 -1 2 0 -1 9 -1 -1
F -1 4 -1 -1 -1 0 1 -1 -1
G -1 -1 -1 14 9 1 0 4 2
H -1 -1 -1 -1 -1 -1 4 0 1
I -1 -1 -1 -1 -1 -1 2 1 0
```

GPT-4 also correctly returns the adjacency list:

Prompt:

```
Convert this into an adjacency list
```

Response:

```
An adjacency list representation of the given adjacency matrix is as follows:
A: [(B, 5), (C, 1)]
B: [(A, 5), (C, 8), (E, 3), (F, 4)]
C: [(A, 1), (B, 8)]
D: [(E, 2), (G, 14)]
E: [(B, 3), (D, 2), (G, 9)]
F: [(B, 4), (G, 1)]
G: [(D, 14), (E, 9), (F, 1), (H, 4), (I, 2)]
H: [(G, 4), (I, 1)]
I: [(G, 2), (H, 1)]
```

Alternative wording for Riddle

We reword the riddle used in the main paper, altering the context to provide a second assessment of GPT-4's ability to answer the riddle correctly with less chance of the exact question/answer being seen during training.

Prompt:

```
You're driving on the surface of the Earth.  After driving one mile south,
one mile west and one mile north, you end up exactly where you started.
Where are you?
```

Response:

```
There are actually multiple locations from which you could start and end up
at the same place after driving one mile south, one mile west, and one mile
north.
The North Pole:  If you start a bit more than one mile south of the North
Pole, you can drive one mile south to reach the North Pole, then drive one
mile west (which would actually just have you moving in a circle around the
North Pole, because all directions point south from there), and then drive
one mile north to return to your original starting point.
Certain points near the South Pole:  This is more complex.  Imagine a circle
around the South Pole that has a circumference of 1 mile.  If you start 1
mile north of this circle, you can drive 1 mile south to reach the circle,
then drive 1 mile west, which will take you all the way around the circle and
back to the same point (because the circumference is 1 mile), and then drive
1 mile north to return to your original starting point.  This same principle
applies to any circle near the South Pole that has a circumference less than
or equal to 1 mile.  For those smaller circles, driving ‘‘west" for 1 mile
would involve making one or more complete laps around the circle.
Note that these answers are based on an idealized, spherical model of the
Earth.  The real world is more complicated due to the Earth's oblate shape,
and the presence of continents and other geographical features.
```

B.14 Landmarks

We use two different prompts to generate SVG code for the outline of landmarks.

Isolated Landmarks

For landmarks that are isolated and far from other landmarks, we used these prompts:

```
What famous landmark is at this location:  <Landmark_Lat/Lon_Coordinates>?
Provide the SVG code for a detailed outline of the landmark.
```

and

```
What famous landmark is at this location:  <Landmark_Lat/Lon_Coordinates>?
Provide the SVG code for a simple outline of the landmark.
```

Where <*Landmark_Lat/Lon_Coordinates*> are the coordinates for the Statue of Liberty and Giza Pyramids Complex. Coordinates were taken from Wikipedia and Google Maps. For these landmarks, we find GPT-4 is able to accurately identify the landmark from the coordinates and provide the SVG code for a recognisable outline.

Dense Landmarks

For landmarks that are located closely with others, we specify the landmark name in the prompt rather than providing a location. For example, "London skyline" is not possible to specify using a single coordinate, and the Gardens by the Bay are located too close to the Marina Bay Sands hotel for a single coordinate to distinguish the two.

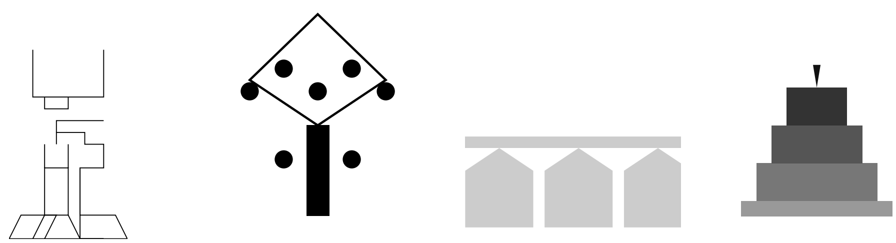
For the London Skyline, we used this prompt:

```
Provide the code for an SVG of the detailed outline of the London skyline.
Use colour where relevant.  Ensure that it's clearly recognisable as London
from the outline.  Include the most famous buildings.
```

For the Gardens by the Bay, we used this prompt:

```
Provide the code for an SVG outline for the Gardens by the Bay in Singapore.
```

We include GPT-4's SVG outlines for additional landmarks that we experimented with in Fig. 30.



(a) Eiffel Tower. (b) St. Basil's Cathedral. (c) Marina Bay Sands. (d) Temple of Heaven.

Figure 30: **Additional Landmark SVGs**. (Countries (a)-(d): France, Russia, Singapore, China).

B.15 Multi-criteria Place Retrieval

We used the following prompt for our experiment on multi-criteria place retrieval:

```
Name all places in the world where skiing and surfing is possible at the same
day in December?  Provide a python list in the format [0.00000N, 0.00000E].
```

```
Name all places in the world where skiing and taking a bath in natural
hot springs is possible.  Provide a python list in the format [0.00000N,
0.00000E].
```

```
Name all places in the world where safe hiking above 3000m is possible in
December.  Provide a python list in the format [0.00000N, 0.00000E].
```

```
Name all places in the world close to a volcano and a city with a population
above 1 million.  Provide a python list in the format [0.00000N, 0.00000E].
```

```
Name all places in the world that lie in a country that has won FIFA World
Cup and where a large cat be seen in the wild.  Provide a python list in the
format [0.00000N, 0.00000E].
```

```
Name all places in the world where whales can be watched and a coastal hike
is possible.  Provide a python list in the format [0.00000N, 0.00000E].
```

We found that asking for a specific coordinate format helped as GPT-4 would sometimes generate inconsistent coordinates otherwise. If the output was not only provided as a Python list, we extracted the coordinates manually.

B.16 Supply Chains

We used the following prompt for analysing the global semiconductor supply chain:

```
I want to construct a map of the semiconductor supply chain, end-to-end.
Please provide the latitude and longitude coordinates and names of the key
elements in the supply chain, including design, manufacturing, materials,
equipment + tools, etc.  If you don't know any coordinates exactly just
estimate, every point needs coordinates.
```

We then ask for assistance with plotting:

```
Provide a python script that uses the Cartopy library to plot these locations
on a world map.  Use different colours for the steps 1-6.  Annotate each
location with the company/industry located there.
```

B.17 Natural World

Wildlife Ranges

We start with the initial prompt:

```
What are the different subspecies of tiger?
```

We then repeatedly use the following prompt, replacing *tiger subspecies* with each subspecies returned in the first answer:

```
I want to create an outline of the range of the tiger subspecies that I can
overlay on a world map that I have.  Provide me with latitude and longitude
coordinates for this approximate outline.  Output the coordinates as a python
list of tuples.  Due to length constraints, just output the list.  Ensure the
outline does not cross over itself.  If the range is not continuous, provide
multiple outlines as separate lists.
```

Bird Migrations

We used the following prompt for evaluating GPT-4's knowledge of bird migrations:

```
I want to plot the migratory routes of various bird species.  Please
provide the latitude and longitude coordinates of the migratory route of
the bird species.  Indicate the start and finish and provide coordinates for
intermediate steps if necessary.  If multiple routes exist, provide separate
lists of coordinates.
Output the coordinates as a python list of tuples.
```