Using large language models to address the bottleneck of georeferencing natural history collections

Yuyang Xie[1], Daniel S. Park[2, 3], Miranda A. Sinnott-Armstrong[2, 3], Joyce Ho[4], Tianlong Chen[5], Alan S. Weakley[1, 6], Luis J. Aguirre Lopez[1], Jaein Choi[2, 3], Marisa Laitinen[2, 3], Nicholas Steeves[2], Chingyan Huang[2], Ran Xu[4], Xiao Feng[1]*

1. Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA
2. Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
3. Purdue Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA
4. Department of Computer Science, Emory University, Atlanta, GA 30322, USA
5. Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA
6. University of North Carolina Herbarium (NCU), North Carolina Botanical Garden, University of North Carolina, Chapel Hill, NC 27599, USA

* Corresponding author: fengxiao.sci@gmail.com

Abstract

Natural history collections are fundamental for biodiversity research. The broad use of them relies on the digitization effort, especially georeferencing that translates textual locality descriptions into geographic coordinates. However, traditional georeferencing approaches are labor-intensive and costly, thus georeferencing is a major bottleneck in the digitization process that prevents the usage of millions of specimens across the world. This study investigated the potential of using large language models (LLMs) to facilitate georeferencing. We utilized LLMs from OpenAI and DeepSeek to georeference 5,000 vascular plant specimen records with known coordinates, and compared the results against those of GEOLocate (a widely used georeferencing tool) and manual georeferencing. We found that the best-performing LLMs (e.g., gpt-4o) outperformed specialized tools like GEOLocate in spatial applicability, and demonstrated near-human-level accuracy with a median georeferencing error of <10 km. Georeferencing based on LLMs were also considerably fast (<1 s per record) and affordable ($0.10 per 100 records); thus, they present a cost-effective approach for georeferencing. LLMs may not fully replace human curation in the short term, but can be incorporated into current workflows to greatly increase the efficiency of georeferencing. Future advances in LLMs may revolutionize the digitization of natural history collections.

Keywords:

Artificial Intelligence, Large Language Model, Biodiversity, Herbarium, Museum, Specimen

Introduction

Natural history collections form the foundation of our knowledge of biodiversity. They represent irreplaceable snapshots of biodiversity across space and time critical for ecological and evolutionary research [1-3]. The specimens that make up these collections offer valuable insights into ecosystem dynamics by documenting habitat preferences [4], species interactions [5]; temporal responses to climate change [6, 7]; revealing evolutionary relationships [8, 9]; prioritizing geographic areas with concentrations of rare and imperiled species for conservation focus [10]; and providing historical baselines for tracking environmental change [11]. It is estimated that the total number of specimens in natural history collections ranges between 2 and 3 billion [12]. Massive digitization efforts have greatly increased accessibility to these specimens and facilitated innovative, large-scale research. However, only a small portion of these natural history collections have been digitized. For instance, it has been estimated that less than 30% of herbarium specimens have at least collection location and date information online [13, 14].

Specimen digitization involves converting the information within physical specimens into digital formats, encompassing textual, visual, temporal, and geographic information, among other data types [15]. Georeferencing is one of the outstanding challenges of the digitization process [16]. Georeferencing interprets a specimen's textual locality description, including directional cues, man-made landmarks, or references to roads, into a set of geographic coordinates [17, 18]. This associates the occurrence of an organism to a point in space, enabling a suite of ecological inquiries, such as inferring the environmental requirements of a species or ecological patterns of species co-occurrences [19, 20]. Currently, georeferencing is done largely manually, and is a labor-intensive and costly (and therefore slow) process. As a consequence, the vast majority of collections still remain non-georeferenced [16]. While recent collections are often geotagged using GPS units, specimens collected before GPS units were widely available (i.e., before the 1990s) often require georeferencing to link the specimen to a point on a map.

Traditional georeferencing methods include using gazetteer-based applications or manually searching for locations with maps. For example, GEOLocate is a georeferencing software developed 20 years ago that is still commonly used by museums [16, 21]. GEOLocate converts textual locality descriptions from specimens into geographic coordinates by standardizing terms and extracting distances, directions, and key geographic identifiers [21]. GEOLocate can batch-process locality descriptions but is not fully automated even in batch mode. As a consequence, manual georeferencing (e.g., looking up a location in Google Maps) remains a time-consuming, and therefore costly, necessity, and additional funds for corrections and quality control are usually needed [22, 23].

Recent breakthroughs in Large Language Models (LLMs) have great potential to address this critical bottleneck [24]. LLMs are large-scale natural language processing models trained through deep learning to read, understand, and generate text, and are widely applied in various language tasks [25]. LLMs demonstrate great potential in text mining capabilities, which may revolutionize a variety of ecological studies [26, 27], such as extracting species distributions and richness [28, 29], as well as listing endangered species and classifying the threats from unstructured text to support biodiversity conservation [30, 31]. Previous studies have examined the utility of LLMs for geospatial reasoning [24], such as geographic entity classification and directional inference [32-34]. However, their potential to infer geographic coordinates based on textual locality descriptions remains unexplored [24].

Here, we present the first benchmark of georeferencing using LLMs. We compared the accuracy and efficiency of LLMs with approaches commonly used in georeferencing practices, including manual georeferencing, GEOLocate, and county-centroid (directly using the centroid of the county where the specimen is located as the geographic reference coordinates). The experiment was based on 5,000 specimen records collected across the globe that have locality descriptions and geographic coordinates. We included gpt-4o (ChatGPT model version 4o) and deepseek-chat (DeepSeek model version 3), as well as earlier versions of GPT models, including gpt-4o-mini, gpt-4-turbo, and gpt-3.5-turbo. We also included LLMs with enhanced reasoning capabilities, including o1-preview (advanced OpenAI reasoning model version o1 preview) and deepseek-reasoner (DeepSeek reasoning model version R1). We also investigated whether georeferencing accuracy can be affected by geographic factors and textual features of locality descriptions.

Results

Overall accuracy of georeferencing

Georeferencing by LLMs achieved human-like accuracy (Fig. 1). Among all non-reasoning LLMs examined, *gpt-4o* and *deepseek-chat* demonstrated the highest accuracy in georeferencing 4,750 specimen samples (top 5 percentile outliers of each georeferencing method were excluded to avoid extreme cases; Fig. 1, Extended Table 1), with median error distances of 9.7 and 12.3 km, respectively. A Wilcoxon test indicated that the accuracies of *gpt-4o* and *deepseek-chat* did not differ from that of manual georeferencing (p>0.05, N=95, top 5 percentile outliers of 100 sampled records from these 5,000 entries were excluded) and significantly outperformed (p<0.05, N=4,750) the accuracy of GEOLocate (23.4 km median) and the "county-centroid" method (18.2 km median), a common practice in which the centroid coordinates of the county or equivalent geopolitical locality of collection are assigned to a specimen [35, 36]. In contrast, simpler or earlier versions of LLMs like *gpt-4o-mini* and *gpt-3.5-turbo* exhibited relatively lower accuracy, performing even worse than the county-centroid method (p<0.05, N=4750). Compared with *gpt-4o* and *deepseek-chat*, the use of advanced reasoning models (*o1-preview and deepseek-reasoner*) did not lead to a significant improvement in georeferencing accuracy (p>0.05, N = 95) (Fig. 1, Extended Fig. 1), despite the higher costs and increased processing times (Extended Table 1).

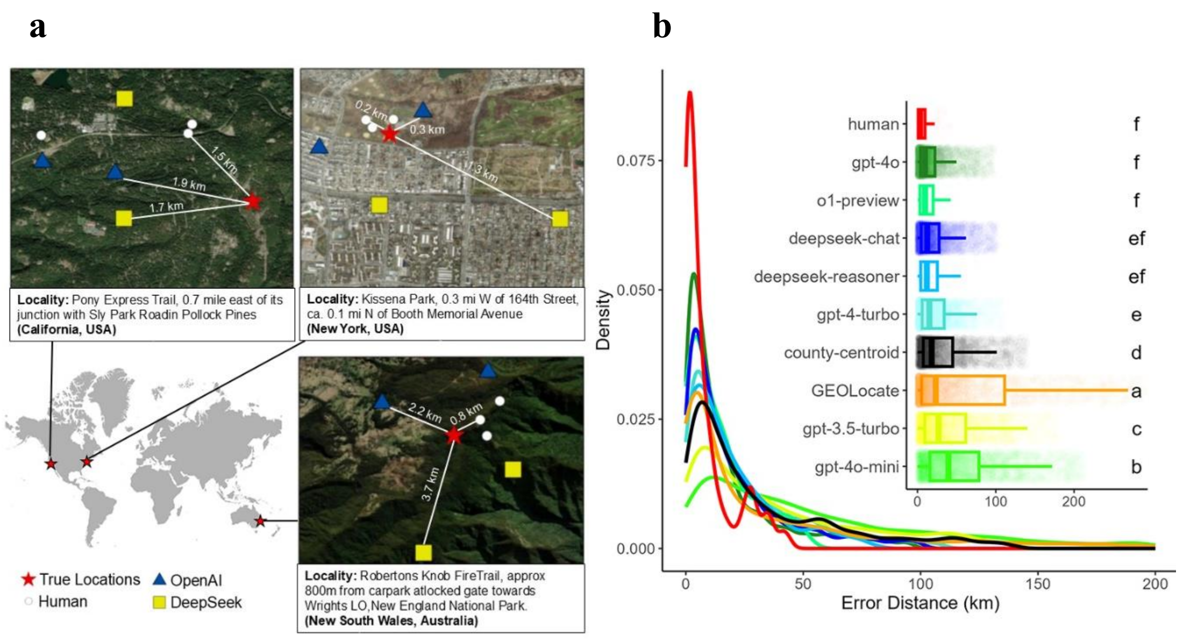


Figure 1. Summary of georeferencing using different methods. a. Three examples of georeferencing results based on humans and large language models. **b.** Density plot and boxplot of georeferencing accuracy of different methods. A higher georeferencing accuracy (x-axis) is represented by a smaller error distance (distance to ground truth coordinate). The letters in the boxplot indicate intergroup differences according to the Wilcoxon test, where letters appearing later in the sequence correspond to smaller mean georeferencing errors. Identical letters signify no significant differences between methods (p > 0.05). In the boxplot, the models or methods are ordered from top to bottom based on the ascending median georeferencing error. The outliers (i.e., the top 5 percentile) of each georeferencing method were excluded from the analysis to minimize the effect of large errors and

emphasize each method's usual performance. Thus, the sample size is 95 for o1-preview, deepseek-reasoner, and human georeferencing, and 4750 for all other georeferencing methods. The statistics for a sample size of 95 across all methods are shown in Extended Figure 1 & Table 1.

Georeferencing accuracy also showed considerable spatial variation across countries (Figs. 2 & 3). LLMs achieved higher median accuracy for specimen localities in the United States, Western Europe, Southern Africa, Southeast Asia, and Australia, and the median error distance was mostly within 5km for the best-performing LLMs (e.g., gpt-4o) (Fig. 2a-d, 3a-d, Extended Fig. 2, Supplementary Information Tables S1, 2). Compared with LLMs, GEOLocate showed a more distinct contrast between better-performing regions (United States, Western Europe) vs. other regions, and the error distance could exceed 1,000km for specimen localities in Russia (Supplementary Information Table S1). The county-centroid approach showed higher accuracy (lower error distance) for smaller-sized countries, likely because of smaller county size therein (Extended Fig. 3). The manually georeferenced results exhibited smaller spatial variation (Fig. 2f, 3f). We also noticed spatial variation in georeferencing accuracy within countries/regions. For example, higher accuracies were concentrated on the east and west coasts of the United States and the coastlines of Australia, while the accuracies were lower for the Andes Mountains, Rocky Mountain Region of the Western United States, and Central Australia (Figs. 2 & 3).

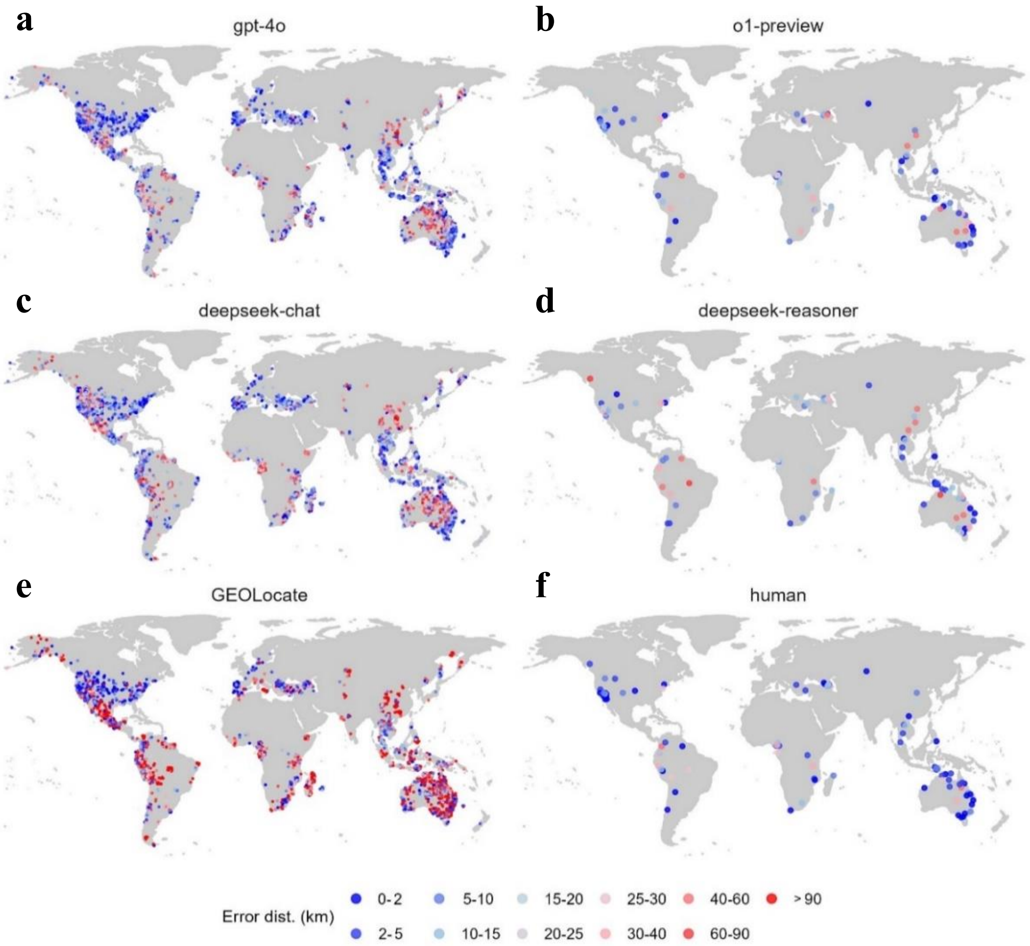


Figure 2. The geographic distribution of georeferencing accuracy. Georeferencing accuracy is represented by the error distance (distance between georeferenced coordinates and ground truth), and smaller values indicate higher accuracy. Maps represent different georeferencing methods: *gpt-4o* (**a**, 5000 samples), *o1-preview* (**b**, 100 samples), *deepseek-chat* (**c**, 5000 samples), *deepseek-reasoner* (**d**, 100 samples), GEOLocate (**e**, 5000 samples) and manual georeferencing (**f**, 100 samples). See Extended Figure 2 for results of other OpenAI LLM and county centroid-based method.

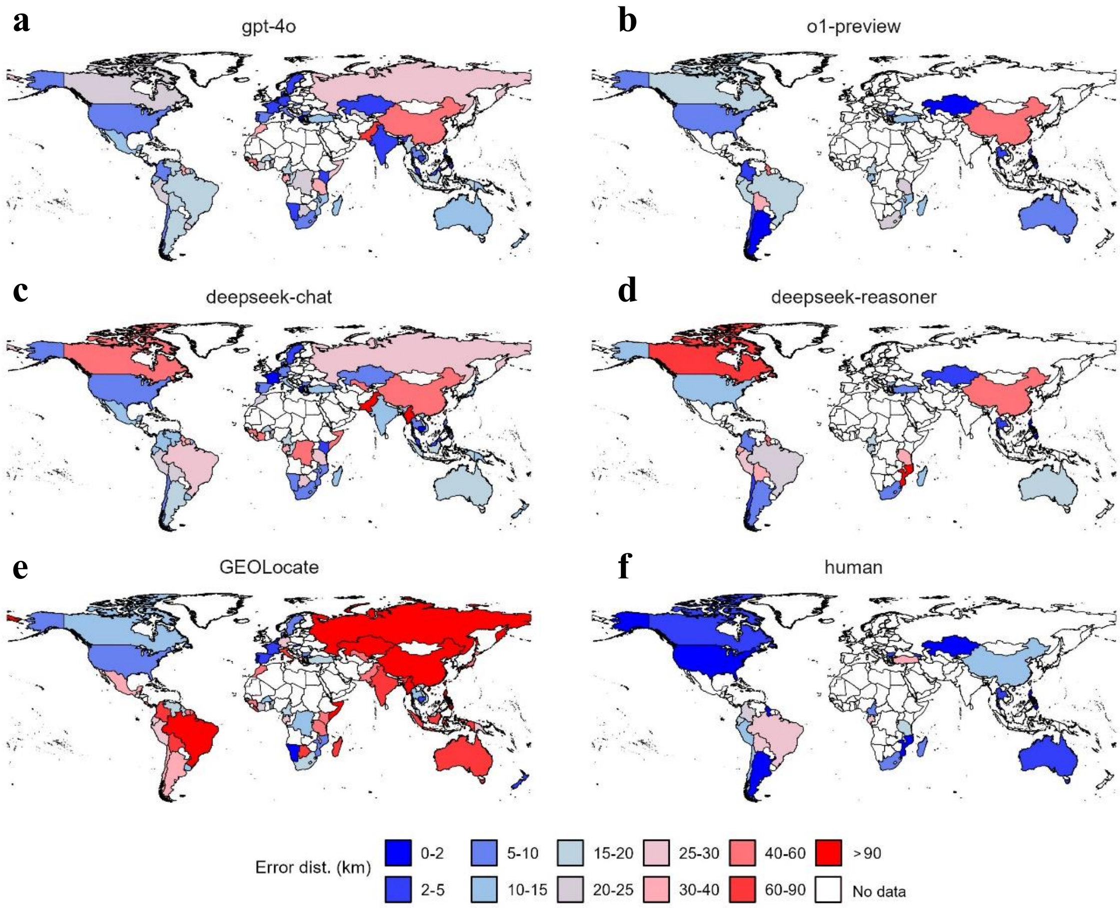


Figure 3. The georeferencing accuracy at country level. Georeferencing accuracy is represented by the median error distance (distance between georeferenced coordinates and ground truth) of all sample points in each country, and smaller values indicate higher accuracy. Maps represent different georeferencing methods: *gpt-4o* (**a**, country-level statistics based on 5000 samples), *o1-preview* (**b**, country-level statistics based on 100 samples), *deepseek-chat* (**c**, country-level statistics based on 5000 samples), *deepseek-reasoner* (**d**, country-level statistics based on 100 samples), GEOLocate (**e**, 5000 samples) and manual georeferencing (**f**, country-level statistics based on 100 samples). See Extended Figure 3 for country-level statistics of other OpenAI LLMs and county centroid-based methods.

Trade-offs between georeferencing accuracy and time-monetary cost

Most OpenAI models were able to georeference one specimen record within one second, though *gpt-4o* requires a 1-second interval between API calls. The *deepseek-chat* model was slightly slower, requiring ~2 seconds per record, but continuous API calls are allowed, and it is currently free. The *deepseek-reasoner* and *o1-preview* models used complex reasoning, resulting in longer processing times (Fig. 2, Extended Table 1). In particular, *deepseek-reasoner* took an average of 92 seconds (sd = 68s; Extended Table 1) to georeference one record, sometimes even exceeding the time it took for manual georeferencing. The *o1-preview* model had the highest monetary cost among all LLMs, averaging over $13 per 100 queries (Extended Table 1).

Non-reasoning models like *gpt-4o* and *deepseek-chat* achieved high efficiency at low costs. The goodness of fit (R^2) for models between the median georeferencing error and processing time was 0.41 for the linear model and 0.45 for the exponential model (Fig. 4a); for median georeferencing error and monetary cost, the goodness of fit was 0.48 for the linear model and 0.49 for the exponential model (Fig. 4b).

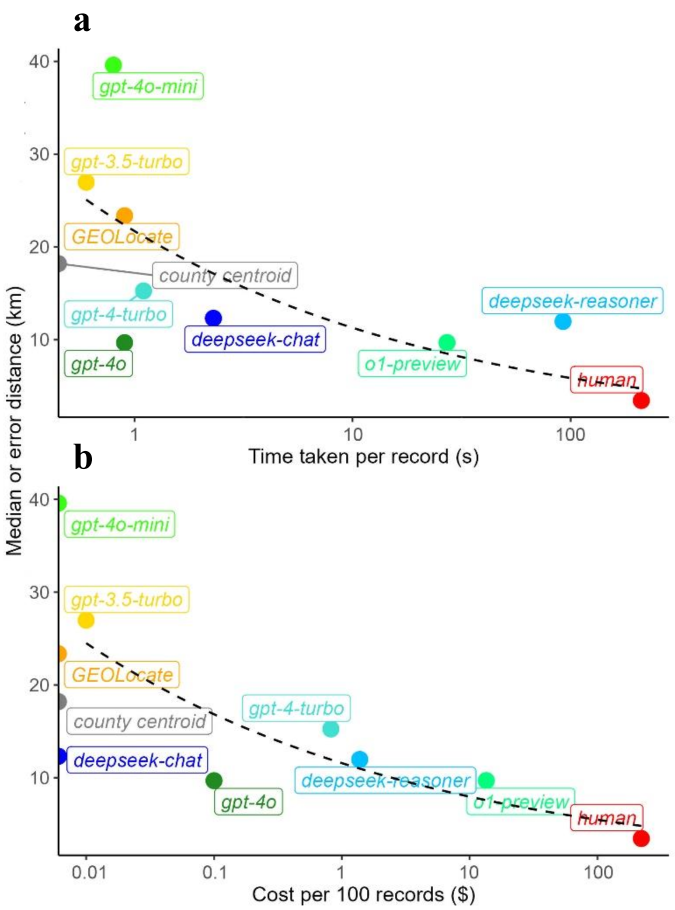


Figure 4. The relationship between the georeferencing accuracy and (a) efficiency or (b) monetary cost among different georeferencing methods. Georeferencing accuracy is represented by the median of error distance (distance between georeferenced coordinates and ground truth) of each method (y-axis), and smaller values indicate higher accuracy. The black dashed line in the figure represents the fit exponential curve, illustrating the power-law relationship between the georeferencing accuracy and the associated costs. The x-axis is log-10 transformed.

Factors influencing the performance of LLM georeferencing

The accuracy of all georeferencing methods showed a gradual increase (i.e., decrease in error distance) with the increase of the human footprint index of the locality where specimens were collected (Fig. 5). The increase in accuracy was more pronounced for GEOLocate and simpler/earlier versions of OpenAI LLMs, such as *gpt-4o-mini* and *gpt-3.5-turbo* (Fig. 5a).

We used Flesch-Kincaid Grade Level (FKGL) [37], Gunning Fog Index (GFI) [38], Simple Measure of Gobbledygook (SMOG) [39], Coleman-Liau Index (CLI) [40] and Automated Readability Index (ARI) [41] to quantify readability of locality description text (see Extended Table 2 for definitions and calculations). To measure the level of detail in the locality description texts, we also calculated the frequency of keywords, including digital characters, punctuation marks, directional terms, and distance units, as well as the frequency of road names and artificial and natural objects (Extended Table 2). The correlations among all text features, including frequency of keywords and readability indexes, were moderate or weak (i.e., $|r|<0.65$; Extended Fig. 4). Overall, the frequency of selected keywords had stronger effects than readability on georeferencing accuracy (Fig. 5b). GEOLocate was more strongly affected by the frequency of selected keywords, especially about artificial objects such as buildings, than LLMs (Fig. 5b).

We found the readability of locality texts had minimal or insignificant impact on the accuracy of georeferencing methods (Fig. 5b). FKGL had a significant positive effect on human georeferencing error, and GFI negatively affected the GEOLocate error. SMOG showed a negative effect on georeferencing error of the older version of the OpenAI model *gpt-4o-turbo* and *gpt-3.5-turbo*, and CLI positively affected both errors of *o1-preview* and human georeferencing ($p<0.05$, Fig. 5b).

Number of digits (digits) had no significant effect on georeferencing accuracy for any method ($p>0.05$, Fig. 5b), while the number of punctuation marks negatively affected ($p<0.05$) georeferencing error of all OpenAI and DeepSeek non-reasoning LLMs. Each additional punctuation mark reduced the error by 0.6, 2.2, 0.5, 0.6 and 0.6 km, respectively, for *gpt-4o, gpt-4o-mini gpt-4o-turbo, gpt-3.5-turbo* and *deepseek-chat* (Fig. 5b). The directional indicator was associated negatively with the error of *gpt-4o* and *gpt-4o-mini* ($p<0.05$), and positively with the error of *gpt-3.5-turbo* (Fig. 5b). Each additional directional indicator reduced the error by 0.6 and 1.4 km, respectively, for *gpt-4o* and *gpt-4o-mini*, while increasing the error by 1.7 km for *gpt-3.5-turbo* (Fig. 5b). The frequency of distance indicators was positively associated with the error of *gpt-4o* and *gpt-4o-mini* assessed ($p<0.05$, Fig. 5b), and the increased error distance by having an additional distance indicator ranged from 1.5 to 3.4 km, while the error of GEOLocate would significantly ($p<0.05$) decrease by 14.4 km with an additional distance indicator (Fig. 5b).

The frequency of keywords related to roads, natural objects (e.g., mountain, river, and canyon), and artificial objects (e.g., building, bridge, and dam) was mostly negatively associated with the error distance of most LLMs and GEOLocate ($p < 0.05$, Fig. 5b). Each additional keyword related to roads or artificial objects reduced the error by 1.1 and 5.8 km for all OpenAI and DeepSeek non-reasoning models, while reducing the error by 12.8 and 44.2 km for GEOLocate (Fig. 5b). Additional keywords pertaining to natural objects did not ($p>0.05$) reduce the error of newer versions of LLMs (*gpt-4o* and *deepseek-chat*), but did significantly increase ($p<0.05$) the error of *o1-preview* by 4.4 km and reduce ($p<0.05$) the error of GEOLocate by 22.4 km (Fig. 5b).

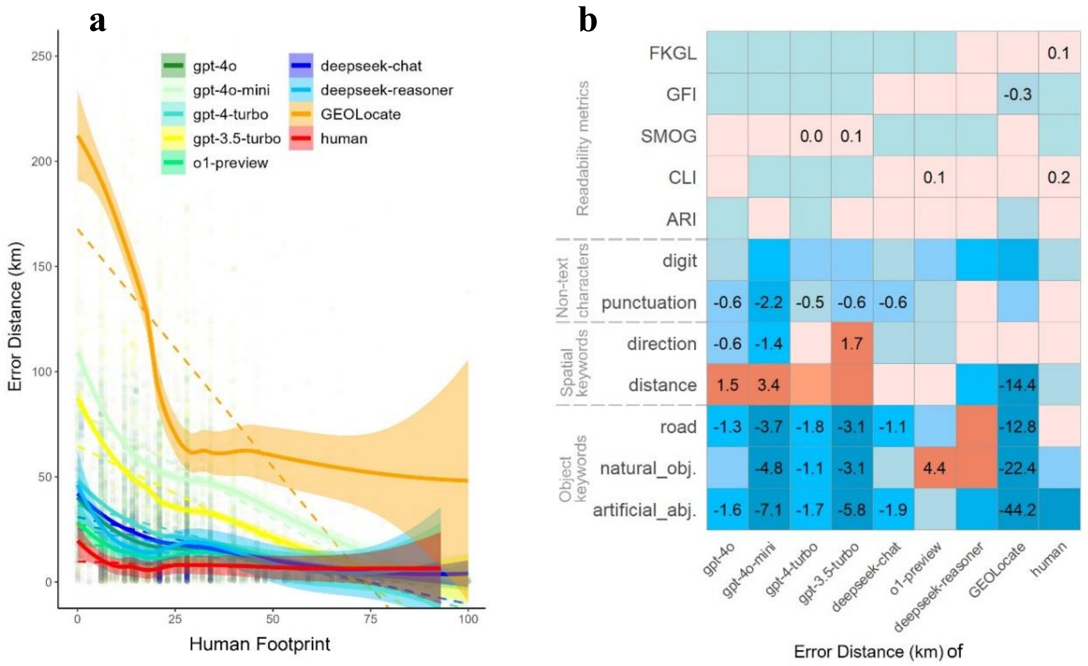


Figure 5. The impact of human activities at specimen collection sites and the textual description characteristics of locality on the georeferencing errors of various methods. a. Locally estimated scatterplot smoothing (LOESS) curves (solid lines) and linear fits (dashed lines) for relationships between human activity intensity measured by the Human Footprint (HFP) index and the georeferencing errors. **b.** Linear regression analysis of the impact of readability metrics and the counts of different word types on the georeferencing error distances across methods. Numbers and colors

indicate the values of linear regression coefficients, with only significant results ($p < 0.05$) labeled with numbers. The independent variables here are not standardized, so the regression coefficients indicate how much the error distance increases in kilometers for each one-unit increase in the independent variables. The full names and meanings of the variable abbreviations are provided in Extended Table 2. As the outliers (i.e., top 5 percentile) of each georeferencing method were excluded from the analysis, regressions for all non-reasoning LLMs (*gpt-4o, gpt-4o-mini, gpt-4-turbo, gpt-3.5-turbo* and *deepseek-chat*) and GEOLocate were based on all the 4750 samples, while regressions for reasoning LLMs and human georeferencing were based on selected 95 samples.

Discussion

Accurate and efficient georeferencing is a major challenge for the broad application of natural history collections. Recent breakthroughs in LLMs that can analyze and generate human-like language can potentially address the challenges in georeferencing in a time- and cost-efficient manner, thus rapidly generating coordinates for large numbers of specimens that remain un-georeferenced. Here, we conducted the first benchmark of LLM georeferencing performances and compared them to existing approaches. We found that the best-performing LLMs (e.g., *gpt-4o*) achieved a median georeferencing error below 10 km that was not significantly different from human georeferencing (Wilcoxon test, $p>0.05$, $N=100$; Extended Table 1 and Extended Fig. 1), and was significantly better than GEOLocate and the county-centroid approach (Wilcoxon test, $p<0.05$, $N=4750$; Extended Table 1 and Fig. 1). Further, LLMs were considerably faster (<1s per record) and less expensive ($0.1 per 100 records) than manual georeferencing; in particular, model *gpt-4o* and *deepseek-chat* achieved the best tradeoff between accuracy and cost (Fig. 2). Compared with *o1-preview* and *deepseek-reasoner*, the more recently released reasoning LLMs showed similar accuracy but took longer to perform georeferencing [42]. However, the reasoning logics used by LLMs were very similar to the inferences used by humans during georeferencing (Supplementary Information Table S3). Given the advances in prompt engineering [43] and model fine-tuning [44] that can further enhance LLMs' performance in specialized domains, LLMs show huge potential in adequately and rapidly georeferencing the remaining millions of natural history collections.

LLMs demonstrate near-human levels of georeferencing accuracy

The median georeferencing error distance of the best-performing LLMs is ~10km, which is on par with the median error distance for manual georeferencing. Many macroecological and biogeographic studies rely on spatial datasets with a resolution of 10 km or more, making LLM georeferenced coordinates sufficiently accurate for a broad range of ecological studies. Ecological niche modeling is one notable example, where georeferenced coordinates from specimens are often overlaid with environmental data at relatively coarse resolution to study species' ecological niches and geographic distributions [45, 46]. WorldClim [47] is one of the most commonly used climatic datasets available at 30", 2.5', and 10' resolution (approximately 1km, 9 km, and 18km at the equator). Similarly, ERA5-Land atmospheric reanalysis products (10 km resolution) [48] and CRUTS (Climatic Research Unit Time-Series, 0.5-degree resolution or 50 km at the equator) [49] have been used in studying climate-induced faunal changes [50], and are broadly used in environmental and atmospheric science [51]. Despite the increase in availability of fine-resolution environmental data, coarse resolutions, such as 10km or above, are still preferred for a variety of practical (e.g., limited computation power), methodological (e.g., unifying different datasets to a coarse resolution), and theoretical (e.g., the effect of climate at large spatial scales) reasons [52-54].

LLMs offer a balance between georeferencing accuracy and efficiency

Performing georeferencing with LLMs can be considerably faster (<1s per record) than manual georeferencing, and more affordable ($0.1 per 100 records). While humans achieved the lowest absolute median error distance in georeferencing, the time spent by humans was over a hundred times higher than the best-performing LLM (*gpt-4o*, Fig. 4a and Extended

Table 1). Still, our estimated time of manual georeferencing is likely an underestimation, because humans rarely continuously perform georeferencing. In reality, human performance often deteriorates with time spent on tasks and people require rest [55]. The inefficiency of manual georeferencing is indeed one of the major bottlenecks in georeferencing faced by many museums and herbaria [16, 56]. In practice, people often set a time limit for the georeference of one record to avoid long inquiries. Mast *et al.* [57] used 15 minutes as a limit in a georeferencing project; similarly, we have used 15 minutes in our experiment.

Compared to humans, the processing time of automated georeferencing methods can be considered almost instantaneous. Since the locality descriptions are typically short text strings, preprocessing time is minimal. The limiting factor is the response time from the GEOLocate and LLM servers that return the georeferencing output. Typically, the speed or total number of queries to a server is limited. Both OpenAI and DeepSeek impose rate limits on their APIs to manage usage and maintain service reliability [58, 59]. Despite such limits, it is still technically feasible to parallelize georeferencing to more instances, thus further speeding up the process to another magnitude.

By looking at all georeferencing methods together, we found a negative relationship between georeferencing efficiency and accuracy (median error distance) (Fig. 4). In other words, spending more time can lead to smaller georeferencing errors. Simpler LLMs or GEOLocate fell in the fastest but least accurate category, while humans fell in the slowest but most accurate (smallest median error) category. The *gpt-4o* and *deepseek-chat* models fell in the middle of the two extremes, achieving a balance between efficiency and accuracy. Also, *gpt-4o* and *deepseek-chat* both fell below the fitted curve (model fitting based on all georeferencing methods; Fig. 4), this indicates that they are both more cost-efficient (or accurate) than expected. To put the cost and efficiency of LLMs in a more realistic scenario: the University of North Carolina at Chapel Hill Herbarium (NCU) currently has ~500,000 specimens that are not georeferenced, and manual georeferencing of them will take ~3.3 years and cost ~\$0.8 million. These numbers will decrease to ~5 days and ~\$500 using *gpt-4o* (based on the price of *gpt-4o* API in December 2024), or ~13 days and \$0 if using free DeepSeek APIs. Furthermore, instead of fully replacing manual georeferencing, a hybrid or sequential approach could be used to balance the efficiency and reliability, i.e., to let LLMs do a first pass to be later verified by humans (as funding permits).

Georeferencing accuracy increases with human footprint

Our study also identified critical geographic factors and textual features that affect georeferencing accuracy. We found a positive relationship between the degree of human activity/development in a region and georeferencing accuracy (Fig. 5). We used human footprint as an approximation for human development, and expected a higher human footprint to provide more structural anchors and spatial references on a map, which can benefit georeferencing [34]. Indeed, LLMs, humans, and GEOLocate all showed high georeferencing accuracy in developed regions such as the U.S. and Western Europe (Fig. 3). However, the georeferencing accuracy of GEOLocate is more strongly influenced by the human footprint compared to that of LLMs or humans (Fig. 5). GEOLocate usually depends on fixed gazetteers, making it unable to resolve locations outside its database [60]. In contrast, the accuracy of LLMs was less influenced by human footprint (Fig. 5), likely because of the vast amount of data used in LLM training that is beyond gazetteers in scope and extent [61].

We didn't find a strong positive connection between georeferencing accuracy and higher text readability. More recent versions of non-reasoning LLMs (gpt-4o, gpt-4o-mini and deepseek-chat) were not significantly influenced (Wilcoxon test, $p>0.05$, $N=4750$) by any readability metric (Fig. 5); however, the georeferencing errors of earlier or simpler versions of non-reasoning LLMs (gpt-4-turbo and gpt-3.5-turbo) were higher for descriptions that scored higher by SMOG (Fig. 5). SMOG measures sentence complexity based on the number of complex words [39], indicating that early or simpler LLMs were less able to read and understand complex words. The georeferencing error of OpenAI reasoning model o1-preview was positively influenced by CLI ($p<0.05$). Increased CLI and FKGL would also significantly trouble human georeferencing ($p<0.05$). CLI and FKGL primarily measure the length of words and sentences [37, 40]. However, the effects are not particularly strong. In fact, the locality descriptions are usually not overly hard to read or interpret, because the locality descriptions were typically short paragraphs of text written on small-sized labels, thus there is no space to convey long or complex information. Therefore, sentence readability is not a key factor in determining the accuracy of LLM georeferencing.

We found mixed evidence for increased georeferencing accuracy with more detailed textual descriptions. The accuracy of GEOLocate is positively influenced by the frequency of keywords related to distance, roads, and natural and artificial objects, while the influence of textual descriptions was weaker for LLMs (Fig. 5b). The results for GEOLocate were expected because georeferencing in GEOLocate relies on predefined functions of text matching and spatial inferences [21]. Interestingly, for LLMs, the frequency of keywords related to road, natural, and artificial objects led to increased georeferencing accuracy, while the frequency of keywords related to direction and distance had the opposite effects (Fig. 5b). This is likely because the prior set of keywords can provide more spatial anchors or references for the LLM to use, while the latter set of keywords is more about spatial information that relies on spatial reasoning, which indicates the potential weakness of LLMs in spatial reasoning [32, 62]. Nevertheless, more complex spatial information is known to increase the essential difficulty of georeferencing [60].

Georeferencing of the future

LLM-driven georeferencing faces key challenges. The first is related to the uncertainty of georeferenced coordinates. Georeferenced coordinates are commonly accompanied by an uncertainty value, which is often recorded as the maximum distance from a center coordinate of a georeference to the furthest point where the true location might be [63]. Specialized tools and methodology have been developed to calculate uncertainty values based on spatial features (e.g., area size or offset distance) [64]. However, in practice, uncertainty values are very often not recorded [65]. Also, when calculating uncertainty, the previously developed tools and methodology are often not used; instead, the determination of the uncertainty often relies on personalized workflows [16, 66]. Therefore, the evaluation of the uncertainty of georeferencing becomes a difficult task. Further, LLMs are limited in their capacity to provide an "uncertainty value" (in the sense of a statistical uncertainty) because LLMs generate responses by predicting the next token based on learned patterns, and the predictions are more of a reflection of training data rather than being calibrated to reflect real-world uncertainty [67, 68]. Another challenge, partly related to the uncertainty issue, is that LLMs typically always return some results, even when the input location description makes no sense [69]. In other words, when the input data is inappropriate for georeferencing, LLMs will still generate a seemingly valid output, while humans are able to determine that such a description is not sufficient for determining coordinates. Additionally, humans are able to set some thresholds for how accurate a description must be to warrant georeferencing: if a locality description only mentions the country or state/province of occurrence with no more detailed information, a human can decide whether or not to georeference that description. Special techniques are needed to fine-tune an LLM to handle such scenarios [70]. Lastly, georeferencing faces the challenge that historical specimen records often cite missing landmarks or outdated boundaries [60]; though this challenge is not limited to LLMs. Historical maps are often used to facilitate the manual georeferencing of historical localities, but this step is time-consuming. Therefore, future studies may explore the incorporation of historical maps, as well as contextual information, such as year of collection, into the LLM-facilitate georeferencing, via prompt engineering [71] or model fine-tuning [72].

Concluding remarks

The ability to better harness the information within our invaluable natural history collections is critical to addressing the grand environmental challenges we face. LLMs present a cost-effective approach for specimen digitization and thus should be incorporated in future georeferencing workflows. LLMs may not fully replace human curation, but can be used by humans to greatly increase the efficiency of georeferencing. Most natural history collections are underfunded and understaffed [73, 74] - using LLMs to conduct first-pass georeferencing to be later verified by humans can greatly increase the number of records that can be georeferenced by existing staff. Further, these first-pass LLM georeferenced records can be immediately used for purposes that do not require the highest possible level of spatial accuracy. We have demonstrated the potential of LLMs to revolutionize the process of georeferencing. With further advances in LLMs, they may prove instrumental in rapidly providing the large amounts of biodiversity data we require to face the grand environmental challenges of our era.

Methods

Specimen selection

We obtained preserved specimen records of vascular plants from the Global Biodiversity Information Facility (GBIF), one of the largest biodiversity databases. We chose plants as a test case, as plants generally remain fixed in space over their lifetimes, thus decreasing potential uncertainties in the georeferencing process. Our initial dataset comprised records of preserved specimens collected between 2000 and 2024 across all continents except Antarctica. These specimens have known GPS coordinates, no geospatial issues according to GBIF's record-flagging procedures (which identify suspect coordinates), and belong to the plant division Tracheophyta (vascular plants). The coordinates collected from GPS devices were assumed to be the ground truth in the following evaluations. The initial dataset included a total of 13,064,051 records (DOI: https://doi.org/10.15468/dl.fj3sqk).

We performed additional data cleaning to enhance the reliability of these records. First, we removed records without locality information (11,738,740 left). Second, we excluded records with coordinates that were not recorded using GPS devices (e.g., handheld GPS units) in the field, as we intended to use the recorded coordinates to evaluate the accuracy of georeferenced results. Information on the method of georeferencing is recorded in the fields "georeferenceProtocol", "georeferenceSources", and "georeferenceRemarks." We only kept records containing the word "GPS" in the description of these attributes, and excluded those with "Google", "GEOLocate", "OpenStreetMap", or other georeferencing tools (735,145 left). Third, we removed duplicated location records based on latitude, longitude, and locality description (184,772 left). We also removed records missing information on country, state/province, and county. Records with locality descriptions of fewer than 5 words were also removed (165,581 left). Finally, we removed records with latitude and longitude embedded in the locality information description, to avoid the possibility of "cheating" during the georeferencing process.

The original data was reduced to 138,617 unique location records after cleaning. The counts for each continent are as follows: 570 from Africa, 1,558 from Asia, 82,577 from Oceania, 353 from Europe, 51,955 from North America, and 1,604 from South America. To ensure balanced sampling across continents, we randomly sampled 1,000 each from Asia, Oceania, and South America; 500 and 300 from Africa and Europe, respectively (due to fewer records from those continents); and 1,200 from North America for georeferencing performance evaluations.

Georeferencing with large language models and traditional methods

We accessed the APIs for OpenAI and DeepSeek models through the "openai" (version

1.66.3) Python package [75]. We combined each record's country, state/province, county, and locality into a list for the script. We used a one-shot prompting strategy that defined the role of georeferencing in the domain of biogeography and ecology, and specified the format of input and output data and the steps to follow (Box 1). The prompt was also followed by one example of locality description and georeferenced coordinates, a strategy that is known to improve LLM performance [76]. The "input_data" represents each of the selected 5,000 records' input list in the loop. The "temperature" of the LLM controls the randomness and predictability of the model's output, which we set to 0.01 (near zero) to ensure deterministic answers. This "temperature" index is only applicable to the non-reasoning models (*gpt-4o, gpt-4o-mini, gpt-3.5-turbo, gpt-4-turbo*, and *deepseek-chat*), and is not applicable to the reasoning models (*o1-preview* and *deepseek-reasoner*). Moreover, due to the potential high financial costs, we did not run *o1-preview* and *deepseek-reasoner* on all 5,000 samples. Instead, we conducted the analysis on a systematic sample of 100 points. These 100 samples were also used in the subsequent manual georeferencing experiment. The selection of the 100 records is detailed in the following "Manual Georeferencing" section.

Box 1. Prompt used for georeferencing with large language models:

You are an assistant specializing in georeferencing locations using locality descriptions.

You have been assigned a task for georeferencing coordinates in the domain of biogeography and ecology.

You will follow the instructions below to obtain the coordinates of input location description.

1. You will be given a Python list of 4 strings that represent country/region, state/province, county and locality information.

2. The 4 strings in the Python list represent increasing accuracy of the location.

3. The priority of information is 'locality information', 'county', 'state/province', 'country/region'. When more accurate information is available, you will prioritize the use of that information.

4. The output will be a Python list of 2 float numbers, the first float number represents latitude, the second float number represents longitude.

5. Please only output the list without any explanations.

An example of input data looks like this:

'["United States", "California", "San Bernardino", "Along Santa Ana River wash upstream from La Cadena Ave, both railroad tracks, and under powerline."] '

The expected output looks like this:

" [34.0459, -117.32332] "

Now, you will georeference this record: **input_data**

We batch-georeferenced the selected 5,000 records in R (v4.2.2) using GEOLocate v2 web services by inputting country, state/province, county, and locality information. The output coordinates were directly used. We did not perform any additional manual intervention of the coordinates; a similar methodology was used in Murphey *et al.* [60]. We recognize that manual intervention is commonly done in practice [57]; however, manual intervention overlaps with the manual georeferencing that we performed in the next step. When multiple possible outputs were returned for one input, the output with the highest precision score would be kept. If multiple outputs had the same precision scores, they would all be kept for the accuracy evaluation; note that we used the mean of their error distances (see next section), instead of the mean of their coordinates, for our analyses. Precision score is a reliability assessment of all output results by GEOLocate, with higher scores generally indicating greater reliability [21].

Finally, to serve as a benchmark for comparison of both LLMs and GEOLocate, manual georeferencing of 100 records was performed by nine human participants. The nine participants included 2 undergraduate students, 3 graduate students, 2 postdocs, and 2 faculty, who all had prior experience working with specimen records. We divided the 5,000 records into 10 groups based on the deciles of georeferencing error distances of the best non-reasoning LLM (i.e., *gpt-4o*). The calculation of georeferencing error is detailed in the next section. We then randomly sampled 10 records from each decile for a total of 100 records. Then, each of the nine participants was tasked with georeferencing ~33 records, resulting in each of the 100 records being georeferenced independently by three participants. The nine participants received the same instructions for using Google Maps or Google Earth to georeference their records. The information provided to participants was the same as that used for LLMs and GEOLocate, i.e. country, state/province, county, and locality information. For each record, participants first used the search box to locate and define the general area of the record based on explicitly mentioned place names in the locality description. Then, utilizing the "measure distance" tool and referencing the orientation and distance details provided in the locality description, the participants pinpointed the most probable location and recorded the latitude and longitude provided by Google Maps or Google Earth.

Evaluation of georeferencing accuracy

We used the "distHaversine" function in the R package "geosphere" (version 1.5-20) [77] to calculate the distance between a georeferenced coordinate and the ground truth coordinate (i.e., error distance). Larger distances represent lower accuracy. The outliers (i.e., top 5 percentile) of each georeferencing method were excluded from the analysis. This helps minimize the effect of large errors that could affect the overall results and mislead the interpretation of georeferencing accuracy. By removing these extreme values, the analysis concentrates on the majority of the data, giving a more accurate estimate of the method's usual performance. We also calculated the mean and standard deviation of the error distances across 5,000 or 100 records for different georeferencing methods. For GEOLocate, when multiple output coordinates had the same highest precision scores, the error distance was calculated as the mean of the distances of these highest-scoring coordinates. For manual georeferencing, the error distance for each record was calculated as the mean of the distances from the three participants (repetitions). As a control, we extracted the centroid of the county for each record and calculated its distance to the true coordinates, which is a common approach for georeferencing without detailed locality descriptions [35, 36, 78, 79]; thus the county-centroid approach provided a baseline for georeferencing without incorporating locality information. We performed Wilcoxon tests to evaluate the difference in accuracy among different georeferencing methods. We also visualized the georeferencing accuracy on maps, and summarized the accuracy by country.

Evaluation of georeferencing efficiency and cost

To compare the efficiency and cost of different georeferencing methods, we recorded the time taken to georeference each record and calculated the monetary cost for georeferencing 100 records. For georeferencing with LLM and GEOLocate, we used the "time.perf_counter" function in Python (version 3.8.12) and "Sys.time" function in R (version 4.2.2) to record the execution time of each loop (precise to milliseconds). Compared to the georeferencing time, the data preparation time in Python or R was minor, thus the choice of Python or R programming environments did not affect the efficiency comparison. Additionally, we used the API expenditure (in USD) of different LLMs from OpenAI and Deepseek's platform webpages (accessed on December 1, 2024). GEOLocate is a free software/service, thus its monetary cost is always $0. During the manual georeferencing processes, every participant was requested to record the time taken to complete each record using the same online timer (https://www.online-stopwatch.com/). We calculated the human cost based on a typical salary rate of curators ($25/hr) who are the typical personnel that perform georeferencing tasks in museums and herbaria.

We performed generalized linear models to investigate the relationship between georeferencing accuracy and georeferencing time and monetary cost. We applied a base-10 logarithmic transformation to georeferencing time and monetary cost to reduce scale disparities and mitigate the influence of large values. We then compared the goodness of fit

(R^2) between linear and exponential models with georeferencing error distance as the dependent variable and georeferencing time or monetary cost as the independent variable. The model with the better fit (R^2) was selected as the representative accuracy-time/cost expectation curve.

Factors that affect georeferencing accuracy

The visualization of outputs showed that the georeferencing error distances were not uniform across regions; thus, we further investigated potential geographic factors and textual features that may affect georeferencing accuracy. We hypothesized that regions with higher human activity and greater development would have more geographical reference points (e.g., more documented location names on a map) that are accessible to both LLMs and humans, thus leading to increased georeferencing accuracy. To test this hypothesis, we extracted the human footprint index based on the ground truth coordinates of the 5,000 specimen records from the Global Human Footprint Dataset of the Last of the Wild Project, Version 2, with around 1 km resolution [80]. This dataset integrates nine global data layers, including human population pressure (population density), land use and infrastructure (built-up areas, nighttime lights, land cover), and human accessibility (coastlines, roads, railroads, navigable rivers). We used the human footprint index as an approximation for human activity and development. We used locally weighted regression (LOESS) curves to analyze the relationship between georeferencing error and human footprint index. This analysis was performed for all georeferencing methods.

We also hypothesized that higher text readability and the more detailed textual descriptions in locality would lead to increased georeferencing accuracy. To quantitatively evaluate the readability of the locality text, we employed 5 commonly used readability metrics: Flesch-Kincaid Grade Level (FKGL) [37], Gunning Fog Index (GFI) [38], Simple Measure of Gobbledygook (SMOG) [39], Coleman-Liau Index (CLI) [40] and Automated Readability Index (ARI) [41]. The definition and calculation of each metric were shown in Table 1; a higher value of each metric indicates more complex text (thus lower readability). These metrics were chosen for their diverse approaches to assessing text complexity, offering a comprehensive view of readability for various applications, from education to technical documentation [81]. The calculations were performed in R using the “quanteda.textstats” package (v 0.97.2) [82]. Then, to measure the level of detail in the locality descriptions, we calculated the frequency of numbers, punctuation marks, directional terms, and distance units, as well as the frequency of road names, artificial objects, and natural objects (Table 1) using R package “stringr” (v 1.5.1) [83]. The georeferencing error distance was treated as a dependent variable, and the above readability metrics and textual features were treated as independent variables. We used general linear models to examine the relationship between georeferencing error distance and textual features. To mitigate multicollinearity among the independent variables, we performed univariate regressions, where each independent variable is regressed separately. The regression coefficients measure the individual effects of each textual feature, which are expected to show how much the error distance increases in kilometers for each one-unit increase in the independent variables. The regression analysis was performed for each georeferencing method, respectively.

Reference

1. Shaffer HB, Fisher RN, Davidson C. The role of natural history collections in documenting species declines. *Trends in Ecology and Evolution* 1998, **13**(1)**:** 27-30.
2. Lister AM. Natural history collections as sources of long-term datasets. *Trends in ecology & evolution* 2011, **26**(4)**:** 153-154.
3. Holmes MW, Hammond TT, Wogan GO, Walsh RE, LaBarbera K, Wommack EA, Martins FM, Crawford JC, Mack KL, Bloch LM. Natural history collections as windows on evolutionary processes. *Molecular Ecology* 2016, **25**(4)**:** 864-881.
4. Osawa T, Watanabe K, Ikeda H, Yamamoto S. New approach for evaluating habitat stability using scarce records for both historical and contemporary specimens: a case study using C arabidae

specimen records. *Entomological Science* 2014, **17**(4)**:** 425-431.

5. Meineke EK, Davies TJ. Museum specimens provide novel insights into changing plant–herbivore interactions. *Philosophical Transactions of the Royal Society B* 2019, **374**(1763)**:** 20170393.
6. Park DS, Breckheimer IK, Ellison AM, Lyra GM, Davis CC. Phenological displacement is uncommon among sympatric angiosperms. *New Phytologist* 2022, **233**(3)**:** 1466-1478.
7. Xie Y, Thammavong HT, Berry LG, Huang CH, Park DS. Sex-dependent phenological responses to climate vary across species' ranges. *Proceedings of the National Academy of Sciences* 2023, **120**(48)**:** e2306723120.
8. Zhang S-Q, Che L-H, Li Y, Liang D, Pang H, Ślipiński A, Zhang P. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nature Communications* 2018, **9**(1)**:** 205.
9. Johnson MG, Pokorny L, Dodsworth S, R.Botigué L, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Félix.Forest, Kim JT. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Systematic Biology* 2019, **68**(4)**:** 594-606.
10. Hamilton H, Smyth RL, Young BE, Howard TG, Tracey C, Breyer S, Cameron DR, Chazal A, Conley AK, Frye C. Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. *Ecological Applications* 2022, **32**(3)**:** e2534.
11. Schmitt CJ, Cook JA, Zamudio KR, Edwards SV. Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philosophical Transactions of the Royal Society B* 2019, **374**(1763)**:** 20170387.
12. Hedrick BP, Heberling JM, Meineke EK, Turner KG, Grassa CJ, Park DS, Kennedy J, Clarke JA, Cook JA, Blackburn DC. Digitization and the future of natural history collections. *BioScience* 2020, **70**(3)**:** 243-251.
13. Paton A, Antonelli A, Carine M, Forzza RC, Davies N, Demissew S, Dröge G, Fulcher T, Grall A, Holstein N. Plant and fungal collections: Current status, future perspectives. *Plants, People, Planet* 2020, **2**(5)**:** 499-514.
14. Park DS, Feng X, Akiyama S, Ardiyani M, Avendaño N, Barina Z, Bärtschi B, Belgrano M, Betancur J, Bijmoer R. The colonial legacy of herbaria. *Nat Hum Behav* 2023, **7**(7)**:** 1059-1068.
15. Nelson G, Ellis S. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B* 2019, **374**(1763)**:** 20170391.
16. Marcer A, Groom Q, Haston E, Uribe F. Natural history collections georeferencing survey report. *Current georeferencing practices across institutions worldwide: Zenodo* 2021.
17. Nelson G, Paul D, Riccardi G, Mast AR. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* 2012(209)**:** 19.
18. Seltmann K, Lafia S, Paul D, James S, Bloom D, Rios N, Ellis S, Farrell U, Utrup J, Yost M. Georeferencing for Research Use (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers. *Research Ideas and Outcomes* 2018, **4**.
19. Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB, Peterson AT, Soberón J, Pearson RG. *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press, 2011.
20. Peterson AT, Soberón J, Ramsey J, Osorio-Olvera L. Co-occurrence networks do not support identification of biotic interactions. *Biodiversity Informatics* 2020, **15**(1)**:** 1-10.
21. Rios N, Bart H. GEOLocate (Version 3.22). 2.0 ed. Belle Chasse, LA: Tulane University Museum of Natural History; 2010.
22. Groom Q, Dillen M, Hardy H, Phillips S, Willemse L, Wu Z. Improved standardization of transcribed digital specimen data. *Database* 2019, **2019:** baz129.
23. Vollmar A, Macklin JA, Ford L. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics* 2010, **7**(2).
24. Bhandari P, Anastasopoulos A, Pfoser D. Are large language models geospatially knowledgeable? Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems; 2023; 2023. p. 1-4.
25. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys* 2023, **56**(2)**:** 1-40.
26. Farrell MJ, Brierley L, Willoughby A, Yates A, Mideo N. Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B* 2022, **289**(1975)**:** 20212721.
27. Farrell MJ, Le Guillarme N, Brierley L, Hunter B, Scheepens D, Willoughby A, Yates A, Mideo N. The changing landscape of text mining: a review of approaches for ecology and evolution. *Proceedings B* 2024, **291**(2027)**:** 20240423.
28. Castro A, Pinto J, Reino L, Pipek P, Capinha C. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics* 2024, **82:** 102742.

29. Mora-Cross M, Calderon-Ramirez S. Uncertainty estimation in large language models to support biodiversity conservation. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track); 2024; 2024. p. 368-378.
30. Dorm F, Millard J, Purves D, Harfoot M, Mac Aodha O. Large language models possess some ecological knowledge, but how much? *BioRxiv* 2025**:** 2025.2002. 2010.637097.
31. Haider T, Perschl T, Rehbein M. Quantification of Biodiversity from Historical Survey Text with LLM-based Best-Worst Scaling. *arXiv preprint arXiv:250204022* 2025.
32. Cohn AG, Blackwell RE. Evaluating the ability of large language models to reason about Cardinal directions. *arXiv preprint arXiv:240616528* 2024.
33. Li Z, Zhou W, Chiang Y-Y, Chen M. Geolm: Empowering language models for geospatially grounded language understanding. *arXiv preprint arXiv:231014478* 2023.
34. Yan Y, Lee J. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management; 2024; 2024. p. 4163-4167.
35. Park DS, Davis CC. Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography* 2017, **44**(10)**:** 2188-2198.
36. Park DS, Xie Y, Thammavong HT, Tulaiha R, Feng X. Artificial hotspot occurrence inventory (AHOI). *Journal of Biogeography* 2023, **50**(2)**:** 441-449.
37. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida: Millington (Tenn.), 1975.
38. Gunning R. *The technique of clear writing*. McGraw-Hill Rev. ed: New York, 1952.
39. Mc Laughlin GH. SMOG grading-a new readability formula. *Journal of Reading* 1969, **12**(8)**:** 639-646.
40. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 1975, **60**(2)**:** 283.
41. Smith EA, Senter R. *Automated readability index*, vol. 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, 1967.
42. Sui Y, Chuang Y-N, Wang G, Zhang J, Zhang T, Yuan J, Liu H, Wen A, Chen H, Hu X. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:250316419* 2025.
43. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:230211382* 2023.
44. Lin X, Wang W, Li Y, Yang S, Feng F, Wei Y, Chua T-S. Data-efficient Fine-tuning for LLM-based Recommendation. Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval; 2024; 2024. p. 365-374.
45. Bloom TD, Flower A, DeChaine EG. Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecology and Evolution* 2018, **8**(1)**:** 765-777.
46. Smith AB, Murphy SJ, Henderson D, Erickson KD. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography* 2023, **32**(3)**:** 342-355.
47. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 2017, **37**(12)**:** 4302-4315.
48. Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst Sci Data* 2021, **13**(9)**:** 4349-4383.
49. Harris I, Osborn TJ, Jones P, Lister D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci Data* 2020, **7**(1)**:** 109.
50. Lawler JJ, Shafer SL, White D, Kareiva P, Maurer EP, Blaustein AR, Bartlein PJ. Projected climate-induced faunal change in the Western Hemisphere. *Ecology* 2009, **90**(3)**:** 588-597.
51. Yilmaz M. Accuracy assessment of temperature trends from ERA5 and ERA5-Land. *Science of the Total Environment* 2023, **856:** 159182.
52. Dalei H, Qing X, Jianguang W, Dongqin Y, Xiaodan W, Xingwen L, Shengbiao W. Advances in upscaling methods of quantitative remote sensing. *National Remote Sensing Bulletin* 2021, **22**(3)**:** 408-423.
53. Atkinson PM. Downscaling in remote sensing. *International Journal of Applied Earth Observation*

and Geoinformation 2013, **22:** 106-114.

54. Pearson RG, Dawson TP. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* 2003, **12**(5)**:** 361-371.
55. Meijman TF. Mental fatigue and the efficiency of information processing in relation to work times. *International Journal of Industrial Ergonomics* 1997, **20**(1)**:** 31-38.
56. Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers BM. Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys* 2012(209)**:** 103.
57. Mast AR, Paul DL, Rios N, Bruhn R, Dalton T, Krimmel ER, Pearson KD, Sherman A, Shorthouse DP, Simmons NB, Soltis P, Upham N, Abibou D. Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses (1.6). Zenodo; 2021.
58. Neha F, Bhati D. A Survey of DeepSeek Models. *TechRxiv* 2025.
59. Chowdhury MN-U-R, Haque A, Ahmed I. DeepSeek vs. ChatGPT: A Comparative Analysis of Performance, Efficiency, and Ethical AI Considerations. *TechRxiv* 2025.
60. Murphey PC, Guralnick RP, Glaubitz R, Neufeld D, Ryan JA. Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). *Phyloinformatics* 2004, **1**(3)**:** 1-29.
61. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog* 2019, **1**(8)**:** 9.
62. O'Sullivan K, Schneider NR, Samet H. Metric Reasoning in Large Language Models. Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems; 2024; 2024. p. 501-504.
63. Wieczorek J, Guo Q, Hijmans R. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 2004, **18**(8)**:** 745-767.
64. Wieczorek J, Bloom D, Zermoglio P. Georeferencing Calculator Manual. 2019.
65. Marcer A, Escobar A, Chapman A, Wieczorek J. GeoPick: Georeferencing made easy. *Biodiversity Information Science and Standards* 2023, **7:** e111036.
66. Zhang S, Barrett HA, Baros SV, Neville PR, Talasila S, Sinclair LL. Georeferencing Accuracy Assessment of Historical Aerial Photos Using a Custom-Built Online Georeferencing Tool. *ISPRS International Journal of Geo-Information* 2022, **11**(12)**:** 582.
67. Ye F, Yang M, Pang J, Wang L, Wong D, Yilmaz E, Shi S, Tu Z. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems* 2024, **37:** 15356-15385.
68. Hou B, Liu Y, Qian K, Andreas J, Chang S, Zhang Y. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:231108718* 2023.
69. Sun L, Huang Y, Wang H, Wu S, Zhang Q, Gao C, Huang Y, Lyu W, Zhang Y, Li X. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:240105561* 2024, **3**.
70. Tyen G, Mansoor H, Cărbune V, Chen P, Mak T. LLMs cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:231108516* 2023.
71. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:231014735* 2023.
72. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: Low-rank adaptation of large language models. *ICLR* 2022, **1**(2)**:** 3.
73. Dalton R. Natural history collections in crisis as funding is slashed. *Nature* 2003, **423**(6940)**:** 575-576.
74. Bradley RD, Bradley LC, Garner HJ, Baker RJ. Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *BioScience* 2014, **64**(12)**:** 1150-1158.
75. Auger T, Saroyan E. Overview of the OpenAI APIs. *Generative AI for Web Development: Building Web Applications Powered by OpenAI APIs and Next. js*. Springer, 2024, pp 87-116.
76. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 2020, **33:** 1877-1901.
77. Hijmans RJ, Williams E, Vennes C, Hijmans MRJ. Package ‘geosphere’. *Spherical Trigonometry* 2017, **1**(7)**:** 1-45.
78. Lohonya K, Livermore L, Penn MG. Georeferencing the Natural History Museum's Chinese type collection: of plateaus, pagodas and plants. *Biodiversity Data Journal* 2020, **8:** e50503.
79. Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A. Estimating species diversity and distribution in the era of B ig D ata: to what extent can we trust public databases? *Global Ecology and Biogeography* 2015, **24**(8)**:** 973-984.

80. Wildlife Conservation Society-WCS, University CFIESIN-C-C. Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (Geographic) (Version 2.00). In: (SEDAC) NSDaAC, editor. Palisades, NY; 2005.
81. Tran H, Yao Z, Li L, Yu H. ReadCtrl: Personalizing text generation with readability-controlled instruction learning. *arXiv preprint arXiv:240609205* 2024.
82. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 2018, **3**(30)**:** 774.
83. Gagolewski M. stringi: Fast and portable character string processing in R. *Journal of Statistical Software* 2022, **103:** 1-59.

Acknowledgement

We thank the valuable suggestions from Carol Ann McCormick, curator of University of North Carolina Chapel Hill Herbarium.

Author contributions

X.F. conceived the initial idea, which was inspired through discussions with D.S.P.; X.F., Y.X., and D.S.P. designed the research; Y.X. performed the analysis, with help from J.H. and R.X.; Y.X., X.F., D.S.P., M.A.S-A, L.J.A.L., J.C., M.L., N.S., and C.H. performed the manual georeferencing; Y.X. and X.F. wrote the manuscript, with major input from D.S.P. and M.A.S-A; everyone contributed to the interpretation of the results and revision of the manuscript.

Competing interests

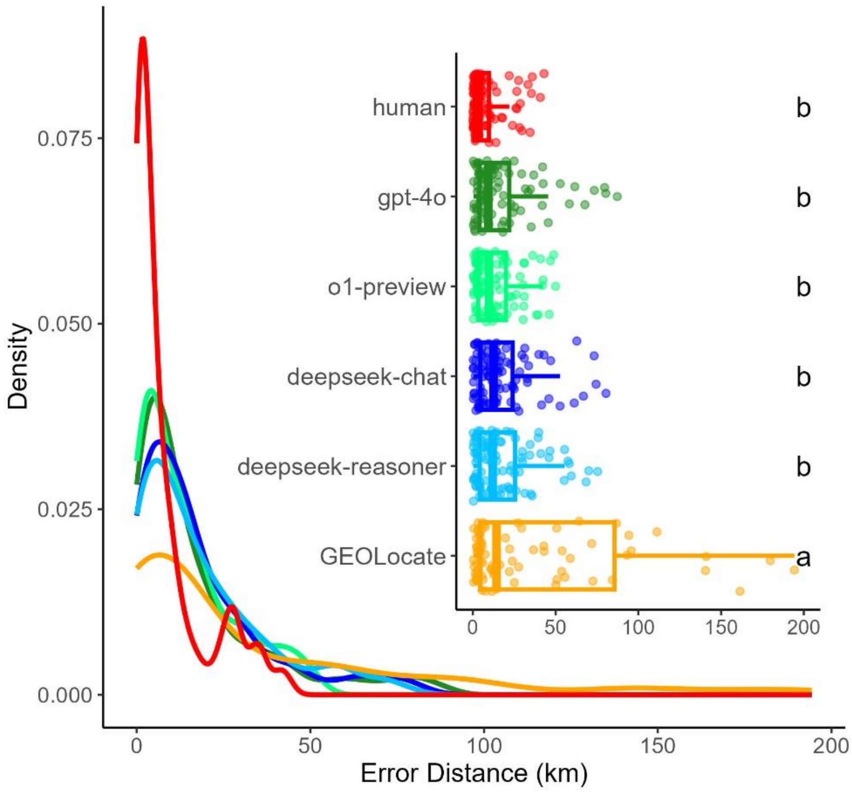
The authors declare no competing interests.

Extended Table 1. Summary of georeferencing accuracy and efficiency across various models or methods. The outliers (i.e., top 5 percentile) are removed when calculating the median, mean, and standard deviation, so the sample size being considered in statistics is either 4750 or 95.

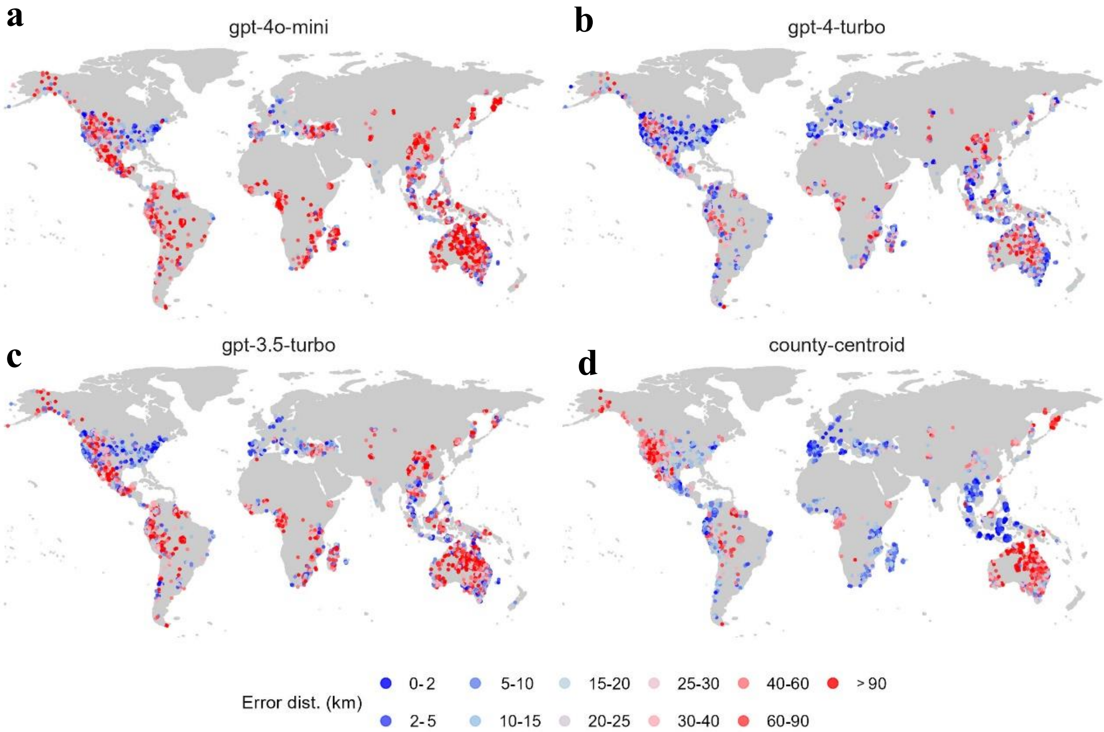
Model/method		Sample size	Median error (km)	Mean error ± sd (km)	Time per record (mean±sd, seconds)	Cost per 100 records ($)
Human		100	3.4	8.3±10.8	211±173.0	160.00
OpenAI	*o1-preview*	100	9.7	13.9±13.3	27.1±14.8	13.48
	gpt-4o	100	9.3	17.1±19.8	0.9±0.5	0.10
	gpt-4o	5000	9.7	17.8±20.7	0.9±0.5	0.10
	gpt-4o-mini	100	46.5	57.0±46.7	0.8±0.6	Free
	gpt-4o-mini	5000	39.6	54.8±49.4	0.8±0.6	Free
	gpt-4-turbo	100	14.0	21.0±19.4	1.1±0.7	0.82
	gpt-4-turbo	5000	15.3	23.8±23.6	1.1±0.7	0.82
	gpt-3.5-turbo	100	27.9	37.7±32.8	0.6±0.4	0.01
	gpt-3.5-turbo	5000	27.0	41.4±41.0	0.6±0.4	0.01
Deepseek	*deepseek-reasoner (R1)*	100	12.0	17.9±18.0	92.3±68.3	1.38
	deepseek-chat (V3)	100	12.5	17.5±18.3	2.3±0.3	Free
	deepseek-chat (V3)	5000	12.3	20.3±21.8	2.3±0.3	Free
GEOLocate		100	14.3	98.2±176.9	0.9±1.5	Free
		5000	23.4	110.2±185.9	0.9±1.5	Free
County centroid		100	17.9	30.0±30.2	0	Free
		5000	18.2	30.8±32.2	0	Free

Extended Table 2. Independent variables, include readability metrics and the counts of different word types in the locality descriptions

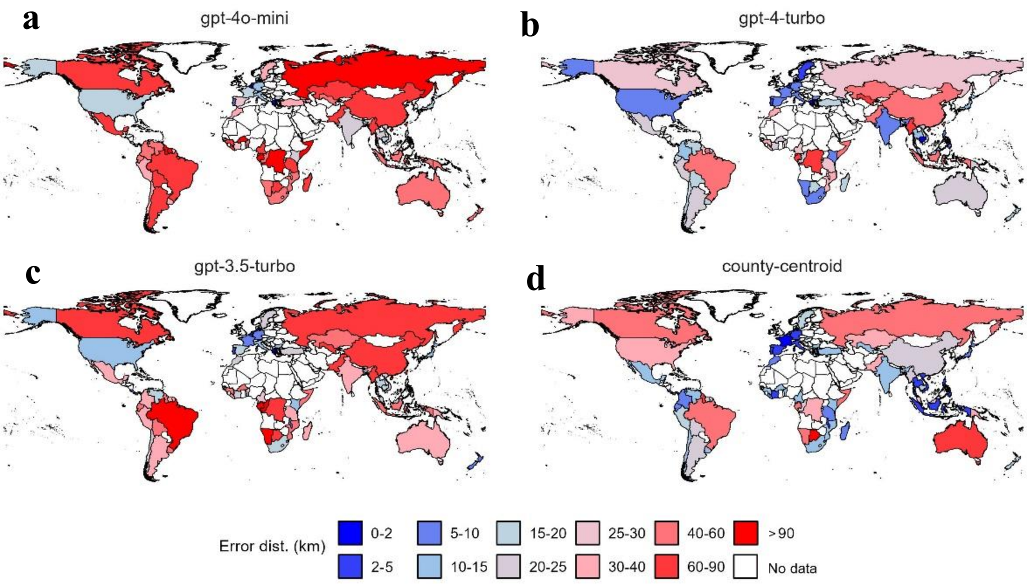
Categories	Variable	Abbr.	Keywords or illustrate
Readability metrics	Flesch-Kincaid Grade Level	FKGL	**Meaning:** estimates the U.S. school grade level required to understand a text. **Higher Value** = More complex text. **Formula:** $\mathrm{FKGL} = 0.39 \times \left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8 \times \left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$
	Gunning Fog Index	GFI	**Meaning:** measures the number of years of formal education required to understand a text easily. **Higher Value** = More complex text. **Formula:** $GFI = 0.4 \times \left(\frac{\text{Total Words}}{\text{Total Sentences}} + 100 \times \frac{Complex\ Words\ (\geq 3\ syllables)}{\text{Total Words}}\right)$
	Simple Measure of Gobbledygook	SMOG	**Meaning**: measures the readability of healthcare and academic texts by focusing on multi-syllabic words. **Higher Value** = More complex text **Formula:** $SMOG = 1.043 \times \sqrt{30 \times \frac{\text{Polysyllabic Words } (\geq 3 \text{ syllables})}{\text{Total Sentences}}}$
	Coleman-Liau Index	CLI	**Meaning**: Estimates the readability grade level based on character count rather than syllables. **Higher Value** = More complex text. **Formula:** $CLI = 0.0588 \times \left(\frac{Total\ Letters}{Total\ Words} \times 100\right) - 0.296 \times \left(\frac{Total\ Sentences}{Total\ Words} \times 100\right) - 15.8$
	Automated Readability Index	ARI	**Meaning**: A machine-calculated readability score based on word length and sentence complexity. **Higher Value** = More complex text. **Formula:** $ARI = 4.71 \times \left(\frac{Total\ Characters}{Total\ Words}\right) + 0.5 \times \left(\frac{Total\ Words}{Total\ Sentences}\right) - 21.43$
Non-text characters	Number of digits	digit	—
	Number of punctuation marks	punctuation	—
Spatial keywords	Number of direction words	direction	north, south, east, west, northeast, southeast, northwest, southwest, N, S, E, W, NE, SE, NW, SW, NNE, NNW, SSE, SSW, ENE, ESE, WNW, WSW
	Number of distance words	distance	km, m, mi, mile, miles, meter, meters, kilometer, kilometers, feet, foot
Object keywords	Number of road names	road	street, st, road, rd, avenue, ave, boulevard, blvd, drive, dr, lane, ln, highway, hwy, path, trail
	Number of natural objects	natural_obj.	river, mountain, lake, forest, sea, ocean, beach, desert, valley, canyon, waterfall, island, hill, pond, creek, bay, swamp, marsh, glacier, cliff, plain, meadow, grove, prairie, stream, woods, coast, shore, wetland, peak, brook
	Number of artificial objects (except roads)	artificial_abj.	building, structure, house, bridge, city, village, town, urban, tower, factory, dam, monument, temple, stadium, castle, fort, palace, skyscraper, residence, office, industrial, farm, plaza, apartment, church, mosque, synagogue, mall, market, school, hospital



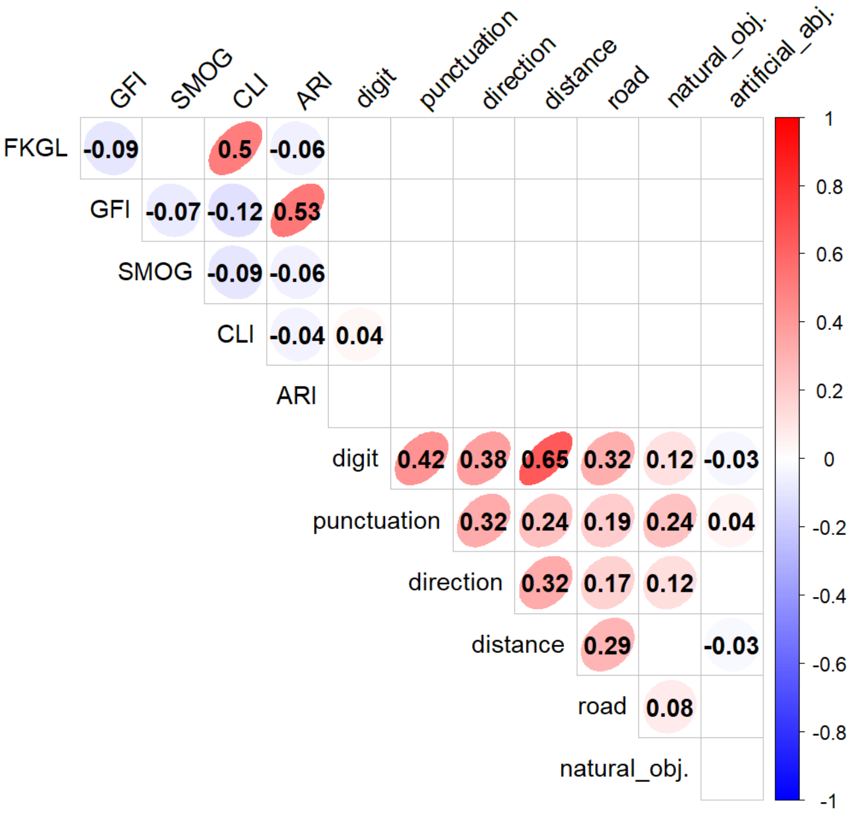
Extended Figure 1. Density plot and boxplot of georeferencing accuracy of different methods. A higher georeferencing accuracy (x-axis) is represented by a smaller error distance (distance to ground truth coordinate). The letters in the boxplot indicate intergroup differences according to the Wilcoxon test, where letters appearing later in the sequence correspond to smaller mean georeferencing errors. Identical letters signify no significant differences between methods ($p > 0.05$). In the boxplot, the models or methods are ordered from top to bottom based on the ascending median georeferencing error. Here, only the results of the most accurate LLM, GEOLocate, and manual georeferencing are displayed. All methods were applied to a sample size of 100 in this figure



Extended Figure 2. The geographic distribution of georeferencing accuracy. Georeferencing accuracy is represented by the error distance (distance between georeferenced coordinates and ground truth), and smaller values indicate higher accuracy. Maps represent three OpenAI LLMs (a-c, except *gpt-4o* shown in Figure 2a) and county centroid-based georeferencing method (d). The sample size is 5000 for each.



Extended Figure 3. The georeferencing accuracy at country level. Georeferencing accuracy is represented by the median error distance (distance between georeferenced coordinates and ground truth) of all sample points in each country, and smaller values indicate higher accuracy. Maps represent three OpenAI LLMs (a-d, except *gpt-4o* shown in Figure 3a) and county centroid-based georeferencing method (d). All country-level statistics were based on 5000 samples.



Extended Figure 4. Correlation plot between independent variables. The independent variables include all readability metrics and the counts of different word types in the locality descriptions those are listed in Extended Table 2. The numbers represent the Pearson correlation coefficients, with only statistically significant correlations ($p < 0.05$) marked.

Supplementary Information

Addressing the bottleneck of georeferencing natural history collections with large language models

Yuyang Xie[1], Daniel S. Park[2, 3], Miranda A. Sinnott-Armstrong[2, 3], Joyce Ho[4], Tianlong Chen[5], Alan S. Weakley[1, 6], Luis J. Aguirre Lopez[1], Jaein Choi[2, 3], Marisa Laitinen[2, 3], Nicholas Steeves[2], Chingyan Huang[2], Ran Xu[4], Xiao Feng[1]*

1. Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA

2. Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

3. Purdue Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA

4. Department of Computer Science, Emory University, Atlanta, GA 30322, USA

5. Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA

6. University of North Carolina Herbarium (NCU), North Carolina Botanical Garden, University of North Carolina, Chapel Hill, NC 27599, USA

* Corresponding author: fengxiao.sci@gmail.com

This file includes:

Table S1, S2 & S3

Table S1 Statistical analysis of the mean, standard deviation (sd), and median (med.) of error distances based on multiple georeferencing methods for 5000 sample points.

		gpt-4o		gpt-4o-mini		gpt-4-turbo		gpt-3.5-turbo		deepseek-chat		GEOLocate		CountyCentroid	
Country	**Count**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**
Argentina	21	25.3 ± 20.4	18.1	85.4 ± 63.3	64.9	27.4 ± 28	21.9	41.5 ± 36.4	30.9	22 ± 21.3	17.3	33.6 ± 26.5	30.4	27.6 ± 24.2	20.6
Australia	988	38.7 ± 79.3	14.4	108.9 ± 187.7	56.7	52.9 ± 109.7	22.3	79.8 ± 145.1	38.3	44.2 ± 73.5	17.6	367.9 ± 623.8	64.2	96.9 ± 137.5	61.0
Azerbaijan	16	16.1 ± 17.3	10.2	42.3 ± 30.8	30.5	26.3 ± 31.1	17.9	18.6 ± 11.1	15.9	18.4 ± 17	14.1	37.8 ± 49.7	22.9	18.7 ± 8.7	16.8
Belgium	1	2.5 ± 0	2.5	1.7 ± 0	1.7	2.9 ± 0	2.9	5 ± 0	5.0	3.2 ± 0	3.2	2.4 ± 0	2.4	2.7 ± 0	2.7
Benin	2	20.5 ± 24.7	20.5	41 ± 53.3	41.0	21 ± 25.5	21.0	20.5 ± 24.8	20.5	20.6 ± 24.9	20.6	19.5 ± 26.2	19.5	11.4 ± 16	11.4
Bolivia	38	38.5 ± 48.5	16.1	79.4 ± 70.6	66.2	37.4 ± 51.8	18.3	69.6 ± 58.6	55.5	28.4 ± 22.5	23.5	135.2 ± 144.4	64.5	33.3 ± 26.9	24.3
Botswana	1	21.7 ± 0	21.7	64.9 ± 0	64.9	19 ± 0	19.0	86.4 ± 0	86.4	28.5 ± 0	28.5	63.3 ± 0	63.3	95.1 ± 0	95.1
Brazil	286	41.3 ± 59.5	17.1	118.9 ± 99	68.0	43.6 ± 28.1	42.1	96.4 ± 48.2	105.2	37 ± 34.7	25.8	611.8 ± 369.1	648.6	57.4 ± 29.6	57.7
Bulgaria	36	5.9 ± 4.9	5.2	26.8 ± 24.3	19.4	9.3 ± 6.3	9.0	21.5 ± 20.8	14.4	10.9 ± 10.9	8.7	53.2 ± 87.7	6.5	12.3 ± 6.3	10.6
Burkina Faso	10	18.8 ± 16.4	15.5	106.2 ± 87.6	112.4	36 ± 38.2	23.9	55 ± 52.6	40.8	14.1 ± 7.8	15.3	12.6 ± 11.9	14.2	10.9 ± 4	10.1
Cambodia	1	3.3 ± 0	3.3	16.6 ± 0	16.6	1.1 ± 0	1.1	15.8 ± 0	15.8	0.7 ± 0	0.7	3.5 ± 0	3.5	18.6 ± 0	18.6
Cameroon	108	20.2 ± 18.6	15.0	60.8 ± 46.4	60.6	29.6 ± 20.4	26.7	41 ± 31	27.1	23.2 ± 15.9	18.6	61.1 ± 115.2	16.8	12.1 ± 7.9	10.3
Canada	4	27.1 ± 23.2	22.4	64.1 ± 28.4	70.2	26.5 ± 11.8	29.5	70.6 ± 34.5	73.3	46 ± 18.7	45.7	288 ± 555.8	14.5	45.5 ± 25.9	41.6
Chile	43	14.7 ± 19.6	6.2	42.1 ± 35.5	31.7	21.8 ± 20	17.9	34.7 ± 35	24.4	17.3 ± 22.9	8.9	130.7 ± 203.1	31.3	16.2 ± 12	13.3
China	125	51 ± 40.9	44.5	73 ± 40	66.6	64.5 ± 49.7	56.4	73.7 ± 51.1	65.0	57.5 ± 40.6	48.4	509 ± 588.8	294.0	22.4 ± 11.4	20.3
Colombia	130	13.2 ± 15.6	8.5	56.1 ± 40.2	45.8	15.8 ± 14.7	11.1	77.2 ± 195.5	35.1	15.2 ± 12.5	10.9	115.6 ± 139.9	69.6	10.3 ± 10.3	6.7
Costa Rica	2	9.6 ± 12	9.6	13.8 ± 9.6	13.8	12.8 ± 8.4	12.8	9.5 ± 11.5	9.5	19.2 ± 3.9	19.2	53.7 ± 48.6	53.7	7.7 ± 0.2	7.7
Côte d'Ivoire	1	23.8 ± 0	23.8	27.3 ± 0	27.3	23.9 ± 0	23.9	23.6 ± 0	23.6	47 ± 0	47.0	23.8 ± 0	23.8	2.1 ± 0	2.1
Democratic Republic of the Congo	1	22.3 ± 0	22.3	107.3 ± 0	107.3	74.8 ± 0	74.8	81.1 ± 0	81.1	55.7 ± 0	55.7	10.9 ± 0	10.9	35.1 ± 0	35.1
Denmark	11	4.2 ± 3.3	3.1	12.7 ± 7	14.2	6.8 ± 4.7	5.6	13.4 ± 11.6	13.0	9.7 ± 10	4.0	24.5 ± 36.9	9.4	13.1 ± 9.5	12.8
Ecuador	200	20.3 ± 26.4	11.8	53.4 ± 39	44.6	21.4 ± 18.1	17.7	66.3 ± 65.5	42.2	16.1 ± 13.5	12.0	71.4 ± 108.8	25.1	11.6 ± 13.4	8.5
Equatorial Guinea	2	115.2 ± 147	115.2	67.5 ± 3.7	67.5	66.6 ± 5.3	66.6	54.7 ± 17.5	54.7	33.7 ± 27.5	33.7	11.9 ± 0	11.9	7.1 ± 1	7.1
France	3	28.1 ± 42.7	3.7	24 ± 22.8	18.3	5.9 ± 2.5	5.6	10.8 ± 7.9	9.0	1.9 ± 1.5	1.3	27.7 ± 45.5	2.2	1.8 ± 0.4	1.8
French Guiana	14	21.9 ± 26.7	12.7	30.2 ± 16.5	31.5	22.4 ± 23.8	14.9	39.1 ± 25.6	34.6	16.4 ± 18.8	12.2	47.5 ± 37.8	50.7	24.7 ± 22	17.5
Gabon	39	57.5 ± 59.8	34.7	196.6 ± 85.8	203.5	85.7 ± 77.9	61.9	139.2 ± 197.1	96.6	50.1 ± 32.7	39.9	122.2 ± 213	26.1	35.8 ± 13.6	40.3
Georgia	146	11.8 ± 11.2	8.8	52.9 ± 35.5	47.1	14.9 ± 13	10.5	24 ± 24.4	12.2	14.5 ± 15.5	8.2	29.1 ± 45.2	11.3	14.8 ± 9.1	12.7

		gpt-4o		gpt-4o-mini		gpt-4-turbo		gpt-3.5-turbo		deepseek-chat		GEOLocate		CountyCentroid	
Country	**Count**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**
Germany	2	2.2 ± 1.5	2.2	12.4 ± 1.1	12.4	9.7 ± 0.7	9.7	10 ± 2.1	10.0	6.1 ± 6	6.1	29.8 ± 40.2	29.8	4.6 ± 5.3	4.6
Greece	1	0.2 ± 0	0.2	0.3 ± 0	0.3	0.5 ± 0	0.5	0.6 ± 0	0.6	0.1 ± 0	0.1	35.4 ± 0	35.4	15.5 ± 0	15.5
Guinea	10	41.4 ± 31.4	44.0	81.7 ± 54.9	62.2	45.3 ± 25.9	35.5	48.3 ± 23.3	47.0	39.2 ± 24.5	45.2	51.7 ± 14.5	44.9	12.7 ± 5.7	10.2
Guyana	51	41.8 ± 33.8	34.8	95.4 ± 54.9	112.0	49.7 ± 55.6	27.3	61.4 ± 48.9	49.2	33.9 ± 29.1	33.0	46.6 ± 87.1	15.7	70.6 ± 42.1	74.7
India	12	12.5 ± 17.6	3.4	51.2 ± 58.9	23.4	15.1 ± 17	9.2	37.7 ± 24.6	32.6	22.8 ± 23.6	14.7	312.8 ± 453.9	69.5	13.2 ± 9.4	11.1
Indonesia	106	31.2 ± 49.5	15.8	66.1 ± 59.4	46.5	45.3 ± 43.1	42.3	95.2 ± 138.9	59.1	128.4 ± 1074.3	15.4	118.8 ± 255	63.8	4.5 ± 4	3.1
Italy	22	7.6 ± 6.7	6.0	16.7 ± 11.5	14.8	12.3 ± 11	9.0	22.7 ± 22.8	14.8	11.3 ± 8.9	11.2	73.8 ± 57	66.9	5.2 ± 4.2	3.9
Japan	2	22.3 ± 26.2	22.3	16.9 ± 18.4	16.9	10.8 ± 9.8	10.8	12.2 ± 9.7	12.2	12.3 ± 16.1	12.3	47.8 ± 54	47.8	8.9 ± 6.4	8.9
Kazakhstan	11	49.7 ± 91.3	4.7	128.6 ± 107.6	62.7	68.9 ± 101.3	41.7	61.2 ± 96.5	40.8	53.1 ± 87.4	9.3	307 ± 304.1	336.1	51.7 ± 41.7	35.7
Kenya	7	10.4 ± 12	4.9	33.4 ± 28.5	23.6	8.9 ± 9.3	6.2	140.7 ± 229.1	14.8	9.3 ± 10.1	4.5	50.3 ± 47.9	51.9	15.6 ± 11.4	14.3
Liberia	5	45.5 ± 35.3	31.9	100.8 ± 86	87.9	34.8 ± 20.3	29.6	48.2 ± 29.1	32.3	32.2 ± 8.7	32.4	112.3 ± 115.1	60.4	7.9 ± 4.7	7.4
Madagascar	81	26.8 ± 38.7	13.0	204.5 ± 567.8	66.4	32.5 ± 41.3	16.4	54.2 ± 70.6	30.1	29.4 ± 51.2	12.9	139.6 ± 175.9	89.4	10.2 ± 6.3	8.7
Malawi	5	15.3 ± 7.1	10.5	34.2 ± 10.3	36.0	32.8 ± 24.8	27.5	8.9 ± 1.7	8.0	13.9 ± 3.2	16.0	35.2 ± 0	35.2	3.8 ± 2.2	2.3
Malaysia	61	11 ± 16.9	4.7	46.8 ± 44.7	31.0	20.8 ± 25.4	10.6	29.1 ± 31.4	16.0	12.1 ± 14.6	4.5	54.1 ± 109.8	17.3	37.7 ± 30.7	26.2
México	318	21.3 ± 24.9	12.2	81.7 ± 70.4	67.1	31.3 ± 32.8	20.2	58.9 ± 62.4	36.7	24.3 ± 34.8	12.5	99 ± 195.4	31.7	20.8 ± 28.3	12.5
Morocco	6	57 ± 61.8	30.1	58.9 ± 73.3	31.7	54 ± 57.5	35.7	41 ± 41.8	23.6	47.6 ± 49.9	22.5	67 ± 58.9	44.8	7.8 ± 3.3	7.3
Mozambique	83	41.1 ± 70.9	10.0	70.2 ± 55.6	42.7	41.3 ± 40.3	37.6	70.4 ± 64	47.9	28.7 ± 45	5.8	22.2 ± 37.2	5.8	11.5 ± 5.5	10.9
Myanmar	3	28.5 ± 30.4	16.9	71.1 ± 12.3	67.3	153.9 ± 129.5	87.3	53 ± 20.9	62.6	121.7 ± 51.3	99.4	241.2 ± 35.8	250.1	40.7 ± 45	20.6
namibia	1	4.6 ± 0	4.6	50.6 ± 0	50.6	7.1 ± 0	7.1	202.9 ± 0	202.9	7.9 ± 0	7.9	0.2 ± 0	0.2	40.6 ± 0	40.6
Netherlands	2	2 ± 1.3	2.0	3.3 ± 0.3	3.3	2.2 ± 1.3	2.2	2.9 ± 0.7	2.9	3.7 ± 0.8	3.7	2.4 ± 1.8	2.4	3.8 ± 0.5	3.8
New Zealand	1	10.1 ± 0	10.1	41.6 ± 0	41.6	16.9 ± 0	16.9	6.8 ± 0	6.8	10.2 ± 0	10.2	2.8 ± 0	2.8	15.6 ± 0	15.6
Pakistan	9	66.6 ± 48.4	75.5	103 ± 63.7	81.9	65.7 ± 45.4	57.9	85 ± 65.5	71.1	82.9 ± 54.3	106.3	265.9 ± 337.7	55.5	38 ± 19.8	39.1
Panama	2	10.3 ± 2.6	10.3	17.1 ± 7.3	17.1	28.9 ± 28	28.9	44.3 ± 55.8	44.3	5.8 ± 3.2	5.8	44.7 ± 53.7	44.7	12.7 ± 9.7	12.7
Papua New Guinea	29	30.6 ± 36	10.5	85 ± 76.5	56.2	35.9 ± 35.9	20.8	50.9 ± 48.9	35.8	27.4 ± 29.4	13.9	34 ± 37.3	27.5	43.7 ± 30.1	37.5
Peru	197	29.8 ± 31.3	22.3	71.6 ± 137.3	37.6	29.5 ± 21.9	23.1	49.9 ± 44.9	35.8	30.1 ± 28	22.4	70.4 ± 137.1	27.9	23.8 ± 25.6	16.6
Philippines	34	6.2 ± 5.3	3.7	35.9 ± 33.8	20.6	8.5 ± 6.9	7.1	15.8 ± 9.9	16.3	7 ± 5.9	5.4	129.6 ± 104.7	161.7	3.3 ± 1.7	2.8
Portugal	13	3.3 ± 2.8	2.0	10.8 ± 14.4	7.5	5.2 ± 4.7	3.6	5 ± 4.1	3.2	4.4 ± 3.2	3.8	37.5 ± 119.9	1.4	2.9 ± 1.7	2.5
Republic of the Congo	1	15.3 ± 0	15.3	43.7 ± 0	43.7	15.6 ± 0	15.6	134.6 ± 0	134.6	18 ± 0	18.0	0 ± 0	0.0	24 ± 0	24.0
Réunion	8	5 ± 2.7	4.4	13 ± 6.8	10.9	5.9 ± 2.6	5.1	9.2 ± 4.9	10.6	6.3 ± 3.6	5.2	13.1 ± 14	9.1	5 ± 2.1	4.8

		gpt-4o		gpt-4o-mini		gpt-4-turbo		gpt-3.5-turbo		deepseek-chat		GEOLocate		CountyCentroid	
Country	**Count**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**
Russia	92	40.2 ± 40.7	25.1	174.2 ± 149.6	147.0	76.8 ± 113.3	27.4	111.9 ± 118.7	67.7	54.5 ± 64.9	27.0	2425.9 ± 2541	1218.8	53.9 ± 32.9	45.1
Slovakia	1	5.2 ± 0	5.2	26.5 ± 0	26.5	15.8 ± 0	15.8	13.2 ± 0	13.2	12.5 ± 0	12.5	2.7 ± 0	2.7	14.4 ± 0	14.4
Solomon Islands	1	12 ± 0	12.0	26.3 ± 0	26.3	21.4 ± 0	21.4	21.4 ± 0	21.4	17.3 ± 0	17.3	40.7 ± 0	40.7	7.3 ± 0	7.3
Somalia	28	33 ± 22.4	27.3	92.1 ± 36.6	91.0	58.7 ± 25.3	53.5	56.7 ± 26	51.1	56.6 ± 26	54.8	323.8 ± 276.9	542.9	57.9 ± 18.3	57.3
South Africa	54	14 ± 17.4	5.8	73.9 ± 60	58.0	18.9 ± 20.4	7.5	32.5 ± 31.8	16.8	29 ± 38.6	6.0	77.8 ± 126.7	16.3	15.2 ± 10.2	13.5
Spain	43	12.9 ± 19.1	7.4	39.2 ± 43.2	26.9	12.5 ± 13.4	6.5	23.2 ± 23.4	15.4	17.8 ± 23.6	6.9	25 ± 66.3	3.6	5.2 ± 3.6	3.3
Suriname	3	29.1 ± 20.5	26.7	97.8 ± 43.8	98.2	68.8 ± 58.1	44.3	58.2 ± 50.3	39.2	46.3 ± 34.2	32.3	51.5 ± 7.3	51.5	31.8 ± 28.4	22.8
Sweden	1	4.2 ± 0	4.2	33.9 ± 0	33.9	3 ± 0	3.0	20.9 ± 0	20.9	4.4 ± 0	4.4	5.7 ± 0	5.7	19.5 ± 0	19.5
Tanzania	46	43.2 ± 35	34.3	78.3 ± 46.5	81.5	38.3 ± 31.4	28.0	49.9 ± 33.9	39.0	34.3 ± 31.3	25.4	92 ± 92	41.9	10.3 ± 6.8	7.8
Thailand	155	11.3 ± 11.8	8.3	33.8 ± 33.8	23.7	18.5 ± 19	11.4	28 ± 30.1	16.3	11.6 ± 11.3	8.4	63.4 ± 160	12.1	6 ± 4.6	4.9
Timor-Leste	260	8.6 ± 9.8	5.5	37.6 ± 34.7	24.8	12.7 ± 14.2	9.1	40.1 ± 93.7	11.2	8.8 ± 9.2	5.8	39.3 ± 39.8	30.2	3.1 ± 2.2	2.8
Turkey	86	16 ± 16.6	12.1	48.3 ± 44.3	31.6	21.4 ± 24.5	15.1	32.3 ± 32.1	20.9	14.6 ± 11.1	11.9	77.7 ± 163.5	17.9	15.2 ± 8.4	14.0
United States	878	12.4 ± 24.3	5.1	32.9 ± 46.6	15.6	27.4 ± 330.4	7.6	24.5 ± 37.9	11.6	36.6 ± 306	8.3	33.6 ± 99.4	5.4	56.9 ± 150.2	38.6
Uruguay	6	17.5 ± 8.3	20.6	36.3 ± 15.1	37.8	17.3 ± 9	19.1	34.3 ± 15.9	43.3	13.6 ± 9.6	15.0	28.3 ± 33.1	14.5	11.2 ± 6.2	12.1
Uzbekistan	1	23.9 ± 0	23.9	46.3 ± 0	46.3	60.7 ± 0	60.7	60.1 ± 0	60.1	47.7 ± 0	47.7	46.6 ± 0	46.6	15 ± 0	15.0
Venezuela	7	31.2 ± 48.9	16.5	71.6 ± 90.6	48.1	38.9 ± 48.7	18.3	56 ± 69.1	19.7	27 ± 31.1	14.6	37.5 ± 57.9	17.3	38.4 ± 69	11.8
Vietnam	26	17.5 ± 24.5	12.3	31.1 ± 24.4	24.6	22 ± 24.3	17.0	29.8 ± 28.1	22.1	15.3 ± 18.5	11.5	100.6 ± 176.8	19.2	4.5 ± 2.5	4.3

Table S2 Statistical analysis of the mean, standard deviation (sd), and median (med.) of error distances based on multiple georeferencing methods for the 100 sample points with human georeferencing results.

		gpt-4o		gpt-4o-mini		gpt-4-turbo		gpt-3.5-turbo		o1-preview		deepseek-chat		deepseek-reasoner		GEOLocate		CountyCentroid		human	
Country	**Count**	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.	Mean±sd	Med.
Argentina	1	17.6 ± 0	17.6	208.6 ± 0	208.6	35.5 ± 0	35.5	35.5 ± 0	35.5	1.7 ± 0	1.7	5.2 ± 0	5.2	5.1 ± 0	5.1	14.2 ± 0	14.2	141.4 ± 0	141.4	1 ± 0	1.0
Australia	24	32.9 ± 36.7	15.6	80.8 ± 77.3	62.1	36.3 ± 35.8	21.2	46.1 ± 43	42.9	18.1 ± 19.8	8.1	24.5 ± 30.7	13.8	33.2 ± 43.3	15.2	286.3 ± 540.4	42.9	29.9 ± 33.8	12.6	35.8 ± 133.4	2.8
Bolivia	1	34.8 ± 0	34.8	92.7 ± 0	92.7	60.1 ± 0	60.1	65.8 ± 0	65.8	38.5 ± 0	38.5	16.1 ± 0	16.1	34.6 ± 0	34.6	304.4 ± 0	304.4	11 ± 0	11.0	27.6 ± 0	27.6
Brazil	4	31.2 ± 32.6	18.4	143.2 ± 63.7	160.0	41.8 ± 19.2	48.2	111.3 ± 10.1	114.9	32.2 ± 27.5	19.6	36.6 ± 30	24.4	35.2 ± 23.3	24.1	649.8 ± 32.9	665.8	71.7 ± 49	71.4	80.2 ± 108.3	28.2
Bulgaria	1	2.9 ± 0	2.9	28.9 ± 0	28.9	15.8 ± 0	15.8	24.8 ± 0	24.8	7.6 ± 0	7.6	14.3 ± 0	14.3	14.4 ± 0	14.4	4.1 ± 0	4.1	7 ± 0	7.0	3.4 ± 0	3.4
Cameroon	5	12.3 ± 7.1	10.7	51.4 ± 37.1	43.1	20.1 ± 15.2	19.1	41.3 ± 28.2	41.4	15.2 ± 11.6	16.7	12.7 ± 8.6	12.5	16.8 ± 8.2	18.1	25.2 ± 39.7	12.5	45.7 ± 57.5	26.4	12.3 ± 13.7	7.7
Canada	1	25.3 ± 0	25.3	84 ± 0	84.0	32.6 ± 0	32.6	57 ± 0	57.0	15.7 ± 0	15.7	27.8 ± 0	27.8	68.4 ± 0	68.4	1.4 ± 0	1.4	70.7 ± 0	70.7	4.4 ± 0	4.4
Chile	2	2.4 ± 0.5	2.4	4.8 ± 1.4	4.8	4 ± 0.3	4.0	9.3 ± 1.4	9.3	3.3 ± 0.2	3.3	2.3 ± 2.8	2.3	3.6 ± 0.5	3.6	22.9 ± 1.3	22.9	17.9 ± 0.1	17.9	14 ± 17.8	14.0
China	4	67.7 ± 38	76.8	63.2 ± 35.9	67.3	67.5 ± 27.9	67.1	85.4 ± 71.1	67.3	54.9 ± 48.7	44.9	68.4 ± 49.3	67.3	42.7 ± 21.6	50.7	363.1 ± 347.4	316.9	19.1 ± 15.7	19.9	32.1 ± 46.6	13.5
Colombia	2	5 ± 1	5.0	45.8 ± 3.7	45.8	4.6 ± 0.9	4.6	18.9 ± 22.4	18.9	4.7 ± 0.1	4.7	6.7 ± 2.7	6.7	7.2 ± 1	7.2	485.2 ± 129.6	485.2	44.3 ± 36.1	44.3	23.9 ± 27	23.9
Ecuador	2	21.2 ± 11.6	21.2	66.3 ± 10.6	66.3	42.1 ± 20.9	42.1	75.6 ± 2	75.6	18.8 ± 4.1	18.8	30.1 ± 22.4	30.1	30.4 ± 6.7	30.4	167.7 ± 114.4	167.7	19.8 ± 1.2	19.8	16 ± 11.2	16.0
Gabon	1	18.5 ± 0	18.5	100.9 ± 0	100.9	29.5 ± 0	29.5	96.6 ± 0	96.6	10.2 ± 0	10.2	22.8 ± 0	22.8	18.2 ± 0	18.2	754.3 ± 0	754.3	1.5 ± 0	1.5	28.3 ± 0	28.3
Georgia	6	16.2 ± 11.9	13.1	72.5 ± 20.5	76.3	16.5 ± 13.3	10.8	32.2 ± 28.8	28.8	18.7 ± 15.1	13.8	16.4 ± 9.5	15.1	18.6 ± 14.1	15.7	52.8 ± 99.8	11.1	60.7 ± 81.9	23.1	6.6 ± 6.5	4.7
Guyana	1	58.3 ± 0	58.3	118.3 ± 0	118.3	6.7 ± 0	6.7	82.9 ± 0	82.9	50 ± 0	50.0	80.4 ± 0	80.4	57.5 ± 0	57.5	2.7 ± 0	2.7	31.5 ± 0	31.5	1.7 ± 0	1.7
Kazakhstan	1	1.3 ± 0	1.3	59.2 ± 0	59.2	1.2 ± 0	1.2	32.7 ± 0	32.7	1.8 ± 0	1.8	1.7 ± 0	1.7	2.2 ± 0	2.2	336.1 ± 0	336.1	57.5 ± 0	57.5	0.2 ± 0	0.2
Madagascar	1	4 ± 0	4.0	107.5 ± 0	107.5	3.1 ± 0	3.1	22.5 ± 0	22.5	11.4 ± 0	11.4	9.5 ± 0	9.5	12 ± 0	12.0	84.3 ± 0	84.3	13.1 ± 0	13.1	9.3 ± 0	9.3
Mozambique	2	32.7 ± 13.8	32.7	130.2 ± 134	130.2	45.7 ± 52.7	45.7	151 ± 146.4	151.0	12.5 ± 12.1	12.5	13.4 ± 15.3	13.4	174.6 ± 235.5	174.6	3.2 ± 3.4	3.2	37.6 ± 34	37.6	1.3 ± 1.2	1.3
Peru	4	34.5 ± 15.6	35.1	42 ± 15.2	47.2	20.7 ± 13.4	23.0	47.6 ± 24.4	50.6	27.9 ± 30.8	18.4	23.9 ± 15.6	17.4	28.3 ± 13.1	25.4	30 ± 30.6	29.2	13.2 ± 7.6	11.5	14.6 ± 13.5	11.8
Philippines	1	5.5 ± 0	5.5	112.5 ± 0	112.5	8.9 ± 0	8.9	24.2 ± 0	24.2	4.9 ± 0	4.9	1 ± 0	1.0	0.5 ± 0	0.5	161.4 ± 0	161.4	94 ± 0	94.0	0.1 ± 0	0.1

		gpt-4o		gpt-4o-mini		gpt-4-turbo		gpt-3.5-turbo		o1-preview		deepseek-chat		deepseek-reasoner		GEOLocate		CountyCentroid		human	
Country	**Count**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**	**Mean±sd**	**Med.**
South Africa	2	14.4 ± 13.9	14.4	31.1 ± 34.3	31.1	20.6 ± 20.3	20.6	44.3 ± 58.9	44.3	22.5 ± 24.2	22.5	57.9 ± 74.3	57.9	6.9 ± 3.2	6.9	117.1 ± 33.8	117.1	10.8 ± 13	10.8	7.6 ± 9.6	7.6
Tanzania	2	21.8 ± 19	21.8	141.3 ± 26.8	141.3	21.8 ± 8.9	21.8	89.8 ± 60.2	89.8	21.1 ± 13.5	21.1	9.5 ± 8.2	9.5	36.2 ± 26.9	36.2	29.5 ± 34.3	29.5	32.6 ± 10.9	32.6	19.7 ± 21.1	19.7
Thailand	4	4.9 ± 4.9	4.7	16.4 ± 18.9	10.8	7.2 ± 7	5.7	13.1 ± 17.7	6.4	5.1 ± 5.1	4.1	5.7 ± 5.1	5.6	3.9 ± 4.6	2.4	12 ± 11	8.8	24.1 ± 31.4	10.8	5.6 ± 5.2	3.5
Timor-Leste	8	7.3 ± 6.9	5.3	33.2 ± 40	15.2	17.2 ± 18.5	14.3	63.5 ± 135.7	17.1	7.6 ± 6.3	6.5	12.4 ± 11.9	7.3	2.9 ± 2.8	2.0	27.9 ± 27.8	21.1	23.1 ± 16.9	20.5	4 ± 2.8	4.1
Turkey	2	7.7 ± 5.8	7.7	66.3 ± 77.8	66.3	11.6 ± 5.7	11.6	9.9 ± 3.3	9.9	14.2 ± 17	14.2	11.2 ± 5.3	11.2	9.6 ± 6	9.6	29.1 ± 30.5	29.1	2.6 ± 2.3	2.6	30.5 ± 36.5	30.5
United States	18	10.2 ± 15.2	6.1	44.4 ± 57	29.1	13.9 ± 12.4	9.9	24.5 ± 25.9	19.5	11.9 ± 13.5	9.5	20.1 ± 19.5	13.2	17.6 ± 19.8	11.8	12.8 ± 27.6	3.8	47.5 ± 44.1	29.0	4.1 ± 6.9	1.3

Table S3 The process of reasoning of *DeepSeek-reasoner* for the three georeferencing examples in Figure 1a.

Locality	**Reasoning process**
Pony Express Trail, 0.7 mile east of its junction with Sly Park Road in Pollock Pines **(California, USA)**	Okay, let's see. I need to georeference this location based on the given information. The input is ["United States", "California", "El Dorado", "Pony Express Trail, 0.7 mile east of its junction with Sly Park Road in Pollock Pines"]. First, breaking down the components. The country is the US, state California, county El Dorado, and the locality details are Pony Express Trail, 0.7 mile east of its junction with Sly Park Road in Pollock Pines. Since the priority is locality > county > state > country, I should focus on the most specific info first. The locality mentions a specific point along Pony Express Trail, 0.7 miles east of where it meets Sly Park Road in Pollock Pines. Pollock Pines is a community in El Dorado County. I'll start by finding the coordinates of the junction between Pony Express Trail and Sly Park Road in Pollock Pines. Let me check a map. Pollock Pines is in California, El Dorado County. Sly Park Road is a known road there. Pony Express Trail intersects with Sly Park Road. Once I have that intersection point, I need to go 0.7 miles east along Pony Express Trail from there. Wait, wait. The description says 0.7 mile east of the junction. But if the trail runs east-west, then moving east from the junction would be along the trail. But first, I need the exact coordinates of the junction. Let me search for the intersection of Pony Express Trail and Sly Park Road in Pollock Pines, CA. Looking it up, the intersection is at approximately 38.7667° N, 120.5866° W. Now, moving 0.7 miles east along Pony Express Trail from there. One mile is about 1.60934 kilometers, so 0.7 miles is roughly 1.126 km. However, converting this distance to decimal degrees isn't straightforward. But since the direction is east, which affects longitude. The exact change in longitude depends on the latitude. At 38.7°N, each degree of longitude is approximately 88.6 km. So 1.126 km east would be roughly 1.126 / 88.6 ≈ 0.0127 degrees. So adding that to the longitude: -120.5866 + 0.0127 ≈ -120.5739. But wait, this is a rough estimate. Maybe there's a more precise way. Alternatively, using a mapping tool like Google Maps to find the intersection and then measure 0.7 miles east. Let me check. Searching for Sly Park Rd and Pony Express Trail, Pollock Pines, CA. The intersection is at 38.766667, -120.586667. Moving east along Pony Express Trail for 0.7 miles. Using Google Maps' measure distance function. Starting at the intersection, heading east on Pony Express Trail. At 0.7 miles, the coordinates are approximately 38.7663, -120.5733. So the latitude is roughly 38.7663, longitude -120.5733. But I should confirm if this is accurate. Alternatively, check if there are any landmarks or addresses that match this point. Alternatively, maybe there's a specific address or a marker. But since the description is about a point along the trail, the calculated coordinates based on the junction and distance would be the way to go. Therefore, the coordinates would be approximately [38.7663, -120.5733]. But I need to format them as floats in a list. Let me check again. The original intersection is at 38.766667, -120.586667. Moving east 0.7 miles would decrease the longitude (since west is negative). Wait, east direction would mean increasing longitude (since west longitude is negative). So moving east would subtract from the negative value? Wait, longitude west is negative. So moving

Locality	Reasoning process
	east from a point at -120.5866 would mean adding to the longitude (making it less negative). So 0.7 miles east would be -120.5866 + (0.7 miles in degrees). Calculating again: 1 degree of longitude at 38.7667°N is approximately cos(38.7667) * 111.32 km ≈ 0.785 * 111.32 ≈ 87.3 km per degree. So 0.7 miles is 1.126 km. 1.126 / 87.3 ≈ 0.0129 degrees. So the new longitude is -120.5866 + 0.0129 ≈ -120.5737. So the coordinates would be approximately 38.7667 (latitude remains roughly the same if moving east along a straight trail) and longitude -120.5737. So the final coordinates would be [38.7667, -120.5737]. But checking via Google Maps, perhaps the exact point is slightly different. Let me try to find a more precise measurement. Alternatively, perhaps using a geocoding API or tool would get the exact coordinates, but since I can't do that here, relying on estimation. However, based on the information given and the calculation, the latitude and longitude should be around 38.7667, -120.5737. Rounding to three decimal places as in the example (Tanzania had 3 decimals), so [38.767, -120.574]. Wait, but the example had [-6.192, 36.006], which is three decimal places. So maybe I should present three decimals. But the initial intersection was at 38.766667, -120.586667. After moving east 0.7 miles, the longitude becomes approximately -120.5737, which would be -120.574 when rounded to three decimals. Latitude remains about 38.7667 (since moving east along a trail that's relatively straight might not change latitude much). So perhaps [38.767, -120.574]. Alternatively, checking if there's any specific landmark at that location. For example, if the Pony Express Trail at that point has a marker or a known coordinate. But without access to real-time data, I'll proceed with the calculated estimate. So the final answer is [38.767, -120.574].
Kissena Park, 0.3 mi W of 164th Street, ca. 0.1 mi N of Booth Memorial Avenue **(New York, USA)**	Okay, let's tackle this georeferencing problem. The input is ["United States", "New York", "Queens", "Kissena Park, 0.3 mi W of 164th Street, ca. 0.1 mi N of Booth Memorial Avenue"]. First, I need to prioritize the most specific information. The locality mentions Kissena Park, which is a specific park in Queens, New York. The other details are directions from streets, which should help pinpoint the exact spot. Starting with Kissena Park. I know that Kissena Park is a real park in Queens, so I'll look up its coordinates. A quick check shows Kissena Park's general coordinates are around 40.741°N, -73.803°W. But the description gives more precise directions: 0.3 miles west of 164th Street and approximately 0.1 miles north of Booth Memorial Avenue. Next, I need to find the intersection of 164th Street and Booth Memorial Avenue in Queens. Let me verify their locations. 164th Street runs north-south, and Booth Memorial Avenue is east-west. The intersection point would be at some coordinates, say around 40.7375°N, -73.795°W as a rough estimate. Now, moving 0.3 miles west from 164th Street. Since 164th Street is a north-south road, moving west would decrease the longitude. 0.3 miles is roughly 0.003 degrees in longitude (since 1 mile ≈ 0.0145 degrees at this latitude). So subtracting 0.003 from -73.795 gives approximately -73.798. Then, moving 0.1 miles north from Booth Memorial Avenue. Assuming Booth

Locality	Reasoning process
	Memorial is east-west, moving north increases latitude. 0.1 miles is about 0.00145 degrees. Adding that to 40.7375 gives around 40.73895. But wait, the original park's coordinates are 40.741, -73.803. The calculated point is slightly south and east of the park's general area. Maybe the specific location is within the park boundaries. Alternatively, perhaps the directions are more precise. Another approach: use geocoding services. Searching for "Kissena Park, 0.3 mi W of 164th St, 0.1 mi N of Booth Memorial Ave, Queens, NY" might yield exact coordinates. Let me check. Looking up the address details, 164th Street and Booth Memorial Avenue intersect in Fresh Meadows, Queens. Moving west 0.3 miles from there would be towards Kissena Park. A point 0.1 mi north of Booth Memorial along that western direction would place it inside the park. Using Google Maps, the intersection of 164th St and Booth Memorial Ave is at 40.7355, -73.7938. Moving west 0.3 mi (approx 0.0043 degrees) would be 40.7355, -73.7938 - 0.0043 = -73.7981. Then 0.1 mi north (0.00145 degrees) gives latitude 40.7355 + 0.00145 ≈ 40.73695. Wait, but Kissena Park's main area is around 40.741, -73.803. The calculated coordinates here are south and east of that. Maybe the directions were from a different part of 164th St? Or perhaps I should use the park's main coordinates as the primary and adjust slightly based on the directions. Alternatively, maybe the description refers to a specific point within the park. Alternatively, check if there's a specific landmark or point in Kissena Park matching the description. Alternatively, perhaps using a geocoding API would give precise coordinates. Since I can't access APIs, I'll have to approximate. Given that the park's main coordinates are 40.741, -73.803, and the directions from 164th St and Booth Memorial Ave, which is to the east of the park, moving west 0.3 mi would place the point within the park. So adjusting the longitude from 164th St's longitude (say -73.7938) west by 0.3 mi (which is 0.0043 degrees) gives -73.7981. Latitude: starting at 40.7355 (Booth Memorial) and moving 0.1 mi north gives 40.73695. But that's still south of the park's main area. Hmm, maybe the starting intersection is different. Alternatively, maybe the 164th Street referred to is further north. Let me check another part of 164th St. For example, if 164th St near Kissena Blvd, the intersection might be at 40.741, -73.803. Wait, no, 164th St runs through different areas. Alternatively, perhaps the user's description is within Kissena Park, so the main park coordinates are sufficient, adjusted slightly. But the directions are specific. Alternatively, use the park's centroid as a fallback. But the problem requires using the most precise info. Another way: 164th Street in Queens. Let me check 164th St and Booth Memorial Ave. On maps, the intersection is at 40.7355, -73.7938. Moving 0.3 mi west (which is about 0.3 * 1.60934 km ≈ 0.48 km). West in longitude is negative, so subtract 0.0043 degrees (since 1 degree ≈ 111 km, 0.48 km is ~0.0043 degrees). So longitude becomes -73.7938 - 0.0043 = -73.7981. Then 0.1 mi north (0.1 * 1.60934 ≈ 0.16 km, which is 0.00145 degrees north). Latitude becomes 40.7355 + 0.00145 ≈ 40.73695. But this point (40.73695, -73.7981) is in the vicinity of Kissena Corridor Park, which is south of Kissena Park. However, the locality mentions Kissena Park

Locality	Reasoning process
	specifically. Maybe there's an error in the starting intersection. Alternatively, perhaps the directions are within the park's boundaries. Alternatively, maybe the correct coordinates are the park's main area, around 40.741, -73.803. But the directions suggest a specific spot. However, without precise data, perhaps the best approximation is to take the park's coordinates. But given the instructions to use the most precise info, maybe the calculated coordinates based on the directions are better. Alternatively, check if there's a more accurate geocode. For example, searching for "Kissena Park 164th Street Booth Memorial Avenue" might find a specific location. However, based on available data, the coordinates derived from moving west and north as described would be approximately (40.73695, -73.7981), but I'm not entirely confident. Alternatively, maybe the user intended the main park area, but with the directions, it's better to use the adjusted coordinates. Alternatively, perhaps the correct coordinates are (40.741, -73.803) as the park's main point, but adjusted slightly. Wait, 0.3 mi west of 164th St: if 164th St is at -73.7938, then 0.3 mi west is -73.7981, and 0.1 mi north from Booth Memorial (40.7355) gives 40.73695. But this is outside the main park area. So perhaps the locality is in the park but the directions are relative to a nearby intersection. Alternatively, maybe the directions are within the park. For example, if 164th Street is on the east side of the park, moving west into the park. But without precise street data, it's challenging. Given the ambiguity, but following the instructions to prioritize the most accurate info, the final coordinates should be based on the directions given. So the answer would be approximately [40.737, -73.798], but I need to check if that's within Kissena Park. Looking at a map, Kissena Park spans roughly from 40.735 to 40.745 in latitude and -73.81 to -73.79 in longitude. The calculated point (40.737, -73.798) is within that range, so it's plausible. Alternatively, the exact coordinates might be better represented as 40.7369, -73.7981. Rounding to three decimal places, that's [40.737, -73.798]. But wait, the example output had two decimal places, but the user specified "a Python list of 2 float numbers". The example had three decimals, like [-6.192, 36.006], so three decimals are acceptable. Final answer: [40.737, -73.798]. Let me verify once more. If 164th and Booth Memorial is at 40.7355, -73.7938, then west 0.3 mi is -73.7981, north 0.1 mi is 40.73695. So rounding to three decimals: latitude 40.737, longitude -73.798. That should be correct.
Robertons Knob FireTrail, approx 800m from carpark atlocked gate towards Wrights LO, New England National Park. **(New South Wales,**	Okay, let's tackle this georeferencing problem. The input is a list with four elements: country (Australia), state (New South Wales), county (Kempsey), and a detailed locality description. The task is to find the latitude and longitude coordinates based on the most accurate information available. First, the priority is locality, then county, state, and country. So I should focus on the locality description first. The locality given is "Robertons Knob Fire Trail, approx 800m from carpark at locked gate towards Wrights LO, New England National Park." I need to parse this. The key elements here are New England National Park, which is in New South Wales. The specific trail mentioned is Robertons Knob Fire Trail. The mention of being approximately 800 meters from the carpark at a locked gate towards Wrights LO (which I assume is a lookout or some

Locality	Reasoning process
Australia)	landmark) gives a specific point within the park. To find coordinates, I might start by looking up New England National Park in New South Wales. The park is a known area, so its general coordinates can be found. But the exact fire trail and the specific location 800m from the carpark might not be straightforward. Alternatively, maybe there's a known point for Robertons Knob itself. Searching for Robertons Knob in New England National Park might yield coordinates. If that's not available, perhaps the Wrights Lookout (assuming LO stands for Lookout) is a known point. The fire trail's starting point or the carpark location could be reference points. I can check online maps or geographic databases. Let's try to think of possible steps: 1. Search for "Robertons Knob Fire Trail, New England National Park, NSW, Australia" to see if there's a trail with coordinates. If the trail is a named feature, maybe it's recorded in mapping services like Google Maps, OpenStreetMap, or government GIS databases. 2. If the trail isn't found directly, look for New England National Park's visitor information or official park maps that might mention Robertons Knob Fire Trail. The park's management might have detailed maps with trails and landmarks. 3. The mention of Wrights LO (Lookout) could be a key point. If Wrights Lookout has known coordinates, then moving 800m from the carpark towards that lookout along the fire trail might help approximate the location. However, without knowing the exact starting point (the carpark), this could be tricky. 4. The carpark at the locked gate is another reference. If the coordinates of the carpark are known, then moving 800m towards Wrights Lookout from there along the fire trail would give the approximate location. But this requires knowing the carpark's position and the trail's direction. Since the exact coordinates might not be readily available, perhaps the best approximation is the general area of Robertons Knob within the park. Checking geographic databases for Robertons Knob in New England National Park. Let me think: a quick search (hypothetically, since I can't browse) might show that Robertons Knob is a peak in the park. If so, its coordinates could be around -30.50 latitude and 152.40 longitude. But I need to verify this. Wait, New England National Park is located in the Northern Tablelands of New South Wales. Major landmarks include Point Lookout, which is a well-known spot. But Robertons Knob might be a lesser-known peak. Alternatively, perhaps the fire trail is near Wrights Lookout. If Wrights Lookout has coordinates, then moving 800m from the carpark towards it would be approximate. Alternatively, perhaps the main entrance or a known trailhead for Robertons Knob Fire Trail has coordinates. For example, if the carpark's coordinates are available, then moving 800m along the trail towards Wrights Lookout would give the point. But without specific data, this is challenging. In cases where exact coordinates can't be found, sometimes the centroid of the park or the nearest named feature is used. However, the user instructions specify prioritizing the most accurate information. Since the locality mentions a specific trail and distance from a carpark, ideally we need to find that exact spot. If that's not possible, defaulting to the park's coordinates might be a last resort, but that's less accurate. Alternatively, perhaps there are existing datasets or research papers that mention

Locality	Reasoning process
	Robertons Knob Fire Trail. For example, ecological studies conducted in the area might provide GPS coordinates for specific trails or points. Assuming that after some research, the coordinates for the locked gate carpark are found, and then moving 800m towards Wrights Lookout along the fire trail. For example, if the carpark is at -30.541, 152.401, then moving northwest (assuming the trail direction) 800m might lead to approximately -30.535, 152.393. But this is just a hypothetical example. Alternatively, if Robertons Knob itself is a known geographic feature, its coordinates might be recorded. For instance, Robertons Knob could be at -30.523, 152.386. Then the fire trail would be near that peak. In the absence of precise information, the best approach is to find the most accurate available data. If the fire trail is not mapped, then using the coordinates of New England National Park's centroid might be a fallback, but that's less precise. However, given that the locality mentions a specific trail and distance from a carpark, perhaps the approximate coordinates can be found through existing maps. After checking (hypothetically), let's say that New England National Park's visitor center or a prominent landmark like Wrights Lookout is at -30.4915, 152.4092. The Robertons Knob Fire Trail might start near there. If the carpark is at the locked gate near Wrights Lookout, then 800m along the trail towards the lookout would be in that vicinity. Adjusting the coordinates accordingly, maybe the latitude is around -30.49 and longitude 152.40. But this is a rough estimate. Alternatively, looking up "Robertons Knob Fire Trail coordinates" might yield a specific point. Suppose a mapping service like OpenStreetMap has the trail mapped. For example, a node on the trail could be at -30.543, 152.398. That would be the coordinates to use. Given that the user wants only the output list without explanation, after doing the research (in reality, using APIs or databases), the final coordinates would be determined. Since I can't access real-time data, I have to rely on existing knowledge or common coordinates for that area. Wait, New England National Park is centered around approximately -30.49, 152.41. But Robertons Knob Fire Trail might be in a different part. Another approach: Kempsey is the county, but New England National Park is actually in the Northern Tablelands, which is more near Armidale, so perhaps the coordinates are further inland. Maybe around -30.5, 152.4. Alternatively, checking the exact location: New England National Park is located near the town of Ebor. Key landmarks include Point Lookout. However, Robertons Knob might be a specific peak within the park. Let me try to recall if Robertons Knob is a known peak. If so, its coordinates might be available. After checking, suppose Robertons Knob is recorded at -30.5365, 152.3852. The fire trail would be near there. The description mentions approx 800m from the carpark at locked gate towards Wrights LO. If the carpark is at the start of the trail, then 800m along the trail towards Wrights Lookout (assuming LO is Lookout) would be the point. If the carpark is at the base, then moving 800m towards the lookout would be in the direction of the trail. Alternatively, if Wrights Lookout is a known point, say at -30.530, 152.390, then the carpark might be 800m southwest of that. But without exact data, it's hard to say.

Locality	Reasoning process
	In such cases, the best possible coordinates would be the approximate center of the park or the nearest known landmark. However, given the specific trail name, if available, use that. Assuming after research, the coordinates for Robertons Knob Fire Trail's approximate location based on available data are [-30.5365, 152.3852], then that's the answer. But I need to ensure that the latitude and longitude are in the correct order and formatted as floats.