Conceptual Vagueness in Human and LLM Representations: Evidence from Migration Distance Categorization

Songlin Wang[1]
Department of Geography and Regional Research, University of Vienna, Austria

Krzysztof Janowicz
Department of Geography and Regional Research, University of Vienna, Austria

Ailin Benitez Cortes
Department of Geography and Regional Research, University of Vienna, Austria

Marion Borderon
Department of Geography and Regional Research, University of Vienna, Austria

Yingjing Huang
Department of Geography and Regional Research, University of Vienna, Austria

Mina Karimi
Department of Geography and Regional Research, University of Vienna, Austria

Zilong Liu
Department of Geography and Regional Research, University of Vienna, Austria

Annika Süß
Department of Geography and Regional Research, University of Vienna, Austria

Patrick Sakdapolrak
Department of Geography and Regional Research, University of Vienna, Austria

Abstract

Conceptual vagueness is a fundamental characteristic of many geographic categories. The varied definitions of "long-distance" and "short-distance" in migration studies provides a typical example. Although interpretations of these categories are inherently context-dependent and heterogeneous, they remain widely used in domain research. This study investigates the extent to which different individuals reach a consensus on conceptualizing these categories. With a migration related corpus where "long" and "short" are generally not explicitly mentioned, we elicit human and LLM judgments in classifying each article into these two categories. Both human-human and human-LLM agreements are quantified. Our results show that human agreements are only moderate, confirming substantial conceptual variability among informed readers. LLM outputs exhibit a comparable level of variability, failing to converge on a stable category boundary. This disagreement arises from multiple sources, from detection of implicit signals to their interpretations. These findings demonstrate the feasibility of using LLMs as probes to characterize conceptual vagueness in geography at scale. Returning to migration distance, we highlight that they should not be treated as universally shared categories in knowledge organization and representation. In contrast, resolving such vagueness requires richer contextual information or additional semantic signals.

2012 ACM Subject Classification Information systems → Geographic information systems; Theory of computation → Categorical semantics

Keywords and phrases qualitative categories, conceptual vagueness, human migration, large language models, agreement

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.21

[1] Corresponding author

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 21; pp. 21:1–21:9

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Category Short Paper

Supplementary Material *Software (Source Code)*: `https://github.com/SonglinWang1111/Long-Short-Distance-Conceptualization` [17]
archived at `swh:1:dir:68cffc055f1b2bcd60c8893eb122fb83588c7fcf`

1 Introduction

Philosophical investigations has long recognized that many concepts and categories do not admit necessary and sufficient definitions. Such vagueness is often regarded as arising from intrinsic properties of human cognition [1]. One influential account is Wittgenstein's notion of "family resemblance", arguing that categories may lack fixed defining features, while features of diverse categories remain overlapping similarities [19]. Similarly, "prototype theory" further claims that categories exhibit graded membership centered around prototypical instances, with other entities organized progressively further, and their borderlines are generally vague [12]. In GIScience, many categories routinely used to represent spatial phenomena also lack precise and universally agreed-upon conceptualizations [9, 14]. Such categories may challenge semantic interoperability, leading to inconsistencies in spatial knowledge organization and representation, information retrieval, and policymaking [2, 6].

This vagueness is also evident in the distinction between "long-distance" and "short-distance" human migration. Since early work such as Ravenstein's laws of migration [11], these terms have been extensively used to describe and analyze mobility patterns. Similar to many other vague categories, their conceptualizations are affected not only by physical magnitudes, but also by cognitive factors such as regional background, culture, society, required time, and accessibility [5, 10, 13, 16]. Consequently, researchers have attempted to investigate how these categories are conceptualized; however, they generally sought to identify optimal metric thresholds [18]. Yet, a more fundamental question remains under-explored: to what extent people actually converge in applying these categories? Answering this question should be useful to uncover the underlying logic that triggers people's decision making.

Recent developments in large language models (LLMs) [3] offer a new methodological opportunity to dive into this research gap. Trained on massive corpora, LLMs encode patterns of language use, therefore, they can be used as statistical mirrors that reflect patterns present in human-generated discourse [15]. By aggregating LLM responses across repeated sampling, it is possible to reflect how concepts are generally represented. Besides, they have demonstrated the ability to extract, summarize, and reason over implicit knowledge in unstructured texts with limited examples of the expected input-output pairs without further training [20]. This makes LLMs promising tools for simulating human understanding of migration types through learning from recurring patterns in given examples to further categorize human migration.

In this study, we investigate how humans and LLMs categorize migration distance when it is described in text (i.e., not as numeric distance values) where, in general, neither "long-distance" nor "short-distance" is explicitly mentioned by authors. We prepare a corpus of migration-related articles, and ask human raters to classify each of them. In parallel, we prompt LLMs using few-shot examples to perform the same task. We then quantify agreements to assess consistency and variability across groups. We emphasize that neither humans nor LLMs are treated as sources of ground truth, but both are only used to reflect disagreement patterns. Meanwhile, we also emphasize that demonstrating migration categories are vague, which is a fact long recognized, is also not our purpose. Rather, we ask a different question: To what extent do humans and LLMs converge with each other when navigating through conceptually vague migration distance categories?

2 Methods

2.1 Data

This study is built upon a systematic literature review in the migration domain, which identified 53 empirical and case-study papers on "climate migration". As part of that review process, the authors initially categorized each article with respect to migration distance type (i.e., "long-distance", "short-distance", "both", or "neither"). We use the reviewed papers as our corpus and the authors' initial categorization to design the prompts and compute agreements. To ensure computational feasibility while preserve diverse contexts, we sample 33 articles from the corpus in which migration distance categories are generally not explicitly mentioned. Three of them are randomly selected as fixed in-context examples, and the remaining 30 articles are used for comparison.

2.2 Preprocessing

Before the analysis, each article is converted from PDF to structured text stored as Python dictionaries, to ensure consistent input across models, as direct PDF ingestion is not uniformly supported. Text extraction is performed using `scipdf_parser`[2]. Figures, tables, and maps are excluded in this conversion, so the categorization relies solely on textual information. We acknowledge that this may introduce slight bias because such visual elements are accessible to humans, although they rarely present explicit distance information.

2.3 Experiment Design

To assess inter-rater agreement between humans, three additional domain researchers independently categorize the selected articles without the assistance of LLMs. In parallel, we query three different LLMs under multiple settings summarized in Table 1. Both human raters and LLMs assign each article to one of four mutually exclusive categories: "long-distance migration", "short-distance migration", "both", or "neither". This scheme matches the original review and the prevailing practice in migration studies.

Table 1 Experimental settings.

Feature	Setting(s)	Rationale
Language models	GPT-4.1-mini; DeepSeek-Chat-V3.2; Gemini-2.5-Flash	Exclude model-specific influence
Prompting approaches	0-shot; 1-shot; 3-shot	Elicit implicit decision patterns
Temperatures	$T = 0;\ 0.4;\ 0.8;\ 1.2;\ 1.6$	Higher T may surface less dominant signals
Iteration	10	Mitigate the influence of stochastic variation or hallucination

Few-shot prompting approaches is used to test whether recurring patterns associated with human categorizations can be detected and reproduced by LLMs. The three fixed examples with their initial labels are provided as paired demonstrations. For the 1-shot condition, one fixed example of those three is selected to maintain coherency. Listing 1 presents the core instruction used across models.

[2] `https://github.com/titipata/scipdf_parser`

Listing 1 Core instruction in the prompt.

```
Please classify the article based on the type of human migration it
addresses by assigning one of the following codes:
'1': It focuses on or observes evidence of long-distance human
migration.
'2': It focuses on or observes evidence of short-distance human
migration.
'3': It focuses on or observes evidence of both long-distance and
short-distance human migration.
'0': It does not mention anything about human migration distance.
Do not infer any missing information. Output only the single
character without explanations, opinions or meta-comments.
```

2.4 Agreement Computation

We quantify agreement at two levels: (1) inter-rater agreements between human raters, and (2) human-model agreements between the initial reviewers and each LLM. Consistent with Section 1, neither humans nor LLMs are treated as sources of ground truth. Agreement therefore serves as a measure of consistency in categorization, not predictive correctness. For any two agents a and b, the simple agreement is defined in Equation 1, where N is the number of articles, I is the number of iterations (for human-human agreements, $I = 1$), $y_i^{(a)}$ and $y_i^{(b)}$ are the labels assigned to the same article at the same iteration by agents a and b, respectively, and $\mathbb{I}()$ is the indicator function, which equals to 1 when the two categorizations are identical and 0 otherwise.

$$\text{Agreement}(a, b) = \frac{1}{N \times I} \sum_{i=1}^{N \times I} \mathbb{I}(y_i^{(a)} = y_i^{(b)}) \tag{1}$$

For agreements between human raters, we further compute Cohen's kappa coefficient [4] which adjusts for chance agreement. Equation 2 illustrates the calculation of κ, where p_o is the observed relative agreement between raters, and p_e is the hypothetical probability of chance agreement.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{2}$$

Simple agreement ranges from 0 to 1, while Cohen's κ ranges from -1 to 1. Higher values indicate strong consistency in how category boundaries are applied.

For LLM-generated classifications, since each configuration is repeated by 10 times, aggregated agreements are reported. Given four possible categories, uniform random assignment yields an expected agreement of 25%, which serves as a baseline.

To examine the structure of disagreement, we also compute multiclass confusion matrices. These matrices help distinguish structured disagreement patterns from degenerate response distributions (e.g., overuse of a single category) and support the interpretation of agreement values as empirical manifestations of conceptual vagueness rather than response collapse.

3 Results

Table 2 summarizes the pairwise agreements and Cohen's κ coefficients among the four human raters. Pairwise agreements range from 0.40 to 0.70, while Cohen's κ ranges from 0.19 to 0.58. Although all four participants are researchers familiar with human migration, none of the pairwise comparisons reaches near-perfect agreement. These results suggest that migration distance categories are not applied according to universally shared criteria, even among domain experts. Rather than being uniformly distributed, disagreements are concentrated in a specific subset of articles, indicating that conceptual ambiguity is context-dependent. Such disagreements arise from multiple recurring forms of conceptual vagueness rather than from a single source. First, regarding several articles, some human raters consider certain linguistic signals to be sufficient in representing migration distance, while others do not think so. The second form concerns the uncertainty about where the conceptual boundary between migration distance categories should be drawn. In such articles, the contextual evidence remains compatible with multiple interpretations. Factors such as geographic scale, regional background, migration purpose, accessibility, etc. lead different raters to classify the same migration event differently.

Table 2 Pairwise agreements and Cohen's κ coefficients among the four human raters.

Human pair	Agreement	Cohen's κ
R1-R2	0.70	0.58
R1-R3	0.53	0.29
R1-R4	0.40	0.19
R2-R3	0.47	0.27
R2-R4	0.47	0.27
R1-R3	0.50	0.33

We next compare human categorizations with classifications generated by LLMs under different prompting strategies and temperatures. Figure 1 presents the agreements between each model and the original reviewers, whose categorizations are also used as demonstrations for few-shot prompting. As this figure presents, across the three models, human-model agreements remain within a range comparable to the agreements observed among human raters. In particular, GPT-4.1-mini consistently achieves agreement values similar to the human inter-rater range. These results indicate that these LLMs reproduce patterns of categorization variability that resemble those observed among human experts.

This figure also illustrates that agreement patterns remain relatively stable across temperature settings, with only minor fluctuations. Increasing sampling diversity therefore produces little influence on the final categorizations, suggesting that the observed disagreements cannot be explained primarily by stochastic generation. Instead, the variability appears to originate from the underlying conceptual ambiguity of the categorization task itself. Prompting strategy exhibits a more noticeable influence. Compared with zero-shot prompting, one-shot prompting frequently reduces agreement, whereas three-shot prompting generally restores agreement to levels comparable to or higher than the zero-shot baseline. This observation suggests that isolated demonstrations may encode localized decision patterns that do not generalize reliably, whereas multiple demonstrations provide more stable guidance regarding the implicit regularities underlying human categorization.

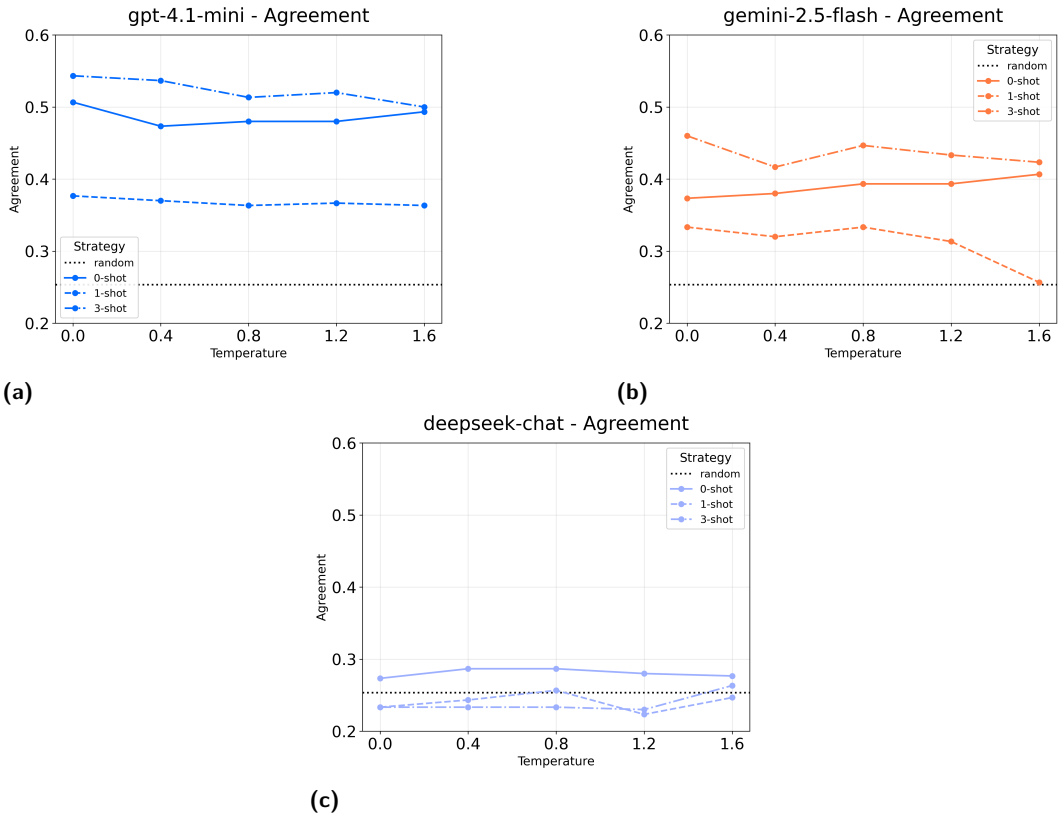

Figure 1 Human-model agreements under different temperatures and prompting strategies: (a) Human-GPT; (b) Human-Gemini; (c) Human-DeepSeek.

The qualitative observations of disagreement patterns among human raters are further supported by the confusion matrix shown in Table 3, which compares the classifications produced by GPT-4.1-mini with those of the original reviewers. Similarly, two dominant disagreement patterns can be observed. First, a considerable proportion of articles labeled as "neither" by human reviewers are classified into one of the other three migration distance categories by GPT-4.1-mini, while articles labeled as "both" by GPT are frequently not classified as this category by human reviewers. This finding indicates that agents sometimes disagree on whether sufficient evidence exists to invoke migration distance categories. Second, disagreements also occur among "long distance" and "short distance", regarding boundary of categories. Taken together, these patterns further support the interpretation that disagreement reflects structured conceptual variability rather than random prediction errors or degenerate response distributions.

Table 3 Confusion matrix comparing classifications of the initial reviewers and GPT-4.1-mini.

Human \ Model	Neither	Long distance	Short distance	Both
neither	65	130	361	344
long distance	0	251	22	177
short distance	163	233	1193	511
both	108	120	265	557

4 Discussion and Conclusion

This study moves beyond the traditional pursuit of fixed thresholds for distinguishing "long-distance" and "short-distance" migration by examining how consistently these categories are actually applied. Rather than asking whether migration distance categories are vague, which is a fact long recognized in philosophy, cognitive science, and GIScience, we investigate the extent to which humans and LLMs converge when interpreting them from textual descriptions. Across both human-human and human-LLM comparisons, moderate agreements are observed, indicating that these categories are not governed by universally shared conceptual criteria. Moreover, our observation reveals that disagreement originates from multiple forms of vagueness, regarding both the applicability of linguistic cues and the conceptual boundary between categories. These findings have implications beyond migration studies. In geographic knowledge representation, categories are often considered to possess stable and universally shared meanings. However, our results suggest that the semantics associated with "migration distance" remain inherently context dependent. Consequently, representing such kind of categories as rigid classes or fixed ontology concepts risks introducing "artificial precision" and masking legitimate conceptual diversity. Instead, geographic knowledge representation should accommodate graded membership here, considering contextual evidence, rather than assuming a single authoritative categorization.

Our findings also contribute methodologically. Rather than evaluating LLMs as classifiers, we demonstrate their potential as computational probes for conceptual analysis. While LLMs inevitably introduce biases [7, 8], several observations support our interpretation. First, similar levels of disagreements emerge among human experts and between humans and LLMs. Second, disagreement exhibits systematic structures instead of random response patterns, as reflected by both confusion matrices and qualitative analysis. Finally, agreement remains relatively stable across different temperatures, suggesting that the observed variability cannot be attributed primarily to stochastic model behavior. Together, these observations indicate that LLMs capture meaningful regularities in human conceptualizations and therefore provide a scalable methodology for investigating conceptual vagueness in geographic discourse.

Despite these contributions, several limitations remain. First, although the corpus contains linguistically rich and diverse migration narratives, it remains relatively small. Future work should investigate whether similar agreement patterns persist across substantially larger collections. Second, our corpus consists exclusively of peer-reviewed scientific publications. The conceptual patterns identified here may not generalize directly to other forms of discourse such as news reports and social media posts. Examining conceptual vagueness across multiple discourses would provide a more comprehensive understanding of how migration distance categories are socially constructed and communicated. Finally, while we identify several recurring forms of disagreement, the contextual factors underlying these patterns remain only partially understood. The exact "traces" linked to socio-cultural or infrastructural factors should be systematically investigated in future works.

In conclusion, this study provides empirical evidence that migration distance categories remain only partially shared among both humans and LLMs. Rather than indicating failures of either human or language models, the observed variability reflects the inherently vague nature of such conceptualizations. By combining human annotation with LLM outputs, we present a scalable framework for investigating conceptual vagueness. This work encourages future research toward more context-aware knowledge representations that acknowledge, rather than eliminate, conceptual uncertainty.

References

1 Ahmed Arara and Djamal Benslimane. Towards formal ontologies requirements with multiple perspectives. In Henning Christiansen, Mohand-Saïd Hacid, Troels Andreasen, and Henrik Legind Larsen, editors, *Flexible Query Answering Systems*, volume 3055, pages 150–160. Springer Berlin Heidelberg, 2004. doi:10.1007/978-3-540-25957-2_13.

2 Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pages 1–6. ACM, 2018. doi:10.1145/3227609.3227689.

3 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. Version Number: 4. doi:10.48550/arXiv.2005.14165.

4 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960. URL: https://api.semanticscholar.org/CorpusID:15926286.

5 Marián Halás and Pavel Klapka. Revealing the structures of internal migration: A distance and a time-space behaviour perspectives. *Applied Geography*, 137:102603, 2021. doi:10.1016/j.apgeog.2021.102603.

6 Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. Geospatial semantics and linked spatiotemporal data – past, present, and future. *Semantic Web*, 3:321–332, 2012. doi:10.3233/SW-2012-0077.

7 Mina Karimi and Krzysztof Janowicz. Exploring challenges of large language models in estimating the distance. *Abstracts of the ICA*, 7:68, 2024.

8 Zilong Liu, Krzysztof Janowicz, Kitty Currier, and Meilin Shi. Measuring geographic diversity of foundation models with a natural language–based geo-guessing experiment on GPT-4. In *AGILE: GIScience Series*, volume 5, pages 1–7, 2024. doi:10.5194/agile-giss-5-38-2024.

9 David M. Mark. Toward a theoretical framework for geographic entity types. In Andrew U. Frank and Irene Campari, editors, *Spatial Information Theory A Theoretical Basis for GIS*, pages 270–283, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. doi:10.1007/3-540-57207-4_18.

10 Daniel R. Montello. The perception and cognition of environmental distance: Direct sources of information. In Stephen C. Hirtle and Andrew U. Frank, editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume 1329, pages 297–311. Springer Berlin Heidelberg, 1997. doi:10.1007/3-540-63623-4_57.

11 E. G. Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235, 1885. URL: http://www.jstor.orglstable/2979181.

12 Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973. doi:10.1016/0010-0285(73)90017-0.

13 Joseph Shingleton and Ana Basiri. How close is "close"? An analysis of the spatial characteristics of perceived proximity using large language models. In *Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025*, volume 6 of *11*, pages 1–14, 2025. doi:10.5194/agile-giss-6-11-2025.

14 Barry Smith and David Mark. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B (Planning and Design)*, 30(3):411–427, 2003.

15 Johannes Treutlein, Dami Choi, Jan Betley, Sam Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 140667–140730. Curran Associates, Inc., 2024. doi:10.52202/079017-4465.

16 Yi-fu Tuan. *Space and place: the perspective of experience*. University of Minnesota Press, 1977.

17 Songlin Wang. Human Migration Distance Conceptualization. Software, swhId: `swh:1:dir:68cffc055f1b2bcd60c8893eb122fb83588c7fcf` (visited on 2026-09-01). URL: `https://github.com/SonglinWang1111/Long-Short-Distance-Conceptualization`, `doi:10.4230/artifacts.27840`.

18 Sam Wilding, David Martin, and Graham Moon. How far is a long distance? An assessment of the issue of scale in the relationship between limiting long-term illness and long-distance migration in england and wales. *Population, Space and Place*, 24(2):e2090, 2018. e2090 PSP-16-0139.R2. `doi:10.1002/psp.2090`.

19 Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA, 1953.

20 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. `doi:10.48550/arXiv.2303.18223`.