DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

Does AI-Encoded Meaning Align with Human Meaning?

Zhenhua Wang [1], **Aixin Yao** [1], and **Ming Ren** [1]

[1] School of Information Resource Management, Renmin University of China

Abstract

AIs are increasingly used to support metadata processing and investigation, which depends on whether AI-encoded meaning aligns with human meaning. However, AI encodes word meaning through distributional and contextual representations, and it remains unclear whether such representations preserve the meaning value of human system. We answer this question through Zipf's meaning law, which links word frequency to number of word meanings. We compare multiple AI-induced meaning estimates with human-measured meaning. To quantify alignment, we propose Meaning-Zipf Deviation (MZD), which covers continuous meaning distributions and measures their divergence with reliability adjustment. Extensive experiments show that human words consistently follow Zipf's meaning law. AI-encoded meanings also exhibit Zipfian regularities, inheriting part of the statistical structure of human language. However, AI meaning distributions remain flatter than human distributions, with lower scaling exponents and non-negligible MZD values. Larger models do not reduce this gap. AI tends to bind words to context-conditioned senses rather than preserve their broader polysemous potential.

1. Introduction

Large language model-based AIs are increasingly used in knowledge organization and metadata applications, including subject metadata creation, keyword recommendation, semantic enrichment, vocabulary mapping, and metadata quality control. These tasks require AI systems to recognize what a document is about, distinguish domain-specific meanings of terms, and connect textual expressions with human-curated semantic structures such as dictionaries, lexical resources, authority files, controlled vocabularies, and knowledge organization systems. A key foundation of these capabilities lies in the computational representation of linguistic meaning. Word meaning is, at least partially, embodied in usage patterns, and thus can be inferred from large-scale textual regularities [1], [2], [3]. However, AI-encoded meaning and human-perceived lexical meaning may follow different organizational principles. Human lexical meaning is grounded in embodied experience, cognitive constraints, social interaction, and curated semantic conventions [4], whereas AI-derived meaning is learned primarily from statistical regularities such as co-occurrence patterns, see *Appendix A.1*. Although the two may yield consistent performance on specific tasks, such behavioral convergence does not entail that AIs and humans organize or quantify word meaning in an identical manner. This leaves a core research question yet to be fully addressed: does AI-encoded meaning align with human lexical meaning, especially when AI is used to support metadata creation and semantic organization? This question is critical to evaluating AI "understanding", ensuring interpretable human-AI collaboration, and defining AI's application boundaries in knowledge production, metadata generation, and semantic interoperability [5], [6].

Full Paper

Available Aug 01, 2026

Correspondence to
Ming Ren
renm@ruc.edu.cn

To address this question, an analytical framework is needed that connects word use with word meaning. Zipf's meaning law provides such a framework, see *Appendix A.2*. This law describes a power-law relationship between word frequency and the number of meanings associated with a word, and has been widely regarded as one of the characteristic regularities of human language [7], [8], [9], [10]. Its significance goes beyond the direct enumeration of senses. Under limited cognitive resources, human communication tends to use economical linguistic forms to carry rich expressive functions, leading high-frequency words to accumulate greater semantic load. Although meaning itself is difficult to observe directly, the long-term organization of language, where fewer forms are used to convey richer content, leaves an identifiable scaling relationship between frequency and number of meanings [11]. Therefore, Zipf's meaning law offers an opportunity to assess whether AI-encoded meaning can support human-centered metadata work.

We examine AI's capacity for meaning encoding by comparing the Zipfian structure of AI-induced word meanings with that of human lexical meanings. Zipf's meaning law is then modelled through the log-linear regression framework. To quantify the degree of alignment between AI-encoded and human-measured meaning, we further propose Meaning-Zipf Deviation. This metric covers continuous meaning distributions, and then measures their discrepancy through an integral form of the Jensen–Shannon distance.

Our experiments across multiple corpora, and with language model families, yield some findings. First, human-measured meanings consistently conform to Zipf's meaning law under literary, encyclopedic, general, and academic lexical systems. Second, AI-encoded meanings display Zipfian regularities, but AI's distributions are systematically flatter than human distributions. AI tends to underestimate the semantic expansion of high-frequency words and produces a more conservative allocation of meaning. Third, increasing model scale does not reduce this deviation, suggesting that larger AIs do not necessarily yield more human-like lexical meaning structures.

The main contributions are as follows. First, it reframes Zipf's meaning law as a new method for comparing human and AI meaning. Second, the Meaning-Zipf Deviation is introduced as a potential diagnostic tool for AI-assisted metadata quality assessment. Finally, it reveals that current language models can reproduce certain Zipfian regularities but still fail to match the human allocation of lexical meaning.

2. Method

2.1. Zipf's law modelling

Let f and m denote the occurrence frequency of a word and the number of its meanings, respectively. Zipf's meaning–frequency law can be formulated as: $m \propto f^{\alpha}$.

Following representative studies [12], this relationship is examined through regression analysis, where f and m serve as the explanatory and dependent variables, respectively. To support robust linear estimation, we have $\log_{10}(m) = \alpha \log_{10}(f) + \beta$, where β is a constant that depends, in part, on the size and composition of the corpus.

Consistent with common practice in previous studies, the regression is conducted using the mean values of m and f within frequency-based bins. Specifically, all word types are sorted in descending order by frequency, and a fixed bin size λ is adopted, such as λ=100. For the j-th bin, word types satisfy: $\lambda(j-1)+1 \le i \le \lambda j$, where $j = 1, 2, 3, \cdots, round(\frac{n}{\lambda})$. The coefficient of determination, R^2, is used to evaluate the extent to which the meaning–frequency law holds. In addition, a Student's t-test is applied to examine whether the exponent α is significantly different from zero.

2.1.1. *AI-encoded meaning*

The quantification of AI-based word meaning is inspired by three streams of research. Şahinuç and Koç [13] suggest that the variance of Gaussian embeddings can be interpreted as a representation of semantic breadth or semantic uncertainty: a larger variance indicates a broader, more generalized, and potentially richer semantic scope. Sun and Platoš [14] construct word sense embeddings from contextual features extracted by bidirectional LSTM and self-attention mechanisms, where clustering centers correspond to induced word senses. Nagata and Tanaka-Ishii [12] argue that each occurrence of the same word in a different context can be regarded as a meaning instance, and that the directional dispersion of these instances in the vector space can be used to measure the contextual diversity of word meaning.

Inspired by these approaches, we propose three refined methods for estimating number of AI-encoded meanings: embedded variance-based uncertainty (EVU), context representation-based clustering (CRC), and contextual direction-based discrepancy (CDD).

EVU. Let word w have n_w contextual instances extracted from the corpus. Their AI-generated contextualized representations, obtained for example from Qwen3-8B, are denoted as: $E(w) = \left\{e_{w,1}, e_{w,2}, ..., e_{w,n_w}\right\} \in \mathbb{R}^d$. For each embedding dimension, we compute the sample variance across all contextual instances and then take the average over all dimensions. The EVU-based number of meanings is obtained as: $m_{EVU} = \frac{1}{d}\sum_{j=1}^{d} Var\left(e_{w,1}^{(j)}, e_{w,2}^{(j)}, ..., e_{w,n_w}^{(j)}\right)$.

CRC. For w, we first collect its set of contextualized instance representations $E(w)$. We then apply DPC-KMeans clustering to partition these representations into: $E(w) = C_1 \bigcup C_1 \bigcup ... \bigcup C_{k_w}$, where each cluster C_j corresponds to a latent word sense. The center of cluster C_j is calculated as: $c_{w,j} = \frac{1}{|C_j|}\sum_{e \in C_j} e$, and is treated as the j-th sense embedding of word w. Accordingly, the CRC-based number of meanings is obtained as: $m_{CRC} = k_w$.

CDD. Let the representations of word w be: $E(w) = \left\{e_{w,1}, e_{w,2}, ..., e_{w,n_w}\right\} \in \mathbb{R}^d$. Each instance vector is first normalized to unit length: $\tilde{e}_{w,i} = \frac{e_{w,i}}{\|e_{w,i}\|_2}$. The mean normalized vector of w is then computed as: $\bar{e}_w = \frac{1}{n_w}\sum_{i=1}^{n_w} \tilde{e}_{w,i}$, $l_w = \|\bar{e}_w\|_2$. The CDD-based number of meanings is obtained as: $m_{CDD} = \frac{1-l_w^2}{l_w(d-l_w^2)}$.

2.1.2. *Human-measured meaning*

To obtain a reliable measure of human word meaning, we draw on manually curated dictionary and lexical knowledge systems. Such resources provide explicitly defined

senses, reflect established linguistic conventions and socially shared semantic distinctions, and are grounded in expert annotation. They therefore offer a suitable gold standard.

We first use WordNet, a widely used large-scale lexical database for English [15]. For word w, the number of synsets associated with it in WordNet can be regarded as the number of senses assigned to that word in a manually curated lexical system. Specifically, for w, we retrieve its sense set: $S(w) = \{s_1, s_2, ..., s_{k_w}\}$.

To obtain a more stable estimate of human-measured number of meanings, we further incorporate BabelNet and Wikidata Lexemes as complementary lexical resources. BabelNet is a large-scale lexical-semantic resource with both lexicographic and encyclopedic coverage, connecting common words, concepts, and named entities [16]. For w, we retrieve its corresponding BabelNet sense set: $B(w) = \{b_1, b_2, ..., b_{r_w}\}$.

Wikidata Lexemes is a large open lexical knowledge base that describes lexical items, word forms, and word senses through the structure of Lexeme, Form, and Sense [17]. For d w, we retrieve its corresponding Wikidata Lexeme sense set: $D(w) = \{d_1, d_2, ..., d_{q_w}\}$.

Finally, to integrate the complementary coverage of the three resources, we merge and deduplicate the sense sets: $U(w) = Dedup(S(w) \cup B(w) \cup D(w))$. The human-measured number of meanings of w is therefore defined as: $m_H = |U(w)|$.

2.2. *Meaning-Zipf Deviation*

For the human-measured meaning distribution H and the AI-encoded meaning distribution AI, we fit Zipf's meaning law: $\hat{m}_S(f) = 10^{\beta_S} f^{\alpha_S}$, $S \in \{H, AI\}$.

To transform the law into a continuous meaning distribution, let $x = \log_{10}(f)$, and the fitted Zipf curve is rewritten as: $\hat{m}_S(x) = 10^{\alpha_S x + \beta_S}$. On the common domain $[x_{\min}, x_{\max}]$, we normalize the fitted curve into a continuous probability density: $p_S(x) = \frac{\hat{m}_S(x)}{\int_{x_{\min}}^{x_{\max}} \hat{m}_S(t)dt}$.

Thus, $\int_{x_{\min}}^{x_{\max}} p_S(x)dx = 1$. Here, $p_S(x)$ represents how the Zipf meaning law allocates meaning mass over the logarithmic frequency space. After normalization, the scale effect of β is removed; the comparison therefore focuses mainly on the distributional shape of the two Zipf laws, especially the allocation tendency determined by the α.

To quantify the discrepancy between the two distributions, we present calculus-based Jensen–Shannon distance. Let $q(x) = \frac{p_H(x)+p_{AI}(x)}{2}$, and we have:

$$JS(p_H, p_{AI}) = \frac{1}{2}\int_{x_{\min}}^{x_{\max}} p_H(x) \log \frac{p_H(x)}{q(x)}dx + \frac{1}{2}\int_{x_{\min}}^{x_{\max}} p_{AI}(x) \log \frac{p_{AI}(x)}{q(x)}dx$$

We normalize it to the [0,1]: $D_{JS}(H, AI) = \sqrt{\frac{JS(p_H, p_{AI})}{log2}}$. $D_{JS} = 0 \rightarrow$ distributions are identical; $D_{JS} = 1 \rightarrow$ maximum possible divergence between them.

If either side has a low R^2, it suggests that the corresponding distribution substantially deviates from Zipf's meaning law. Therefore, R^2 is introduced as a reliability penalty. Let ε be a small positive constant to avoid division by zero, and let $\gamma \in$

Table 1. *Description of corpus*

Corpus	Character count	Word count	Sentence count	Top100 word share	Hapax type share
arxiv	14,169,879,358	2,226,885,777	172,766,686	0.475	0.399
brown	6070766	1005968	58511	0.479	0.425
gutenberg	25,317,305,855	4,421,317,699	305,063,970	0.497	0.463
wiki	558,215,509	88,972,472	4,559,093	0.439	0.391

$[0, 1]$ control the strength of the penalty. The reliability weight is: $W_{R^2}(H, AI) = \left(\max\left(\varepsilon, R_H^2\right)\max\left(\varepsilon, R_{AI}^2\right)\right)^{\gamma/2}$.

The Meaning-Zipf Deviation is then defined as:

$$MZD(H, AI) = 1 - \left(1 - D_{JS}(H, AI)\right)W_{R^2}(H, AI)$$

MZD lies in the interval [0,1]. If the two meaning distributions are identical and both fitted Zipf laws have perfect goodness of fit, then MZD = 0. If the two distributions are highly divergent, or if either fitted law has a low R^2, then MZD $\to$ 1. Therefore, MZD jointly captures two aspects: whether the Zipfian structure encoded by the AI resembles the human-measured distribution, and whether the two fitted laws themselves are statistically reliable.

3. Experiment and Result

3.1. Data and Setting

We use four representative corpora covering distinct language settings: Project Gutenberg for literary texts [18], WikiText for encyclopedic writing [19], Brown Corpus for balanced general English [20], and arXiv for academic papers [21]. These corpora differ in style, topical structure, domain specificity, and lexical distribution, providing a heterogeneous basis for examining meaning–frequency relationships. All texts are lowercased, and English word types are extracted using regular expressions by retaining only alphabetic tokens. Special symbols are removed, and spaCy lemmatization is applied to reduce inflectional variants to lemma forms. Table 1 reports the statistics of the cleaned corpora, where Hapax type share denotes the proportion of word types that occur only once.

To examine the effect of model choice on AI-based meaning estimation, we evaluate 3 model families: BERT (small, base, large), GPT-2 (small, medium, large), and Qwen3 models ranging from 0.6B to 8B parameters. For each model, contextualized word representations are obtained by averaging hidden states from the final several layers. We use the same occurrence-level extraction strategy across models, with batch size and maximum input length adjusted according to model scale. Corpus frequency is computed at the word level rather than the token or subword level, so frequency and meaning count share the same unit. Each experiment is repeated 5 times, and mean values are reported.

3.2. Human-measured meaning and Zipf's meaning law

We first examine whether human-measured meaning conforms to Zipf's meaning law. Figure 1 (*Appendix B*) shows that human lexical meanings consistently exhibit a Zipfian meaning–frequency pattern. The estimated scaling exponent α is positive in all corpora, with all $p < 10^{-4}$. The goodness of fit is also high: $R^2 = 0.870$ for arXiv, 0.916 for Brown, 0.931 for Gutenberg, and 0.926 for Wiki. These results confirm the robustness of Zipf's meaning law in human lexical systems.

The α varies across corpora, suggesting that meaning allocation is shaped by corpus type. To interpret this variation, we divide each corpus into 500 fixed-length chunks and compute lexical indicators including MSTTR, normalized lexical entropy, top-100 word share, low-frequency word share, Gini coefficient, long-word proportion, top-50 bigram share, corpus-specific log-odds, and cross-entropy relative to Brown. Kruskal–Wallis tests show significant cross-corpus differences for all indicators, with all $p < 10^{-4}$ (Figure 2 in *Appendix B*), confirming that the corpora differ substantially in lexical organization.

arXiv has the lowest exponent, $\alpha = 0.173$, together with the highest domain log-odds, cross-entropy to Brown, long-word proportion, top-100 word share, and Gini coefficient. This indicates a highly specialized and concentrated vocabulary structure. Brown has the highest exponent, $\alpha = 0.343$, along with the highest MSTTR and normalized lexical entropy, and the lowest concentration and domain-shift indicators, reflecting balanced and dispersed general-language usage. Gutenberg shows an intermediate exponent, $\alpha = 0.283$, with high lexical diversity but the lowest long-word proportion, suggesting that literary diversity mainly arises from stylistic and narrative variation rather than technical terminology. Wiki has a lower exponent, $\alpha = 0.229$, and high cross-entropy to Brown and top-50 bigram share, consistent with standardized encyclopedic writing and recurrent knowledge expressions.

Overall, human-measured meanings robustly follow Zipf's meaning law, but the strength and interpretation of the scaling relationship vary with corpus type. This establishes a necessary human baseline for comparing AI-encoded meaning distributions.

3.3. AI-encoded meaning and Zipf's meaning law

Figure 3–Figure 5 (*Appendix B*) show that AI-based meaning estimates can also exhibit Zipfian meaning–frequency patterns: more frequent words generally receive higher estimated meaning quantities. However, the strength of this relationship varies substantially across measurement methods, which is expected given that different operationalizations capture different aspects of AI-encoded meaning.

Clear differences emerge among the three AI-based meaning measures. The EVU- and CRC-based measures are relatively stable, often reaching good ($R^2 > 0.8$) or even excellent ($R^2 > 0.9$) goodness of fit, and the CDD-based measure shows weaker fitting performance. This suggests that language models, after being trained on large-scale human-generated corpora, can inherit certain structural regularities of human lexical

meaning. Nevertheless, the degree to which such regularities are expressed depends strongly on how model-based number of meanings is defined.

Although larger AIs are often associated with stronger general intelligence and broader semantic understanding [22], our results indicate that this assumption does not straightforwardly hold for number of word meanings. Larger models do not necessarily conform more closely to Zipf's meaning law. In the BERT series, some corpora and measurement settings show improved fitting performance as model size increases. In the GPT-2 series, GPT-2-large tends to be more stable than the small and medium variants. However, in the Qwen series, larger models do not consistently produce higher slopes or better R^2 values; in some cases, both even declines. These results indicate that greater model scale does not automatically translate into a meaning structure that more closely resembles the human lexical system.

Corpus type also plays an important role. AI's Zipfian behaviour is not constant across corpora. In knowledge-oriented or literary corpora, the relationship between number of AI-estimated meanings and word frequency is more likely to display a stable Zipfian pattern. By contrast, balanced general corpora, such as Brown, can be more difficult for some models to capture consistently.

3.4. Alignment between AI-encoded and human-measured meaning

Figure 6 (*Appendix B*) visualizes the Meaning-Zipf Deviation (MZD) between AI-encoded and human-measured meaning distributions. Across the 120 model–corpus combinations, the mean MZD is 0.331 and the median is 0.246, indicating a substantial deviation between AI and human meaning structures.

At the corpus level, Wiki has the lowest average MZD, 0.284, suggesting that AI-encoded meaning structures are relatively closer to human-measured meaning in encyclopedic texts. Brown has the highest average MZD, 0.365, indicating that balanced general-language corpora place stronger demands on AI–human meaning alignment. The average MZD values for arXiv and Gutenberg are 0.341 and 0.336, respectively, falling between Wiki and Brown. These results suggest that the alignment between AI and human meaning systems is shaped by the way words are used across different textual environments.

4. Conclusion

This study examines whether AI-encoded word meaning aligns with human meaning using Zipf's meaning law. We compare multiple AI-induced meaning measures with human meaning quantities from curated lexical resources, and propose MZD to quantify the distributional gap between the two systems. Across four corpora, human meanings consistently follow Zipf's meaning law, and AI meanings also show Zipfian regularities. However, this does not imply alignment. AI meaning distributions are systematically flatter than human distributions, with lower scaling exponents and non-negligible MZD values. This suggests that AI underestimates the semantic expansion of high-frequency words and allocates meaning more conservatively. Larger models do not necessarily

reduce this gap, indicating that scale alone does not recover the human allocation of lexical meaning.

These findings are relevant to metadata studies. Subject metadata, vocabulary mapping, and semantic enrichment depend on the ability to preserve the semantic breadth of frequent, polysemous, and domain-sensitive terms. If AI encodes high-frequency terms in flatter and more context-bound ways, AI-generated metadata may overlook broader, narrower, or related meanings, leading to semantic compression, incomplete subject description, and weaker interoperability across collections and domains. MZD can therefore serve as a diagnostic signal for metadata quality assessment. Terms, corpora, or domains with high MZD values may require stronger support from controlled vocabularies, lexical resources, provenance documentation, and human review. These results suggest that reliable AI-assisted metadata work should consider meaning alignment in addition to model scale, embedding quality, or surface-level lexical diversity.

Appendix A. Related Work

A.1. AI-encoded meaning

Word meaning is central to AI language understanding, since tasks such as retrieval, question answering, knowledge extraction, and text generation all require models to determine how words refer, associate, and shift across contexts [22], [23], [24]. Early work relied on the distributional hypothesis, representing words as static vectors in models such as Word2Vec and GloVe. This made word meaning computationally measurable, but compressed all senses of a word into a single representation [25].

Pretrained language models changed this view by introducing contextualized representations. Models such as BERT, ELMo, and GPT-2 assign different representations to the same word in different contexts, with upper layers becoming more context-dependent. Subsequent studies show that such representations can partially encode polysemy, sense relatedness, and sense separability [1], [2], [26]. Recent work further derives explicit sense representations through word sense induction and word sense embeddings, or measures semantic breadth through non-point embeddings and contextual vector dispersion [12], [13], [14].

Overall, existing research shows that AI can represent word meaning, differentiate polysemy, and reveal meaning-related statistical patterns. Yet a fundamental question remains unresolved: whether the meanings encoded by AI can reach the human word meanings.

A.2. Zipf's meaning law

Zipf's meaning law links word use with word meaning. It states that more frequent words tend to have more meanings or broader semantic applicability, while this semantic expansion grows sublinearly. The law reflects the principle of linguistic economy: human language balances expressive efficiency, cognitive cost, and communicative demand. High-frequency forms tend to carry richer semantic functions, whereas low-

frequency forms are usually more specific. This pattern has been related to the tension between unification and diversification in communication [3], and provides an empirical entry point for measuring how semantic load is distributed across the lexicon [9].

Empirical studies have provided cross-linguistic evidence. Casas et al. define the meaning–frequency law as a positive association between word frequency and WordNet polysemy, and test it in English, Spanish, and Dutch [8]. Bond et al. further examine the law with multilingual wordnets across languages including English, Spanish, Portuguese, French, Polish, Japanese, Indonesian, and Chinese [7]. These studies show that the law is more stable at the aggregate level, especially when average meaning counts and frequency ranks are used, but weaker for predicting the number of senses of individual words.

Accordingly, Zipf's meaning law should be treated as a macro-level regularity rather than a word-level rule. Individual sense counts may vary with lexicographic conventions, part of speech, register, corpus source, and semantic change [27]. This is crucial for our study: when comparing human and AI meaning distributions, the central question is not whether each individual word has the same sense count, but whether systematic deviations emerge in the overall meaning–frequency structure [28]. Communication-optimization models further support this view by showing that dependencies between form frequency and number of meanings can emerge from constraints on expressive economy, information transmission, and form–meaning mapping [10].

Appendix B. Experimental Result

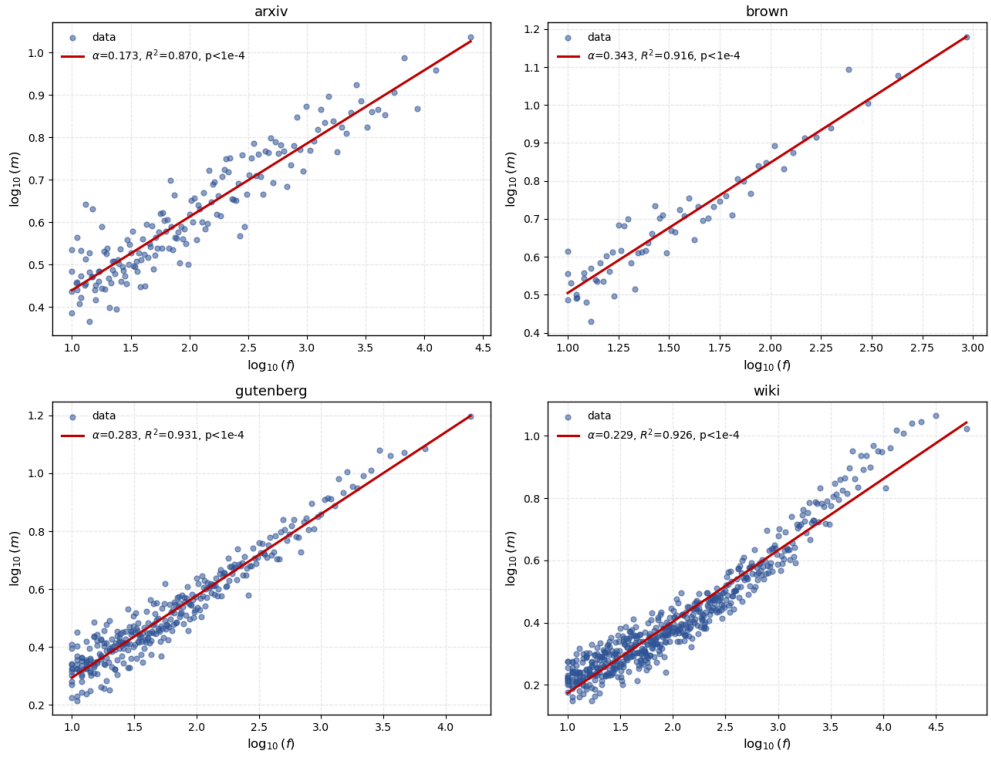

Figure B1. *Zipfian fitting of human-measured meaning quantities.*

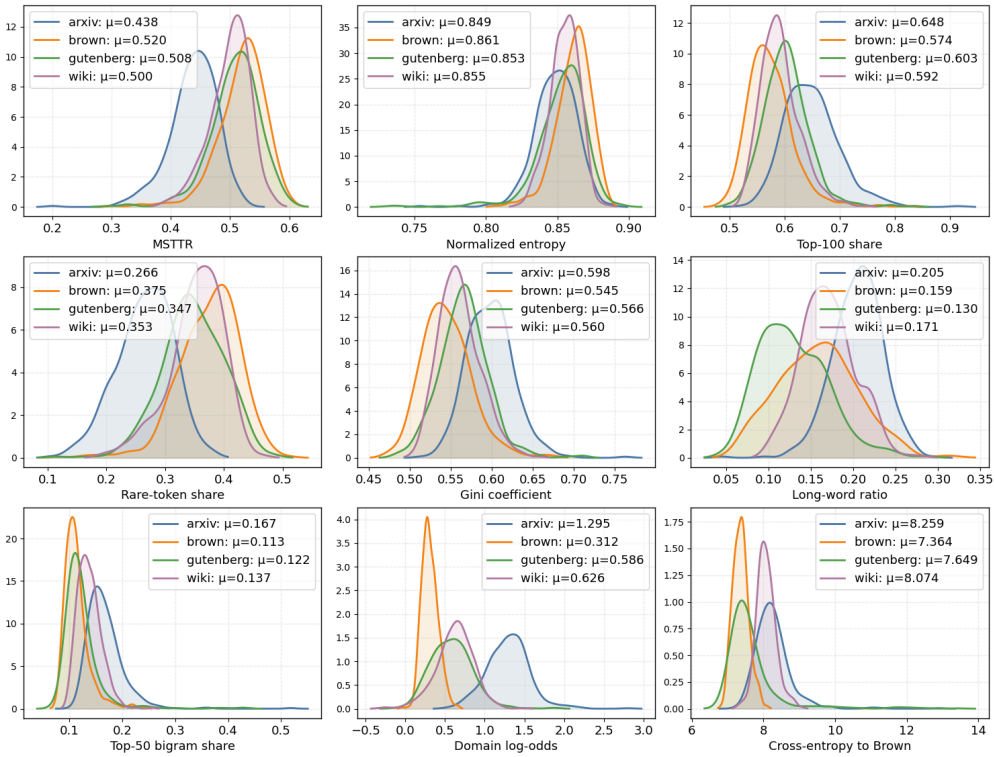

Figure B2. *Corpus effects on the interpretation of Zipf's meaning law. The y-axis represents density, and, μ denotes the mean value. All $p < 10^{-4}$.*

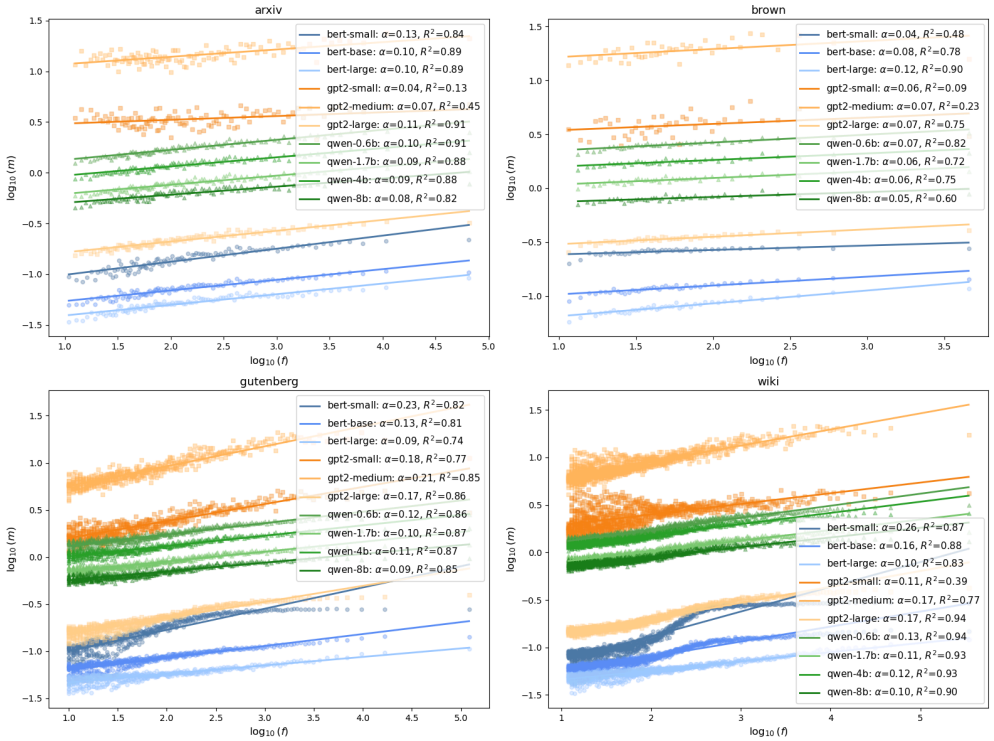

Figure B3. *Zipfian fitting based on m_{EVU}. All estimations are statistically significant, $p < 10^{-4}$.*

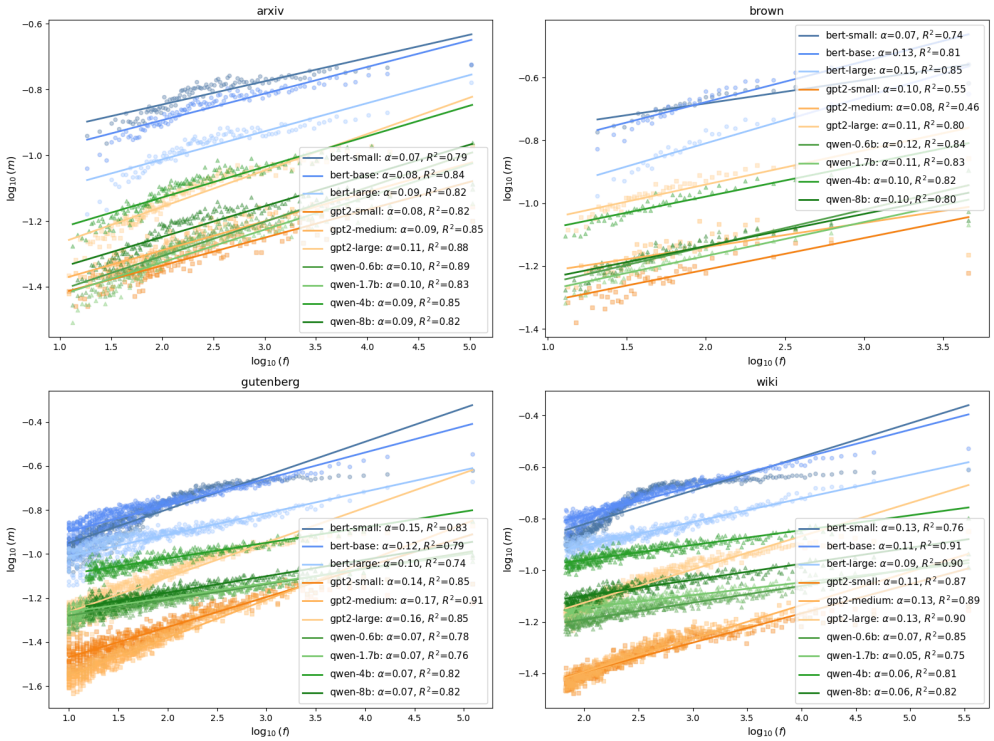

Figure B4. *Zipfian fitting based on m_{CRC}. All estimations are statistically significant, $p < 10^{-4}$.*

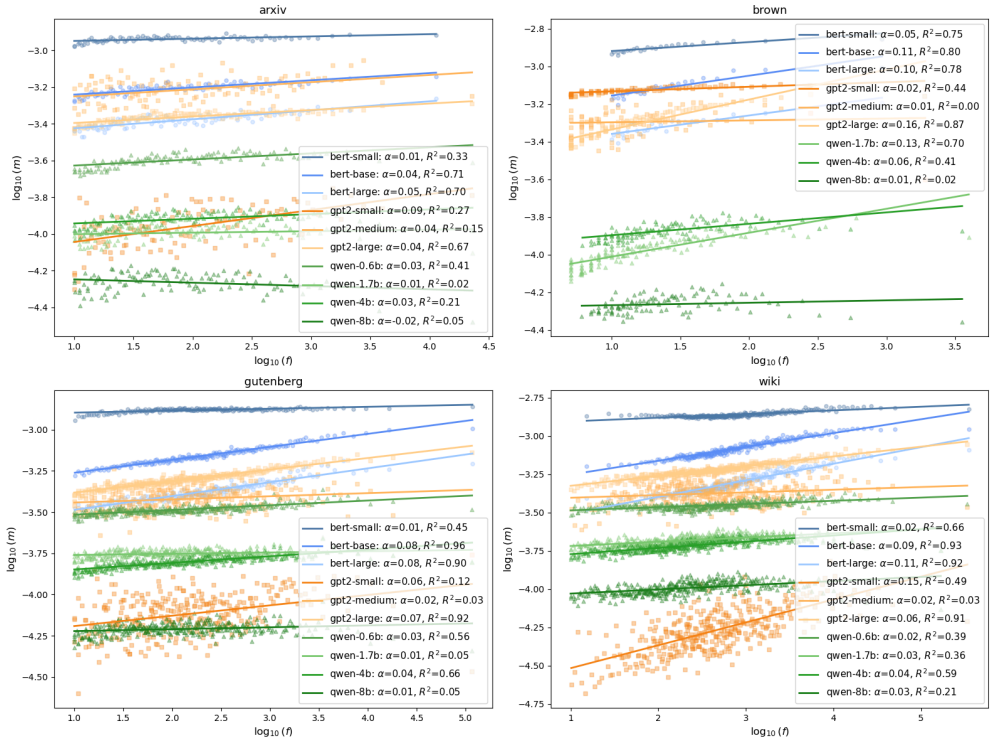

Figure B5. *Zipfian fitting based on m_{CDD}. All estimations are statistically significant, $p < 10^{-4}$.*

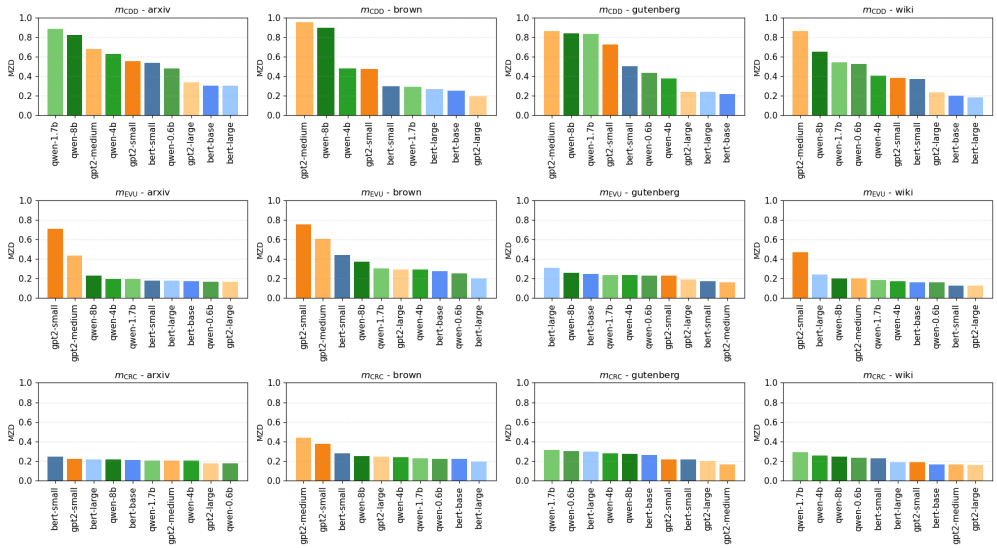

Figure B6. *Distribution of Meaning-Zipf Deviation.*

References

[1] K. Ethayarajh, *How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings*. in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019, pp. 55–65.

[2] A. Garí Soler and M. Apidianaki, "Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 825–844, 2021.

[3] S. Trott and B. Bergen, "Languages are efficient, but for whom?," *Cognition*, vol. 225, Art. no. 105094, 2022.

[4] B. M. Lake and G. L. Murphy, "Word meaning in minds and machines," *Psychological Review*, vol. 130, Art. no. 401, 2023.

[5] M. R. Qorib, G. Moon, and H. T. Ng, *Are decoder-only language models better than encoder-only language models in understanding word meaning?*. in Findings of the Association for Computational Linguistics: ACL 2024. 2024, pp. 16339–16347.

[6] D. Loureiro, K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados, "Analysis and evaluation of language models for word sense disambiguation," *Computational Linguistics*, vol. 47, pp. 387–443, 2021.

[7] F. Bond, A. Janz, M. Maziarz, and E. Rudnicka, *Testing Zipf's meaning-frequency law with wordnets as sense inventories*. in Proceedings of the 10th Global WordNet Conference. 2019, pp. 342–352.

[8] B. Casas, A. Hernández-Fernández, N. Català, R. Ferrer-i-Cancho, and J. Baixeries, "Polysemy and brevity versus frequency in language," *Computer Speech & Language*, vol. 58, pp. 19–50, 2019.

[9] A. Koshevoy, I. Dautriche, and O. Morin, *Why do some words have more meanings than others? A true neutral model for the meaning-frequency correlation*. in Proceedings of the Annual Meeting of the Cognitive Science Society, no. 45. 2023.

[10] R. Ferrer-i-Cancho and M. S. Vitevitch, "The origins of Zipf's meaning-frequency law," *Journal of the Association for Information Science and Technology*, vol. 69, pp. 1369–1379, 2018.

[11] S. Thurner, R. Hanel, B. Liu, and B. Corominas-Murtra, "Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation," *Journal of the Royal Society Interface*, vol. 12, 2015.

[12] R. Nagata and K. Tanaka-Ishii, *A new formulation of Zipf's meaning-frequency law through contextual diversity*. in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, pp. 15323–15335.

[13] F. Şahinuç and A. Koç, "Zipfian regularities in "non-point" word representations," *Information Processing & Management*, vol. 58, Art. no. 102493, 2021.

[14] Y. Sun and J. Platoš, "A method for constructing word sense embeddings based on word sense induction," *Scientific Reports*, vol. 13, Art. no. 12945, 2023.

[15] R. H. Maudslay, S. Teufel, F. Bond, and J. Pustejovsky, *ChainNet: Structured metaphor and metonymy in WordNet*. in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024, pp. 2984–2996.

[16] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, and F. Cecconi, *Ten years of BabelNet: A survey*. in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. 2021, pp. 4559–4567.

[17] C. Möller, J. Lehmann, and R. Usbeck, "Survey on English entity linking on Wikidata: Datasets and approaches," *Semantic Web*, vol. 13, pp. 925–966, 2022.

[18] J. Brooke, A. Hammond, and G. Hirst, *GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus*. in Proceedings of the Fourth Workshop on Computational Linguistics for Literature. 2015, pp. 42–47.

[19] A. Baevski and M. Auli, *Adaptive input representations for neural language modeling*. in International Conference on Learning Representations. 2019.

[20] N. Ide and C. Macleod, *The American National Corpus: A standardized resource of American English*. in Proceedings of Corpus Linguistics, no. 3. 2001, pp. 1–7.

[21] A. Cohan *et al.*, *A discourse-aware attention model for abstractive summarization of long documents*. in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Volume 2 (Short Papers). 2018, pp. 615–621.

[22] A. Singh *et al.*, "OpenAI GPT-5 system card," 2025.

[23] Z. Wang, H. Chen, G. Xu, and M. Ren, "A novel large-language-model-driven framework for named entity recognition," *Information Processing & Management*, vol. 62, Art. no. 104054, 2025.

[24] V. C. Storey *et al.*, "Large language models for conceptual modeling: Assessment and application potential," *Data & Knowledge Engineering*, vol. 160, Art. no. 102480, 2025.

[25] S. Teglia, S. Tedeschi, and R. Navigli, *How much do encoder models know about word senses?*. in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025, pp. 2266–2277.

[26] S. Nair, M. Srinivasan, and S. Meylan, *Contextualized word embeddings encode aspects of human-like word sense knowledge*. in Proceedings of the Workshop on the Cognitive Aspects of the Lexicon. 2020, pp. 129–141.

[27] Z. Wang, G. Xu, and M. Ren, "Data augmentation with large language models: A scaling law-guided approach," *ACM Transactions on Knowledge Discovery from Data*, vol. 36, pp. 1–37, 2026.

[28] Z. Wang, A. Yao, and M. Ren, "AI-generated text detection via multilevel Zipf's law," *Journal of Informetrics*, vol. 20, Art. no. 101799, 2026.

[29] Z. Wang, M. Ren, D. Gao, and Z. Li, "A Zipf's law-based text generation approach for addressing imbalance in entity extraction," *Journal of Informetrics*, vol. 17, Art. no. 101453, 2023.

[30] Z. Wang, S. He, G. Xu, and M. Ren, "Will sentiment analysis need subculture? A new data augmentation approach," *Journal of the Association for Information Science and Technology*, vol. 75, pp. 655–670, 2024.