# Surprisal reveals diversity gaps in image captioning and different scorers change the story

**Nikolai Ilinykh and Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
{nikolai.ilinykh,simon.dobnik}@gu.se

## Abstract

We quantify linguistic diversity in image captioning with **surprisal variance** – the spread of token-level negative log-probabilities within a caption set. On the MSCOCO test set, we compare five state-of-the-art vision-and-language LLMs, decoded with greedy and nucleus sampling, to human captions. Measured with a caption-trained n-gram LM, humans display roughly twice the surprisal variance of models, but rescoring the same captions with a general-language model reverses the pattern. Our analysis introduces the surprisal-based diversity metric for image captioning. We show that relying on a single scorer can completely invert conclusions, thus, robust diversity evaluation must report surprisal under several scorers.

## 1 Introduction

Every person is unique in their linguistic behaviour: the same image may evoke many different yet valid descriptions depending on their different personal contexts, beliefs, attention, and background knowledge. This "natural variability" has been documented in linguistic annotation (Plank et al., 2014; Plank, 2022), inference (Pavlick and Kwiatkowski, 2019), object naming (Silberer et al., 2020), and image captioning (Bernardi et al., 2016). On the other hand, although vision-and-language models may capture fragmets of individual styles, they are normally not trained with capturing diversity in mind. They are optimised for maximum-likelihood estimation (MLE), a function that rewards word combinations already dominant in the training corpus and penalises rarer phrasing (Devlin et al., 2015; Dai et al., 2017). Decoding tweaks (e.g. nucleus sampling (Holtzman et al., 2020)) or alternative objectives (Welleck et al., 2020; Li et al., 2016a) may mitigate this effect. Still, transformer-based models trained with MLE now achieve impressive accuracy on caption benchmarks (Alayrac et al., 2022; Li et al., 2023; Bai et al., 2025; Zhu et al., 2025). However, such benchmarks assume a single ground truth and often focus on targeting this very benchmark, meeting a certain metric score (Schlangen, 2021). As a result, the diversity of model language is often sidelined – even though prior studies argue that missing variability is critical for evaluating multi-modal models (Castro Ferreira et al., 2016; Li et al., 2016b; van Miltenburg et al., 2018). Existing diversity metrics for tasks such as image captioning (Shetty et al., 2017; van Miltenburg et al., 2018), rely on surface counts (length, type–token ratio, distinct-n) and lack a principled link to linguistic processing.

In this study we examine diversity of image captions through the notion of surprisal, a probability-weighted and context-sensitive probe. The surprisal, a negative log-probability of a word in context, has been widely used in linguistic analysis to quantify the processing cost of a message (Hale, 2001; Levy, 2008) and has proved to be a robust metric across languages (Pimentel et al., 2021; Wilcox et al., 2023). Generic, more commonly used words and phrases across speakers or models are easier for either to produce, hence any potential differences in the measured surprisal between these two groups naturally reveal any differences in diversity of their linguistic behaviour (Gehrmann et al., 2019; Venkatraman et al., 2024; Xu et al., 2024). We extend this line of work from text-only settings to image captions.

We estimate the probabilities used to measure surprisal using two probabilistic scorers (i) an n-gram model trained on a large corpus consisting of a balanced collection of human and model generated image captions (in-domain expert), and (ii) GPT-2 (Radford et al., 2019), a general-purpose transformer trained on broad English. As we do not have access to the true population of all possible captions for images, we follow earlier work (Smith and Levy, 2013; Wilcox et al., 2023; Giulianelli et al., 2023) and approximate the underly-

*Proceedings of the 18th International Natural Language Generation Conference*, pages 366–375
October 29–November 2, 2025. ©2025 Association for Computational Linguistics

ing distribution with two different language models. Inspired by the scorer-sensitivity findings of Arora et al. (2022), our dual-scorer design tests whether conclusions about diversity change with the change of the evaluator of surprisal.

Our research questions are:

1. Are human descriptions of individual images more diverse than those generated by language models in terms of the estimated surprisal of the in-domain unbiased scorer?
2. Do predictions of this scorer align to those of a general language model scorer which has been trained on open text?

As artificial describers (description generators) we analyse five state-of-the-art vision-and-language models and compare their descriptions with human captions. Like Zamaraeva et al. (2025), who used linguistic theory to identify syntactic gaps in LLM-generated news compared to human-authored news, we resort to the linguistic theory of surprisal to evaluate multi-modal language, showing how scorer choice interacts with caption distributions. Our study reveals a clear difference in surprisal and therefore diversity between human and model captions, and, importantly, that the direction of this diversity flips with the chosen scorer. These differences between evaluation scorers underscore the need for future benchmarks to include surprisal-based diversity scores computed under several, well-motivated scorers.

## 2 Materials and methods

### 2.1 Data

We use the Karpathy test-split of MSCOCO (Lin et al., 2014), which consists of 5000 images each paired with five independent human captions (25000 descriptions total). MSCOCO is one of the popular and commonly used image captioning benchmarks. It provides multiple human references per image – exactly what we need to quantify how diversely an image can be described by humans and models[1].

### 2.2 Models

We use five vision–language models to generate one description per image: Qwen2.5-VL-72B-Instruct (Bai et al., 2025), InternVL3-78B-78B-Instruct (Zhu et al., 2025), Llama-4-Scout-17B-16E-Instruct[2], Claude Sonnet 4[3], and GPT-4o (OpenAI et al., 2024). These include both closed-source systems and large open-source models, all ranking among the top performers on the MMMU benchmark (Yue et al., 2024). Our choice of five models is intended to correspond to five human describers per image in the MS COCO dataset (Lin et al., 2014). This way we attempt to mirror a scenario of difference in experience between between different human describers, as the models are based on different architectures and training regimes. More on technical details can be found in Appendix A.We produce two different sets of texts by running models with two decoding algorithms: greedy search and nucleus sampling (Holtzman et al., 2020). Different decoding strategies let us assess model captions in a more nuanced way since decoding has a strong effect on the output in NLG (Zarrieß et al., 2021).

### 2.3 Scorers of caption probability distributions

**Overview** For each image in our dataset, we train a single Kneser–Ney-smoothed n-gram model (bi- and tri-) on the *union* of all human and model captions *except* those associated with that target image which will be our test set. As the probabilities estimated by this model are based on both human and artificially generated captions we use this "balanced" language model to score each of the the held-out captions (five human, five machine), giving us one surprisal value per caption. Our zero hypothesis is that this model by virtue of being balanced) will predict no differences in surprisal value between human and model generated captions of individual images.

**Training details** We train both bigram and trigram models to check that our findings are robust to context size. All models are implemented with NLTK's `KneserNeyInterpolated` language model API (Bird et al., 2009; Kneser and Ney, 1995). With 5000 images × 10 captions each, each leave-one-image-out model is trained on $(5000 - 1) \times 10 = 49\,990$ captions. The vocabulary of the model is built from the training pool (excluding the target image). Pooling human and machine

[1] We note that MSCOCO is widely incorporated in training of multi-modal LLMs.

[2] https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[3] https://www.anthropic.com/news/claude-4

captions ensures both groups contribute the same n-gram types and counts, so rare-word penalties are applied equally. Also, since all captions are scored by the same language model, per-token surprisal is directly comparable between human and model outputs. Even if models saw MSCOCO during pre-training, our leave-one-image-out n-gram setup excludes each target image's captions from training, so repeated references are still treated as new, ensuring surprisal reflects style rather than memorisation.

### 2.4 Surprisal as evaluation metric

In psycholinguistics and computational linguistics, *surprisal* (Shannon, 1948) is a well-established measure of the information conveyed by a word in context. Formally, the surprisal of a linguistic unit $w_t$ given preceding units $w_{<t}$ is defined as the negative log-likelihood of the word conditioned on previous context (Hale, 2001; Levy, 2008):

$$I(w_t) = -\log P_\theta(w_t \mid w_{<t}), \tag{1}$$

where $P$ is the underlying probability distribution. Intuitively, treating units as words, predictable words carry less information (lower surprisal), while unexpected words convey more (higher surprisal). We use surprisal to quantify linguistic unpredictability in image captions under a model $\theta$ trained to approximate $P$. We compute word-level surprisals and average them across each caption to estimate how predictable the caption is under a given evaluator. Captions with higher average surprisal are considered more lexically or structurally novel, while captions with lower surprisal follow patterns captured in $P_\theta$.

## 3 Experiments and results

Before moving to our surprisal-based experiments, we first ask if model captions even look like human captions on lexical level. We compute several metrics described in van Miltenburg et al. (2018); the results are shown in Table 1. A table with per model lexical diversity metrics is available in Appendix C. The results show that human and model captions look different. Humans produce concise, about 10 words long captions with little variation, whereas models spin out strings three to four times longer and far less consistent in length. Despite this gap, models list far more unique words than humans, possibly because their captions are several times longer. However, once normalised by output length (TTR metrics), the groups converge. This interesting result suggests that the apparent richness of model vocabulary is largely an artifact of their verbosity. When captions are several times longer, the chance of introducing new word types naturally increases, even if much of the added material is repetitive. Once we control for output length, however, the underlying lexical behaviour looks remarkably similar. This convergence implies that models, like humans, operate within a comparable core vocabulary for the task, but their training encourages them to elaborate rather than compress. In other words, verbosity does not necessarily translate to greater genuine lexical creativity – instead, it may reflect a tendency to "pad" responses with familiar constructions, a behaviour aligned with next-word prediction training objectives rather than pragmatic efficiency in describing images.

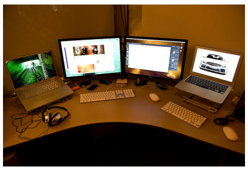
| Source | ASL ± SDSL | #Types | TTR1 | TTR2 |
|---|---|---|---|---|
| Human | $10.44 \pm 2.36$ | 7,252 | 0.28 | 0.66 |
| `greedy` | $39.19 \pm 53.35$ | 10,783 | 0.29 | 0.66 |
| `nucleus` | $40.00 \pm 27.94$ | 13,082 | 0.33 | 0.73 |

Table 1: Overall lexical statistics for human captions and all model captions combined under `greedy` and `nucleus` decoding. ASL = Average Sentence Length (tokens per caption), SDSL = Standard Deviation of Sentence Length, #Types = number of unique word types, TTR1 = type-token ratio (per 1000 tokens), TTR2 = bigram type–token ratio (per 1000 bigrams).

Overall, the results in Table 2 demonstrate that models lengthen their captions mostly by repeating words known to them, so the proportion of genuinely new vocabulary remains virtually the same as in the brief human captions. Such extensive text generation is not pragmatically required as demonstrated by the much shorter human-generated texts for the captioning task. These findings can also be linked to differences in captioning guidelines: those presented to humans for MSCOCO data collection (i.e., "focus on visual content") versus those followed by instruction-tuned models, which were trained to pursue a different goal from that of humans. The reason for the models' verbosity could thus be their tendency to generate texts features different from those found in human-generated texts, including syntactic structures and vocabulary diversity (Muñoz-Ortiz et al., 2024) as well as grammatical differences (Zamaraeva et al., 2025). The effect of the task provided to the image describers

| $P_\theta$ | Data | Human | Model | $t$-value | $d_z$ |
|---|---|---|---|---|---|
| | | Mean ± SD | Mean ± SD | | |
| Bi-gram LM | $\mathcal{H} \cup \mathcal{M}_{\text{greedy}}$ | 5.20 ± 5.04 | 2.17 ± 2.00 | $40.88^{***}$ | 0.58 |
| Tri-gram LM | $\mathcal{H} \cup \mathcal{M}_{\text{greedy}}$ | 7.39 ± 6.04 | 3.71 ± 3.04 | $39.45^{***}$ | 0.56 |
| Bi-gram LM | $\mathcal{H} \cup \mathcal{M}_{\text{nucleus}}$ | 5.03 ± 4.95 | 4.08 ± 3.30 | $11.50^{***}$ | 0.16 |
| Tri-gram LM | $\mathcal{H} \cup \mathcal{M}_{\text{nucleus}}$ | 7.34 ± 6.09 | 6.74 ± 4.49 | $5.70^{***}$ | 0.08 |
| GPT-2 (small) | $\mathcal{H} \cup \mathcal{M}_{\text{greedy}}$ | 1.565 ± 1.521 | 2.015 ± 1.323 | $-16.37^{***}$ | 0.23 |
| GPT-2 (medium) | $\mathcal{H} \cup \mathcal{M}_{\text{greedy}}$ | 1.485 ± 1.420 | 2.016 ± 1.316 | $-20.31^{***}$ | 0.29 |
| GPT-2 (XL) | $\mathcal{H} \cup \mathcal{M}_{\text{greedy}}$ | 1.438 ± 1.344 | 1.954 ± 1.243 | $-20.85^{***}$ | 0.29 |
| GPT-2 (small) | $\mathcal{H} \cup \mathcal{M}_{\text{nucleus}}$ | 1.565 ± 1.521 | 2.906 ± 24.788 | $-3.82^{***}$ | 0.05 |
| GPT-2 (medium) | $\mathcal{H} \cup \mathcal{M}_{\text{nucleus}}$ | 1.485 ± 1.420 | 2.903 ± 24.824 | $-4.03^{***}$ | 0.06 |
| GPT-2 (XL) | $\mathcal{H} \cup \mathcal{M}_{\text{nucleus}}$ | 1.438 ± 1.344 | 2.911 ± 23.710 | $-4.39^{***}$ | 0.06 |

Table 2: Variance in surprisal for human and model image descriptions ($n = 5{,}000$ images; paired $t$-test, $df = 4\,999$). $^{***}$ indicates $p < .001$. $d_z \geq 0.50$ indicates moderate effect size, $d_z < 0.20$ indicates small effect size. For n-gram models, training data for each language-model order $n$ (under "Data" column) is the union of the human corpus $\mathcal{H}$ and machine-generated texts $\mathcal{M}$ produced with the indicated decoding strategy (greedy or nucleus sampling). The same data was used to test pre-trained GPT-2.



**Human caption:**
"The two monitors are sitting among the laptops."

**GPT4o:**
"The workspace setup includes two laptops, two external monitors, two full-size keyboards, and various accessories. One laptop on the left is open and displaying a green-themed image, while another on the right sits elevated on a stand showing a white car. The two central monitors display websites and applications, likely for multitasking. The desk also holds a pair of headphones, a computer mouse for each setup, a smartphone, and other small electronics, indicating a high-performance workstation for media, development, or design tasks."

Figure 1: An example image from MSCOCO test set with one human reference and one model caption.

on the properties and features of the produced captions is also important (Ilinykh et al., 2018). Two captions can be similar or different in terms of length and number of types, yet be radically different in how predictable or information-rich they are. Look at the example in Figure 1. Both captions mention the same objects (monitors, laptops) and share word types, but the model's version explores far less frequent phrases ("workspace setup". "green-themed image") and long chains of modifiers that could carry a different amount of information content. Surprisal under scorers can capture these probability-weighted differences that surface counts alone cannot. In other words, surprisal lets us measure not just how many different words appear, but how unexpectedly they are combined.

One striking difference is that models tend to *overdescribe* visual information exhaustively, while humans focus on particular elements. This selectivity may also allow for greater diversity since different humans can choose to highlight different aspects of an image. If humans are behaving pragmatically, they generate optimal (concise) captions, and anything longer would be pragmatically inappropriate. Multi-modal LLMs, by contrast, struggle to capture captioning intent because they produce much longer texts. This is an interesting finding as it raises the question of what constitutes a good model-generated description. We leave this discussion for future work. Next, we move to our main experiments.

### 3.1 Surprisal within groups: in-domain

We compute surprisal variance across five captions per group for each image and use a paired t-test (Gosset, 1908) to check if human and model variances differ. As shown in Table 2, for both n-gram orders under both decoding methods, human descriptions have shown roughly two times higher variance in surprisal than model descriptions. Switching to `nucleus` narrows the gap, but only slightly as humans still remain the more variable group. This result suggests that different models (although the differ in size, training data and performance on benchmarks) tend to generate image descriptions with similar information content, whereas humans offer more variance in the information they express. Per model descriptive analysis is available in Appendix B.

### 3.2 Surprisal within groups: general-domain

The analysis in Section 3.1 characterises surprisal variance under the probability distribution $P_\theta$ that is estimated by a simple n-gram language model trained in-domain. While these models provide interpretable baselines, their probability estimates are limited by local context and the sparsity of the caption corpus. To test whether our findings generalise under a richer representation of linguistic structure, we replace the caption-trained n-gram models with a large pre-trained language model (GPT-2 (Radford et al., 2019)). This substitution changes the underlying probability distribution to the one trained on a much broader and more expressive corpus, that is more of a general English corpus rather than caption-related one. By doing so, we can examine whether the observed human–model differences in surprisal variance persist when surprisal is computed under a generally more powerful, but different model of language. To compute surprisal with GPT-2, we use codebase provided by Oh et al. (2024)[4]. Recently, previous research has used LLMs as surprisal scorers in the context of second-language writing development (Hu and Cong, 2025). According to Table 2, when surprisal is estimated with the general English language scorer, the trends reverse. Across all GPT-2 configurations, human surprisal variance was lower than model variance, with significant paired differences. One possible explanation for this reversal lies in the training objectives and data distributions of the respective models. The in-domain n-gram model trained only on captions is highly sensitive to local lexical patterns and reflects the narrow regularities of the captioning domain. In contrast, GPT-2 is trained on broad and heterogeneous corpora with the objective of next-word prediction. Under this objective, model-generated captions may appear more variable because they deviate from the stylistic and structural norms GPT-2 has internalised from general English texts, whereas human captions – short, formulaic, and pragmatically efficient – are closer to those norms and thus exhibit lower surprisal variance. In this sense, the two scorers are complementary rather than directly comparable: the n-gram model emphasises within-domain variation, while GPT-2 highlights divergence from general-purpose English usage. The methodological implication is that conclusions about human–model variability depend strongly on the choice of reference probability distribution.

## 4 Conclusion

Our study provides three main messages. First, using surprisal as a measure of diversity gap in image captioning requires reporting under multiple scorers. Doing so will (i) prevent over-interpretation of scores tied to one distribution, and (ii) encourage future models to match human-level variability inside the caption genre rather than describing images in general language. The apparent reversal in variance of surprisal within describer groups is a diagnostic of whose expectations one chooses as their scorer. The two different scorers we employ in this short study should be seen as two complementary metrics. Our work is thus helpful when deciding which scorer to choose given a particular task.

## Limitations

Our analysis centers around two probability models to compute surprisal: the caption-domain n-gram language model and GPT-2. Using additional scorers, including larger or instruction-tuned LLMs will test whether the observed patterns generalise under different scorers. We use MSCOCO Karpathy test split because it offers five independent human references per image. Running our experiments on other datasets with different stylistic conventions and multiple references such as Flickr 30k (Plummer et al., 2016) would test robustness of our method. Linking variance in surprisal to accuracy metrics such as BERTScore (Zhang et al., 2020) or MoverScore (Zhao et al., 2019) will show whether diversity gains align with or trade off against semantic quality. Finally, both captions and scorers are English-based; the interaction between genre-specific and general-language scorers may differ in morphologically richer or typologically distant languages.

## Acknowledgments

The work reported in this paper has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. The

[4] The codebase is available here: `https://github.com/byungdoh/llm_surprisal`

computations and data storage were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.*, 55:409–442.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *Preprint*, arXiv:1504.00325.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2989–2998. IEEE Computer Society.

Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *ArXiv*, abs/1505.04467.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

William Sealy Gosset. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25. Originally published under the pseudonym "Student".

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jingying Hu and Yan Cong. 2025. Modeling Chinese L2 writing development: The LLM-surprisal perspective. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 172–183, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 397–402, Tilburg University, The Netherlands. Association for Computational Linguistics.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10).

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Preprint*, arXiv:1505.04870.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. *Preprint*, arXiv:1703.10476.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. Detecting subtle differences between human and model languages using spectrum of relative likelihood. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing llm-generated and human-authored news text using formal syntactic theory. *Preprint*, arXiv:2506.01407.

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. *Information*, 12(9).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

## A Technical details

Each model was prompted with the following instructions:

**Prompt Setup**

**System prompt:**
You are a helpful annotator tasked with describing images. You will see one image and you will be asked to describe it.

**User prompt:**
Describe all the important parts of the scene. Do not start the sentence with "There is". Do not describe unimportant details. Do not describe things that might have happened in the future or past. Do not describe what a person might say. Do not give people proper names. The sentence should contain at least 8 words.

These instructions replicate those given to human annotators in the MS COCO captioning dataset (Chen et al., 2015). Each model received both the prompt and the image as input.

The open-source models were loaded from HuggingFace repositories. We used `OpenGVLab/InternVL3-78B-78B-Instruct` and `meta-llama/Llama-4-Scout-17B-16E-Instruct` and served them locally using `vLLM` (Kwon et al., 2023) for efficient inference. We used the official repository of Qwen2.5-VL[5] to run `Qwen/Qwen2.5-VL-72B-Instruct`. We used `claude-sonnet-4-20250514` accessed through the Anthropic API[6]. GPT-4o was accessed through the OpenAI API[7], using version `gpt-4o-2024-08-06`.

[5] `https://github.com/QwenLM/Qwen2.5-VL.git`

[6] `https://docs.anthropic.com/en/api/messages`

[7] `https://platform.openai.com/docs/api-reference/`

All open-source models were run offline on four NVIDIA A100 GPUs (80 GB each, except Qwen2.5-VL which used $4 \times 40$ GB). InternVL3-78B and Llama-4-Scout were run with a maximum context length of 8192 tokens. Qwen2.5-VL was run with a maximum output length of 300 tokens and a minimum of 50 new tokens. A full pass over 5000 images took approximately 12 – 14 hours per model. To ensure comparability across models (including those without beam search support), we used two decoding modes for all models:

- `greedy search`:
  temperature 0.0, top_p 0.0
- `nucleus sampling`:
  temperature 1.0, top_p 0.92

Each generation was capped at 300 tokens. All runs used a fixed random seed (42) to ensure reproducibility.

**Pre-processing** We observed that Qwen2.5-VL occasionally hallucinates, producing emoji unicodes, Chinese characters, extra spaces, and line breaks. To clean these artifacts, we process all models' outputs so that they contain only ASCII letters, digits, standard punctuation, and spaces. We tokenise all model-generated captions with PTBTokenizer (using the original MS COCO evaluation code[8]) so that they share exactly the same tokenisation as the human references.

## B Variance in surprisal per model

Descriptively, the models show clear differences in how predictable or variable their outputs are. Across all decoding configurations, Llama 4 Scout-17B-16E shows the lowest variance in surprisal suggesting more uniform descriptions across images. In contrast, Qwen2.5-VL-72B shows the highest mean surprisal and higher variance values. Claude Sonnet 4 and GPT4o fall in between but show an increase in variance from bigram to trigram contexts.

## C Lexical-based diversity metrics per model

[8] [https://github.com/tylin/coco-caption

| Configuration | Model | Mean surprisal | Variance ± SD |
|---|---|---|---|
| Bi-gram LM, greedy | Claude Sonnet 4 | 8.864 | 3.575 ± 1.891 |
| Bi-gram LM, greedy | GPT4o | 8.732 | 2.551 ± 1.597 |
| Bi-gram LM, greedy | InternVL3-78B | 7.995 | 3.130 ± 1.769 |
| Bi-gram LM, greedy | Llama 4 Scout-17B-16E | 8.175 | 1.442 ± 1.201 |
| Bi-gram LM, greedy | Qwen2.5-VL-72B | 10.159 | 3.124 ± 1.767 |
| Tri-gram LM, greedy | Claude Sonnet 4 | 8.385 | 6.106 ± 2.471 |
| Tri-gram LM, greedy | GPT4o | 8.179 | 4.460 ± 2.112 |
| Tri-gram LM, greedy | InternVL3-78B | 7.042 | 5.444 ± 2.333 |
| Tri-gram LM, greedy | Llama 4 Scout-17B-16E | 7.316 | 2.876 ± 1.696 |
| Tri-gram LM, greedy | Qwen2.5-VL-72B | 9.964 | 5.047 ± 2.246 |
| Bi-gram LM, nucleus | Claude Sonnet 4 | 8.742 | 3.311 ± 1.820 |
| Bi-gram LM, nucleus | GPT4o | 9.146 | 2.742 ± 1.656 |
| Bi-gram LM, nucleus | InternVL3-78B | 8.837 | 3.866 ± 1.966 |
| Bi-gram LM, nucleus | Llama 4 Scout-17B-16E | 8.458 | 1.310 ± 1.145 |
| Bi-gram LM, nucleus | Qwen2.5-VL-72B | 12.086 | 4.219 ± 2.054 |
| Tri-gram LM, nucleus | Claude Sonnet 4 | 8.269 | 5.763 ± 2.401 |
| Tri-gram LM, nucleus | GPT4o | 8.817 | 4.695 ± 2.167 |
| Tri-gram LM, nucleus | InternVL3-78B | 8.329 | 6.253 ± 2.501 |
| Tri-gram LM, nucleus | Llama 4 Scout-17B-16E | 7.840 | 2.609 ± 1.615 |
| Tri-gram LM, nucleus | Qwen2.5-VL-72B | 12.667 | 5.197 ± 2.280 |

Table 3: Per-model surprisal statistics across images. For each configuration and model, the table reports the mean surprisal and the variance in surprisal across images with its standard deviation (shown as variance ± SD), computed over 5,000 images.

| Model | ASL | SDSL | #Types | TTR1 | TTR2 |
|---|---|---|---|---|---|
| | | | greedy | | |
| Claude Sonnet 4 | 24.41 | 7.09 | 5,431 | 0.40 | 0.79 |
| GPT4o | 35.59 | 15.63 | 6,507 | 0.38 | 0.79 |
| InternVL3-78B | 15.64 | 7.48 | 3,960 | 0.38 | 0.74 |
| Llama 4 Scout-17B-16E | 77.08 | 107.82 | 6,702 | 0.27 | 0.64 |
| Qwen2.5-VL-72B | 43.23 | 3.99 | 7,021 | 0.46 | 0.86 |
| | | | nucleus | | |
| Claude Sonnet 4 | 24.64 | 7.30 | 5,460 | 0.40 | 0.78 |
| GPT4o | 35.39 | 15.70 | 7,031 | 0.40 | 0.81 |
| InternVL3-78B | 17.77 | 10.36 | 4,991 | 0.41 | 0.80 |
| Llama 4 Scout-17B-16E | 79.04 | 33.79 | 7,570 | 0.32 | 0.73 |
| Qwen2.5-VL-72B | 43.15 | 8.32 | 8,910 | 0.51 | 0.92 |

Table 4: Per-model lexical statistics under greedy and nucleus decoding. ASL = Average Sentence Length (tokens per caption), SDSL = Standard Deviation of Sentence Length, #Types = number of unique word types, TTR1 = type–token ratio (per 1000 tokens), TTR2 = bigram type–token ratio (per 1000 bigrams).