University of California
Santa Barbara

Measurement Challenges for Multimodal Model Capabilities

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Michael Saxon

Committee in charge:

Professor William Yang Wang, Chair
Professor Shiyu Chang
Professor Yao Qin

June 2025

The Dissertation of Michael Saxon is approved.

Professor Shiyu Chang

Professor Yao Qin

Professor William Yang Wang, Committee Chair

June 2025

Measurement Challenges for Multimodal Model Capabilities

Copyright © 2025

by

Michael Saxon

Acknowledgements

I have found the social element of science to be even more fulfilling than research itself. Socialization is not trivial—the most important contribution the modal scientist makes is their engaged presence in their community, serving as one of many minds who form a living repository of knowledge. I am so happy to have made great friends of so many amazing scientists: from labmates next door to collaborators across the globe. I will open by acknowledging them in a few unordered lists.

Starting my Ph.D., the first NLPers I got to know were my labmates Sharon Levy, Wenda Xu, Xinyi Wang, Alon Albalak, Wanrong Zhu, Yi-lin Tuan, and Zhiyu Chen. Starting my degree during the pandemic could have been isolating; commiserating over work and chatting about frivolities remotely with them made it less so. Later I made many more researcher friends at UCSB, including Weixi Feng, Jiachen Li, Liangming Pan, Kexun Zhang, Qiucheng Wu, Gyuwan Kim, Alfonso Amayuelas, Antonis Antoniades, Deepak Nathani, Xiao Pu, Tengxiao Liu, Gurusha Juneja, Kaijie Zhu, Yuxi Xie, Sitao Cheng, Alex Mei, and Avani Tanna.

By participating in conferences, workshops, internships, and symposia I got to make even more great friends in the community, with whom I have been regularly chatting, meeting up, and collaborating. It was a blast befriending and working with Justin Cho, Tejas Srinivasan, Peter West, Isabelle Lee, Simran Khanuja, Shaily Bhatt, Emmy Liu, Shramay Palta, Gowthami Somepalli, Juhyun Oh, Ari Holtzman, Inha Cha, Brihi Joshi, Rupayan Mallick, Naomi Saphra, and Yiran Lawrence Luo on myriad research projects and proposals. To list just a few great friends I've made outside these collaborations, meeting Sayan Ghosh, Janet Liu, Yanai Elazar, Lucy Li, Noah Ziems, and Yoshitomo Matsubara was a pleasure. Honestly, I could list so many more of you but had to keep it short (using a communication frequency heuristic).

I learned a lot through mentorship, with my Early Research Scholars Program mentees Matthew Ho, Aditya Sharma, Vaishnavi Himakunthala, Daniel Rose, Andy Ouyang, and Ryan He; my Fatima Al-Fihri Predoctoral Fellowship mentees Fatima Jahara and Mahsa Khoshnoodi; my other undergraduate mentees Siddharth Yayavaram, Arnav Yayavaram, and Eshaan Tanwar; and my UCSB masters' thesis mentees Alex Mei and Avani Tanna.

I am very grateful to many supportive faculty, both within and without UCSB, who have given me time and effort for research assistance, career advice, and encouragement, including Diba Mirza, Yezhou Yang, Lei Li, Tejas Gokhale, Yejin Choi, Xin Eric Wang, Wenhu Chen, Diyi Yang, Sarah Bargal, Najoung Kim, Greg Durrett, Graham Neubig, Jon May, and my committee members Shiyu Chang and Yao Qin. I was fortunate to work with some great industry collaborators during my Ph.D. as well, including Luca Soldaini, Alessandro Moschitti, Zicheng Liu, and Ximeng Sun.

I could not have entered AI research at all without several pivotal mentors including my early research collaborators Todd Houghton, Hanshuang Liang, Wang Chen, Mingming Li, and Hongbin Yu; my master's thesis mentors, advisors, and collaborators Visar Berisha, Sethuraman Panchanathan, Troy McDaniel, Mark Naufel, Hemanth Venkateswara, Julie Liss, and Meredith Moore; and my Amazon applied science mentors Samridhi Choudhary, Joe McKenna, and Athanasios Mouchtaris.

Of course, I could not have made this dissertation without my advisor William Wang, who welcomed me into the NLP research community and gave instrumental research feedback, moral support, financial support, and guidance of all sorts.

My family and friends have kept me sane. Among many, thank you Alex, Alexander, Wilson, Drea, Serena, Lukas, Jaylia, John, Abhik, Evan, Scott, Karthik, Gab, Dallon, Calvin, Conor, and Chester. Thank you Mom, Dad, Mary, and my many aunts, uncles, and cousins. Thank you all for your support, camraderie, hospitality, and love.

Curriculum Vitæ
Michael Saxon

Education

2025	Ph.D. in Computer Science, University of California, Santa Barbara.
2020	M.S. in Computer Engineering, Arizona State University.
2018	B.S.E. in Electrical Engineering, Arizona State University.

Publications

1. Justin Hyundong Cho, Spencer Lin, Tejas Srinivasan, **Michael Saxon**, Deuksin Kwon, Natali T. Chavez, Jonathan May, "Can Vision Language Models Understand Mimes?" To appear in the Findings of the Association for Computational Linguistics, 2025.
2. Weixi Feng, Jiachen Li, **Michael Saxon**, Tsu-jui Fu, Wenhu Chen, William Yang Wang, "TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation," To appear in Findings of the Association for Computational Linguistics, 2025.
3. **Michael Saxon***, Mahsa Khoshnoodi*, Fatima Jahara*, Yujie Lu, Aditya Sharma, William Yang Wang Wang, "Who Evaluates the Evaluations? Assessing the Faithfulness and Consistency of Text-to-Image Evaluation Metrics with T2IScoreScore," **Proc. Conference on Neural Information Processing Systems (NeurIPS) 2024**, Spotlight, Top 2.5% of 15,000 submissions, Dec 2024.
4. **Michael Saxon***, Aditya Sharma*, William Yang Wang, "Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts,", **Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024**.
5. **Michael Saxon**, Ari Holtzman, Peter West, William Yang Wang, Naomi Saphra, "Benchmarks as Microscopes: A Call for Model Metrology," **First Conference on Language Modeling (COLM) 2024**
6. **Michael Saxon***, Yiran Luo*, Sharon Levy, Chitta Baral, Yezhou Yang, William Yang Wang, "Lost in Translation? Translation Errors and Challenges for Fair Assessment of Text-to-Image Models on Multilingual Concepts," **Proc. North American Chapter of the Association for Computational Linguistics 2024**, *Oral (5% of subs.)* June 2024.
7. Liangming Pan, **Michael Saxon**, Wenda Xu, Deepak Nathani, Xinyi Wang, William Yang Wang, "Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies," **Trans. of the Association for Computational Linguistics (TACL)** May 2024.
8. Vaishnavi Himakunthala*, Andy Ouyang*, Daniel Rose*, Ryan He*, Alex Mei, Yujie Lu, C. Sonar, **Michael Saxon**, William Yang Wang, "Let's Think Frame by Frame with VIP: A Video Infilling and Prediction Dataset for Evaluating Video Chain-of-Thought," **Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023**, Dec 2023

9. Xinyi Wang, Wanrong Zhu, **Michael Saxon**, Mark Steyvers, William Yang Wang, "Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning," **Proc. Conference on Neural Information Processing Systems (NeurIPS) 2023**, Dec 2023
10. **Michael Saxon**, William Yang Wang, "Multilingual Conceptual Coverage in Text-to-Image Models," **Proc. Association for Computational Linguistics; FAccT 2023 Oral** Jul 2023.
11. Yi-lin Tuan, Alon Albalak, Wenda Xu, **Michael Saxon**, Connor Pryor, Lise Getoor, William Yang Wang, "CausalDialogue: Modeling Utterance-level Causality in Conversations," **Findings of the Association for Computational Linguistics**, Jul 2023.
12. Matthew Ho, Aditya Sharma, Justin Chang, **Michael Saxon**, Sharon Levy, Yujie Lu, William Yang Wang, "WikiWhy: Answering and Explaining Cause-and-Effect Questions", to appear in Proceedings of Eleventh International Conference on Learning Representations (**Proc. International Conference on Learning Representations 2023**), Oral Paper: Top 5% out of all 4019 submissions, Kigali, Rwanda, May 1st to 5th.
13. Xinyi Wang, **Michael Saxon**, Jiachen Li, Hongyang Zhang, Kun Zhang, William Yang Wang, "Causal Balancing for Domain Generalization", to appear in Proceedings of Eleventh International Conference on Learning Representations (**Proc. International Conference on Learning Representations 2023**), Kigali, Rwanda, May 1st to 5th.
14. **Michael Saxon**, Xinyi Wang, Wenda Xu and William Yang Wang, "PECO: Examining Single Sentence Label Leakage in Natural Language Inference Datasets through Progressive Evaluation of Cluster Outliers", in Proceedings of The 17th Conference of the European Chapter of the Association for Computational Linguistics (**Proc. European Chapter of the Association for Computational Linguistics**), long paper, Dubrovnik, Croatia, May 2-6 2023, ACL.
15. Wenda Xu, **Michael Saxon**, Misha Sra, William Yang Wang, "Self-Supervised Knowledge Assimilation for Expert-Layman Text Style Transfer", to appear in Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (**AAAI 2022**), long paper, Vancouver, BC, Canada.
16. Xinyi Wang, Wenhu Chen, **Michael Saxon**, William Yang Wang, "Counterfactual Maximum Likelihood Estimation for Training Deep Networks", (**Proc. Conference on Neural Information Processing Systems (NeurIPS) 2021**), long paper, online.
17. **Michael Saxon**, Sharon Levy, Xinyi Wang, Alon Albalak and William Yang Wang, "Modeling Disclosive Transparency in NLP Application Descriptions", to appear in (**Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021**), *Oral (8% of subs.)* long paper, online, ACL.
18. **Michael Saxon**, Samridhi Choudhary, Joe McKenna, Athanasios Mouchtaris, "End-to-End Spoken Language Understanding for Generalized Voice Assistants," **Interspeech 2021**, pp. 4738–4742.
19. Sharon Levy, **Michael Saxon** and William Yang Wang, "Investigating Memorization of Conspiracy Theories in Text Generation", to appear in Findings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the

11th International Joint Conference on Natural Language Processing (**Findings of ACL-IJCNLP 2021**), long paper, online, ACL.

20. **Michael Saxon**, Ayush Tripathi, Yishan Jiao, Julie Liss, Visar Berisha, "Robust Estimation of Hypernasality in Dysarthria," **IEEE Trans. on Audio, Speech, and Language Processing** 2020, Vol. 28, pp. 2511–2522.
21. **Michael Saxon***, Joe McKenna*, Samridhi Choudhary*, Grant Strimel, Athanasios Mouchtaris, "Semantic Complexity in End-to-End Spoken Language Understanding," **Interspeech 2020**, pp. 4273–4277.
22. Meredith Moore, P. Papreja, **Michael Saxon**, Visar Berisha, Sethuraman Panchanathan, "UncommonVoice: A Crowdsourced Dataset of Dysphonic Speech," **Interspeech 2020**, pp. 2532–2536.
23. Meredith Moore, **Michael Saxon**, Hemanth Venkateswara, Visar Berisha, Sethuraman Panchanathan, "Say what? A dataset for exploring the error patterns that two ASR engines make," **Interspeech 2019**, pp. 2528–2532.
24. **Michael Saxon**, Julie Liss, Visar Berisha, "Objective Measures of Plosive Nasalization in Hypernasal Speech," 2019 **IEEE ICASSP 2019**, pp. 6520–6524.
25. Todd Houghton, **Michael Saxon**, Zeming Song, Hoa Nyugen, Hanqing Jiang and Hongbin Yu, "2D Grating Pitch Mapping of a through Silicon Via (TSV) and Solder Ball Interconnect Region Using Laser Diffraction" **IEEE 66th Electronic Components and Technology Conference (ECTC) 2016**, pp. 2222–2227. *(Texas Instruments Best Student Interactive Paper Award)*

Select Honors

- **Rising Star in Generative AI**, 1/9 selectees, UMass Amherst (2024)
- **Google Ph.D. Fellowship Nominee**, 1/4 selected by UC Santa Barbara (2024)
- **Neal Fenzi—Resonant Founder Fellowship**, UC Santa Barbara (2024)
- **Outstanding Reviewer Award**, ACL (2023)
- **NSF Graduate Research Fellowship Award**, NSF GRFP (2020)
- **Responsible ML Fellowship**, UCSB Center for Responsible Machine Learning (2020)

Abstract

Measurement Challenges for Multimodal Model Capabilities

by

Michael Saxon

A fundamental open question in AI research is: "how can we confidently know what a model can reliably do?" Currently, we often discover model limitations through post-deployment failures rather than proactive assessment. I will present novel methodologies for rigorous pre-deployment evaluation, focusing on generative text-to-image models in multilingual and multicultural contexts. I will introduce new frameworks for analyzing these models' cross-cultural knowledge representation and performance, presenting previously unidentified optimization challenges in balancing multiple competing objectives: cross-language consistency, preservation of cultural nuance, and mitigation of harmful biases. Through careful empirical analysis, this work has revealed significant gaps in existing evaluation approaches and challenged several commonly held assumptions about model capabilities. The methodologies introduced here extend beyond text-to-image models, offering broader insights into AI system assessment. I will demonstrate how more precise evaluation techniques can uncover unexpected failure modes and provide clearer pathways for future research in reliable AI development, leading to better empirical progress, scientific understanding, developer informedness, and public awareness.

Contents

Curriculum Vitae **vi**

Abstract **ix**

1 Introduction **1**
1.1 The measurement problem in AI 3
1.2 Measurement challenges in multimodal systems 4
1.3 Multilingual and multicultural multimodal systems 5
1.4 Permissions and Attributions 6

Part I The measurement problem in AI **7**

2 Model metrology **8**
2.1 Problems with current benchmarks for LMs 9
2.2 Qualities of useful, concrete benchmarks 14
2.3 Implications for my research 15

3 Analyzing datasets **16**
3.1 Quantifying NLI Dataset Bias 20
3.2 Analyzing NLI Dataset Bias 28
3.3 Results 34
3.4 Discussion 36
3.5 Conclusion 40

Part II Measurement challenges in multimodal systems **41**

4 Text-to-image faithfulness **42**
4.1 Background 45
4.2 `T2IScoreScore` meta-metrics 47
4.3 The `T2IScoreScore` Dataset 50

4.4 Experiments . . . 53
4.5 Results . . . 56
4.6 Discussion . . . 60
4.7 Conclusion . . . 64

5 Long-context vision-language models 65
5.1 LoCoVQA Generation Method . . . 68
5.2 Experiments . . . 72
5.3 Results . . . 74
5.4 Discussion . . . 76
5.5 Conclusion . . . 78

Part III Multilingual and multicultural multimodal systems 79

6 Multilingual text-to-image models 80
6.1 Introduction . . . 82
6.2 Motivation & Related Work . . . 84
6.3 Definitions & Formulations . . . 86
6.4 Approach . . . 89
6.5 Findings . . . 94
6.6 Conclusion . . . 101

7 Problems in multilingual evaluation 103
7.1 Introduction . . . 104
7.2 Motivation & Approach . . . 105
7.3 Results & Analyses . . . 108
7.4 Discussion & Conclusions . . . 111
7.5 Supplementary materials . . . 115

Part IV Conclusion 122

8 Conclusion 123
8.1 Building an Evaluation Discipline . . . 124
8.2 Future Directions . . . 125

Bibliography 128

Chapter 1

Introduction

It is trite to say that we are living through a revolution in artificial intelligence. This idea is pervasive for a reason—language models, multimodal language & vision systems, and other forms of "generative AI" are being developed, improved, and adopted at a rapid clip. Many of the technologies we refer to as AI indeed have real promise for advancing the pace of technological innovation, freeing knowledge workers from mundane tasks, and building an abundant future. However, these technologies also have the potential for real peril. The tools that can automate mundane paperwork may be the tools for mass layoffs. The abundance brought about by their deployment might drive severe inequality. The same models that can synthesize complex writings from myriad sources are a dream tool for an authoritarian, enabling a citizen's every word to be scrutinized by a machine minder.

It is incumbant upon us as computer scientists to further a *rigorous understanding of what these models can do*—before wide deployment—to inform the public, developers, and policymakers in the interest of ensuring the perils are averted. While has been a big part of the "why" I have pursued a Ph.D, these huge questions cannot be answered conclusively in a single doctoral dissertation.

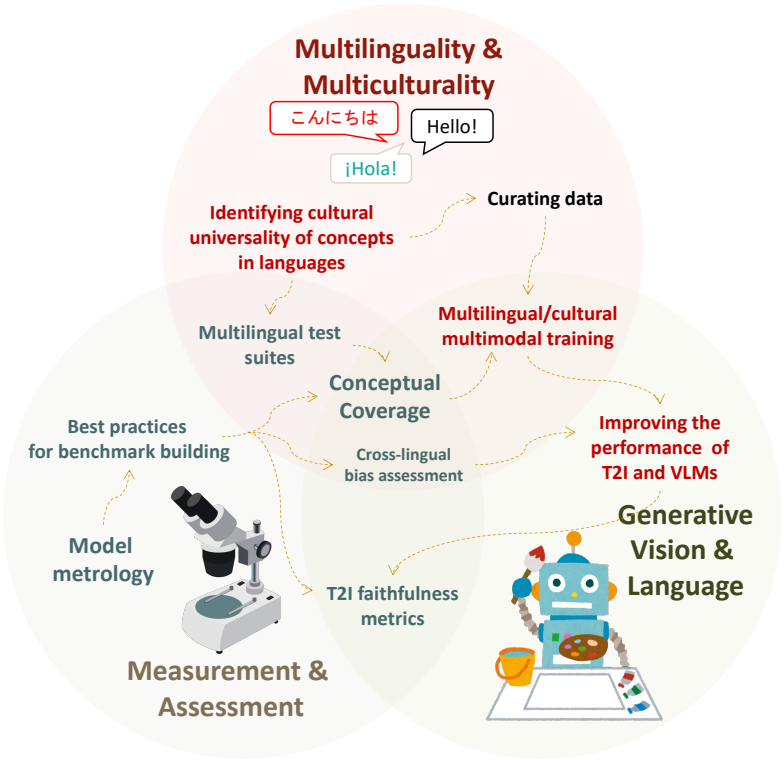


Figure 1.1: An overview of the interlinked foci of this dissertation.

Consequently, I have pursued a few approachable directions toward this greater vision of **scientifically rigorous, practical evaluation of natural language systems**. In this dissertation I will present findings from six conference-published research works toward three overarching directions:

1. **The measurement problem in AI**: what these rigorous evaluations should look like, and how to build them.

2. **Measurement challenges in multimodal systems**: identifying core issues with evaluations in generative vision & language models, and proposing fixes.

3. **Multilingual and multicultural multimodal systems**: making generative vision & language systems ethically accessible for users from around the world.

1.1 The measurement problem in AI

Recent striking advancements in generative AI systems such as language models (LMs) and text-to-image (T2I) models have driven both excitement and concern from the public. By outputting lengthy and coherent text, detailed imagery, or convincing and useful answers to diverse questions, these systems exhibit intriguing capabilities—yet they also continue to yield puzzling failure modes which confound easily analogizing these capabilities to abstract elements of human intelligence. The prevailing method of AI capability assessment nowadays involves gathering benchmark datasets—often based on academic or professional tests for humans—and simply assuming that high accuracy on them represents a transferable measure of one of these high-level, abstract abilities such as "understanding" or "reasoning." Unfortunately, this method is fundamentally flawed. While the community recognizes this issue, it has yet to overcome it. **Part I** addresses these issues.

First, in **Chapter 2** I provide an in-depth argument that LM evaluation research is moving toward being an organized, mature discipline, and provide discussion of how this can be achieved and how it would allow us to address the issues mentioned above. In particular, I argue that scientifically-sound evaluations of LM capabilities must be *dynamic* and *constrained*. Additionally, the piece calls for more *meta-evaluation research*: work that interrogates the quality of evaluation resources to drive insights into making better ones.

In **Chapter 3** I present an example meta-evaluation project in natural language inference (NLI). I demonstrate a simple, two-model approach to remove statistical cheating feature biases from NLI datasets.

I carry the philosophy of these works forward into the remaining chapters.

1.2 Measurement challenges in multimodal systems

Multimodal generative AI systems present unique evaluation challenges. Vision-language models (VLMs) are a class of LMs that can ingest images within the input token stream, enabling them to perform generalist multimodal tasks such as captioning and visual question answering (VQA). Text-to-image models (T2I) synthesize images from textual prompts.

Both of these system classes rely on learned interfaces between the vision and language modalities, which are extremely hard to train robustly. Abundant web captions used for training these interfaces are bound by human notions of salience. While an avid birdwatcher may note the species name in every photo, she will probably not put in her caption which side of the nest the bird is on. Consequently, a vision-language model will have a much easier time identifying a roadrunner as *Geococcyx californianus* than it will deciding if that bird is to the left or the right of a stick. Opportunities to identify and mitigate such failings are thus abundant.

In **Part II** I present two works performing such meta-evaluation in the vision & language domain. In **Chapter 4** I present work on T2I faithfulness metrics, identifying a crucial flaw in how these metrics have been evaluated—nobody evaluated how well these error-detecting metrics can actually order images by counts of errors. This work overturned the conventional wisdom of T2I evaluation by demonstrating that the supposed best metrics, using VLMs as their backbones, conferred negligible performance gains over much simpler and cheaper baselines.

In **Chapter 5** I will present work that pokes holes in the popular "needle in a haystack" evaluation paradigm for long-context VLMs. We demonstrate that frontier VLMs have strikingly bad performance under simple visual distractors.

1.3 Multilingual and multicultural multimodal systems

Multimodal systems are globally deployed. It is important that we ensure their accessibility to all users, regardless of language. In the already-complicated space of multimodal model evaluation, in *multilingual* settings there are even more unforseen edge behaviors to identify and mitigate. In **Part III** I present a line of work that aims to assess these failures rigorously.

In **Chapter 6** I introduce *CoCo-CroLa*, the first benchmark for multilingual conceptual knowledge in text-to-image models. CoCo-CroLa meets many of the desiderata lain out in Chapter 2. The system prompts any T2I model under test as a black box, sampling populations of images for some concept in a language, and compares this population to populations for other concepts and languages to identify the correctness, consistency, and distinctiveness of a model's "understanding" of each concept. Though CoCo-CroLa tests simple tangible noun knowledge, its metrics predict generalization to complex prompts. A high possession score is necessary for a model to be able to use some object in novel, creative imagery when prompted in said language. Thus CoCo-CroLa even has value when it saturates by showing multilingual competence of any saturating model. Additionally, CoCo-CroLa is dynamically and automatically generated, allowing it to be expanded to any language or new set of concepts with minimal human effort.

Unfortunately, no system is perfect. In **Chapter 7** I turn the critical lens applied in Part II on my own work, and identify flaws in the dynamic generation pipeline which cause CoCo-CroLa to contain small mistranslations and word sense errors. I diagnose the causes of these errors and characterize their severity, enabling us to build a stronger version of the benchmark.

1.4 Permissions and Attributions

This paper contains material from six peer-reviewed archival conference publications.

Chapter 2 contains *Benchmarks as Microscopes: A Call for Model Metrology*, presented at the 1st Conference on Language Modeling, (COLM 2024).

Chapter 3 contains *PECO: Examining Single Sentence Label Leakage in Natural Language Inference Datasets through Progressive Evaluation of Cluster Outliers*, presented at the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023).

Chapter 4 contains *Who Evaluates the Evaluations? Objectively Scoring Text-to-Image Prompt Coherence Metrics with T2IScoreScore (TS2)*, a Spotlight presentation at the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024).

Chapter 5 contains *Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts*, published in the 2024 Findings of the Association for Computational Linguistics and presented at EMNLP 2024.

Chapter 6 contains *Multilingual Conceptual Coverage in Text-to-Image Models*, presented at the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

Chapter 7 contains *Lost in Translation? Translation Errors and Challenges for Fair Assessment of Text-to-Image Models on Multilingual Concepts*, an oral presentation at the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024).

All graphics are either produced by me, from prior datasets distributed under an open license, or reproduced from other scientific presentations with attribution.

Part I

The measurement problem in AI

Chapter 2

Model metrology

I open my dissertation with excerpts from a position piece laying out a research philosophy and vision for how we can make evaluation research more useful and systematized. Subsequent chapters will present technical work focused on evaluation—of language models, multimodal systems, or evaluation resources themselves—which draw on intellectual themes presented in this chapter.

Just how good are current language models? It's hard to say. Even the latest benchmarks are not scalable, relevant, or durable enough to predict performance in real-world settings. Although engineering [1, 2], research [3, 4], and policy [5, 6] decisions are grounded in the supposed capabilities of AI systems—particularly language models [7, 6]—even experts disagree about the nature [8, 9] and extent [10, 11, 12] of these capabilities.

Most LM breakthroughs are judged by *aggregate performance on narrow benchmarks* [13] or *one-off manual analyses* [14]. These assessments inform conversations around scaling [15], risk [16], and deployability. Popular static benchmarks inevitably **saturate** [17, 18] as each consecutive generation of models over-optimizes for performance on its evaluation sets—exacerbated by those datasets contaminating future training data—

without resolving fundamental impasses over the nature of LMs or usefully informing their deployment. *We need evaluations practices that yield meaningful observations.*

How measurement matured in astronomy

In 1609, Galileo Galilei built one of the first optical telescopes. When he turned it around, Galileo found that he could also observe very small objects close up—and so microscopy was born [19]. For centuries, these tools were built by the same scientists using them to make fundamental discoveries [20].

Eventually, the expertise required to design the precise tools to advance science outstripped scientists' glassworking skills. By the 20th century, large teams of specialists built orbital space telescopes [21] and microscopes became mass-manufactured commodities [22]. The science and engineering of measurement tools have coalesced into specialized disciplines.

At present, our LM evaluation practices resemble the state of astronomy and microbiology in the early 17th century—the same community analyzing the object of study (models) is also building the tools for that analysis (benchmarks). LMs require an independent professional and scientific community dedicated to building analytic tools just as microscope and telescope building have. In this chapter, I will briefly characterize the shortcomings of current evaluation practices and lay out concrete desiderata for improved evaluations which I carry forward into my work throughout.

2.1 Problems with current benchmarks for LMs

While benchmarks have long driven AI progress, they are now used to support increasingly grandiose claims. When research communities believe that "solving"

a benchmark represents core progress toward generalized intelligence, interest and investment naturally follow. Raji et al. [23] document how the *common task framework*—public contests between systems assessed on common train and test sets [24]—enabled advancements in concrete and tightly-scoped problems such as automatic speech recognition and machine translation, but has since been inappropriately extended to claim generalized capabilities in pretrained vision [25] and language [26] models.

Current LM practices have shifted from training and testing on specific benchmark datasets to testing models in a zero-shot setting. Consequently, many researchers assume that *because* LMs aren't deliberately trained on task-specific train sets [27, 28], performance on these benchmarks is *stronger evidence* of general capability than for fine-tuned models [29, 30]. Though these assumptions are controversial—evidenced by the remarkably divergent perspectives of NLP and AI researchers [31]—the glamour of the promise of general intelligence has carried claims of generalized intelligence to a credulous public [32], driving anxieties over AI risk [33].

These attempts to assess general capabilities address a legitimate need: evaluation is important for guiding advances and comparing models [34]. Because modern LMs are used as *everything systems*, we may naturally wish to characterize their general capabilities [9]. However, *when attempting to assess general capabilities, evaluations neither capture general competency nor predict performance on many downstream applications*.

2.1.1 Generalized capabilities are hard to define and contentious.

Narrow LM benchmarks are often derived either from tests for humans—e.g., GSM8K [35], MMLU [36]—or from *existing constrained benchmarks* for specific engineering problems in natural language processing such as question answering [37] or entailment recognition [38]. Performance on these benchmarks often predicts

performance on similar test sets on the same task, to the degree that a violation of this expectation can be evidence of test-set contamination [39, 40]. However, claims regarding the real-world reliability of these systems are often unsupported [41]. We believe these benchmarks cannot characterize broader capabilities, even when aggregated.

Critics of generalized capabilities benchmarks note that they lack **construct validity**—strong evidence that any evaluation represents a capability [42] that they claim to measure [43]. This discordance between metrics and resulting claims was already present for fine-tuned model evaluation [23] but has since worsened, as LM developers and their allies claim generalized intelligence [14], often based on a huge set of limited benchmark scores [44, 13]. Similar claims are made for more specific abstract capabilities when researchers attribute benchmark performance to faculties like abstract reasoning [45], language understanding [46], or common sense knowledge [47].

We are at an impasse. Though these claims of generality are contested [48], they are hard to conclusively disprove. Bender argues axiomatically that understanding cannot be acquired through the LM objective [11]. Poor generalization across time [49], tasks [50], and heuristics [51] have also been provided as counterarguments to LM capability claims. Given that humans struggle to evaluate intelligence in other animals [52], how can we assess slippery abstract notions like reasoning [53] in AI?

2.1.2 Benchmarks can aim for generality—or be valid and useful.

When benchmarks claim to test abstract capabilities, critics often question whether that capability is *necessary* or just *sufficient* for their solution [54]. It is probably impossible to prove whether most tasks actually require some abstract capability. After all, both a studious human scholar and an answer key achieve 100% accuracy on an exam, but a piece of paper clearly does not possess the scholar's intelligence.

Consider a developer of a real-world application based on an LM, which we dub *builder-consumers*. For a builder-consumer, a useful benchmark must simply test if a system—regardless of abstract capabilities—is performant on their task. Meaningfully representing the deployment setting makes a benchmark **ecologically valid** [55].

Are generality and ecological validity fundamentally in tension? Recent efforts to unify these goals such as HELM [56] provide a large collection of scores by harvesting existing benchmarks divided into specific scenarios (e.g., news domain tasks). By contextualizing these tasks within categories—and providing disaggregated scores over them—they aim to preserve task-level construct validity [57] while capturing a "holistic" view of an LM's capabilities. Though many of these tasks may have construct validity, these holistic evaluation attempts do not capture generality—each benchmark represents a tiny view of a broader "task universe" [41]. Ecological validity is particularly problematic when discussing "AGI-level" capabilities [9], though discussing the credibility of those notions is not central here.

Developers are building applications which rely on predictable LM behavior and therefore need to evaluate consistent, constrained capabilities for their domain. However, developers such as OpenAI [13], Anthropic [58], and Google [59, 60] instead continue to focus on the same MMLU, GSM8K, and HumanEval test suites rather than on application-specific tests. These platforms provide API access to these models as a paid service, so *why aren't they benchmarking customer-relevant capabilities?* Maybe they are convinced that general intelligence is quantifiable by these benchmarks. Perhaps the deeper issue is that building bespoke evaluations is hard, and the domains are innumerable—will a collection of constrained tests ever be large enough to placate critics? Do scientists have a role to play in producing these gap-closing evaluations?

2.1.3 Why do we keep using flawed benchmarks?

These criticisms are not novel—indeed, they're commonly expressed sentiments. Regardless, these flawed benchmarks remain dominant. Saturation is broadly acknowledged as a problem whose mechanisms present a fundamental challenge to benchmark validity. GSM8K, long used to assess mathematical reasoning, is fully saturated for "frontier" models—GPT-4 achieves near-100% accuracy with prompting and decoding tricks [61]. We know models overfit even to hidden (but static) test sets [62]. We know that the long tail of incomprehensibly large pretraining datasets [63, 64] inevitably enables answer memorization [65] or heuristic learning [66, 67, 68, 69] through similar examples [70, 71]. Poor construct validity is widely noted [72], as is the futility of measuring generalized capabilities [73]. So why do we still rely on these benchmarks? Perhaps:

1. Misalignment of interest/incentives for researchers and needs of users.

2. Fundamental difficulties in building benchmarks that meet our desiderata.

3. The allure of general intelligence attracts public, media, and investor attention.

As long as the LM community is the primary benchmark building community, these problems will persist. Academic researchers build benchmarks for 'impact' in citations and public use [74], and the primary incentive for industrial actors is to demonstrate the superiority of their latest product. Unsurprisingly, 'top benchmarks' from a small number of elite institutions have become primary measures of AI progress [75]. In the short term, a carefully-scoped and rigorously designed benchmark is as impactful as a well-hyped but soon-to-saturate one—but the former is much harder to make than the latter. The development of best practices for benchmark building cannot rely on the incentives of scientific machine learning research.

2.2 Qualities of useful, concrete benchmarks

The problems above can only be solved if we *abandon general metrics when making deployment decisions*. We need a benchmarking culture and practice that empowers consumers to *specify their desired constraints* and *generate their own benchmarks*. Model producers should track progress by these scenario-driven evaluations. Good LM capability evaluations are:

Constrained, or characterizing model behavior on a concrete task where domain experts can describe the boundaries a good system should stay within. Often, concrete problems can be defined in terms of verifiable rules. Given these boundaries as an objective, issues of scenario innumerability and ecological validity become much more tractable, as we can rely on domain expert understanding of real-world inputs.

Dynamic. A fixed dataset is easily memorized. To avoid stale metrics, we prefer dynamically generated test simulations where performance is measured by task-specific constraints.

Plug-and-play. Benchmark generating processes must be plug-and-play, i.e., easy to run and adapt to new systems. Expensive, one-off benchmarks are unlikely to foster open academic discourse—or provide value to downstream developers—and human preferences are expensive to gather and difficult to replicate. Accessible evaluation setups will entice domain experts to apply and even publish their constraints.

These desiderata are mutually reinforcing and enabled by *abandoning generality* and centering the needs of real-world builder-consumers and end users. Because model developers and AI researchers usually aim to improve general capabilities, focused real-world evaluation must instead rely on dedicated practitioners, i.e. metrologists.

2.3 Implications for my research

The AI community is sharing observations about LMs too fast for any one researcher to follow. Given their role in disseminating information about advances through the community, flawed benchmarks and metrics lead researchers to waste time and effort developing methods that are ill-suited to any real-world application [76]. A mindset shift from *evaluation for NLP research* to *evaluation research in NLP* must take place to produce evaluations and benchmarks that satisfy the desiderata above[1]. How does this vision shape my own research? My research is aimed at answering fundamental LM evaluation questions in specific, constrained contexts, such as:

What techniques can we use to refine flawed evaluation resources? **Chapter 3** details a such a technique, which iteratively trains paired models and uses differences in their sample-level judgments to eliminate samples exhibiting cheating spurious correlations in natural language inference datasets.

How can we demonstrate the suitability of an evaluation for a specific capability? **Part II** presents meta-evaluation work analyzing and critiquing benchmarking and evaluation methods for multimodal models. Specifically, I will present conventional wisdom-overturning results on the relative performance of text-to-image faithfulness metrics (**Chapter 4**) and demonstrations of the surprisingly poor long-context understanding exhibited by vision-language models (**Chapter 5**).

How can we frame evaluative targets in a realistic and feasible manner? In **Part III** I will present evaluation work which extends best-practice dynamic benchmarking methods into a previously unexplored area—quantifying the multilingual understanding of text-to-image models (**Chapter 6**)—and perform meta-evaluation on this very resource which I produced (**Chapter 7**).

[1]For a full treatment of this argument, see *Benchmarks as Microscopes: A Call for Model Metrology* [77]

Chapter 3

Analyzing datasets

Having established the **need for and utility of** research on scientifically developing and understanding evaluations, I will now present a practical research project which addresses a common limitation of static benchmarks—cheating shortcut features.

Natural language inference (NLI) is a fundamentally pairwise task, wherein a logical relation between two statements is predicted. Progress on NLI benchmarks was long considered an important proxy for advancements in natural language reasoning by machines. In NLI, models are trained to process the two statements simultaneously, in a *paired sentence condition* (PSC). Unfortunately many modern NLI datasets exhibit **single sentence label leakage**. When this leakage is present, models are able to accurately predict the pairwise relation encoded by the labels in a *single sentence condition* (SSC)—where the model is only shown one of the statements [66]. This is a serious problem, rendering NLI datasets' capture of reasoning questionable, and limiting the robustness of models trained on them.

NLI is formalized as predicting a relation $r \in$\{neutral, entail, contradict\} from a pair of sentences (premise s_1 and hypothesis s_2). An ideal NLI benchmark without single sentence label leakage will have distribution of r that is conditionally dependent on

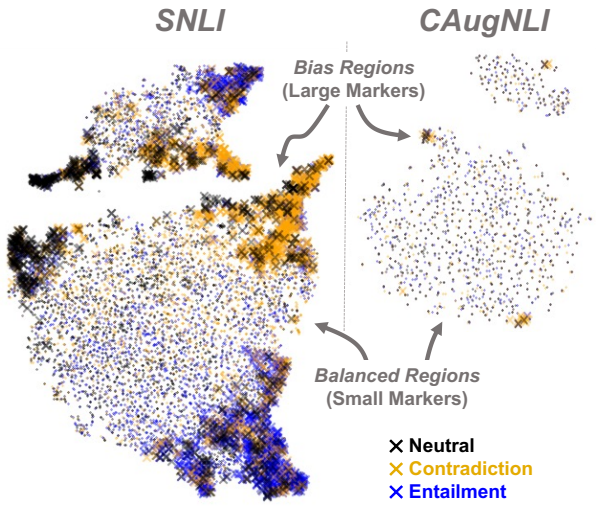


Figure 3.1: A T-SNE projection of the SNLI and CAugNLI test sets in PECO's model–driven *single sentence condition* (SSC) embedding space, showing **entailment**-, **contradiction**-, and **neutral**-labeled samples. This model was trained on the paired-sentence condition (PSC) where the relation between sentences in a sample are observable. In a leakage-free dataset, a model should be unable to separate subpopulations that disproportionately exhibit one label class in the SSC. Local regions exhibiting an imbalanced label distribution in this subpopulation are considered "biased regions" and plotted with large markers. SNLI exhibits large, continuous regions in the hypothesis only space disproportionately exhibiting the same label, compared to CAugNLI. Accordingly, SNLI exhibits higher single sentence label leakage than CAugNLI (table 3.3).

the **pair** of sentences, but independent from either individual sentence [68]. In practice this is difficult to achieve, particularly when constructing usefully large datasets.

Most large-scale NLI datasets are produced by sourcing *seed* sentences from an existing text population to serve as initial premises or hypotheses. Each seed sentence is then assigned one or more relations r, which annotators use to write new *elicited* sentences satisfying each selected relation relative to the seed. Many datasets exclusively build either the hypothesis or premise population from seed sentences, leaving the other exclusively elicited.

Systematic, shared biases in the words, sentence structures, or ideas that crowd-workers consider when given a logical relation, coupled with the exclusively elicited

Dataset	Reference	Seed sentences from	Language	Imbalance on	# Train	# Dev	# Test
SICK	[78]	Image+Video Captions	En	s_2	4.4k	0.5k	4.9k
SNLI	[79]	Image Captions + KB	En	s_2	550k	10k	10k
MNLI	[80]	Multiple Genre	En	s_2	393k	20k	20k
↪MNLI-m					—	10k	10k
↪MNLI-u					—	10k	10k
XNLI	[81]	MNLI	14 Langs.	s_2	—	70k	35k
$\text{FEVER}_{\text{NLI}}$	[82]	Wikipedia	En	s_1	208k	20k	20k
ANLI	[83]	Wikipedia+HotpotQA	En	s_2	163k	3.2k	3.2k
↪A1					17k	1k	1k
↪A2					45k	1k	1k
↪A3					100k	1.2k	1.2k
OCNLI	[84]	Multiple Genre	Zh	s_2	47k	3k	3k
CAugNLI	[85]	IMDb+SNLI	En	s_2 & s_1	8.3k	1k	8.3k
$\text{SNLI}_{\text{debiased}}$	[86]	SNLI, generated	En	s_2 & s_1	1.14M	—	—
$\text{MNLI}_{\text{debiased}}$		MNLI, generated			744k	—	—

Table 3.1: Information on the NLI datasets we compare in this study. MNLI has "matched" (m) and "unmatched" (u) test sets that we evaluate separately. The ANLI dataset is decomposed into partitions "A1," "A2," and "A3". A Vertical line denotes identical value in cell as above (sub-elements of same dataset).

nature of one of the sentence populations, then drive relation leakage [87]. For example, a slight preference for words like "not" or "doesn't" when given *contradict* as opposed to *entail* would lead to a bias in the n-gram distribution between the classes in the SSC. Simple heuristics inspired by these findings can produce challenging test sets that hobble models trained on these biased datasets [67], but they require manual guesswork, don't generalize, and may miss higher-level, more nuanced semantic shortcuts and biases.

These "leakage features" encoded in the elicited sentences are visible to NLI models [88], enabling them to "cheat" by attending to them as shortcuts rather than logical correspondences between the two sentences, calling into question the appropriateness of NLI datasets as benchmarks for language understanding [89]. In this work we rigorously analyze this problem of single sentence relation leakage in both popular and recent NLI datasets using novel techniques to enable targeted interventions and create higher quality future resources.

Further NLI datasets have been proposed to tackle these problems using machine-in-the-loop adversarial sentence elicitation, [83], counterfactual augmentation [85], and learning dynamic-based debiasing [86]. These datasets are purported to provide more challenging generalization scenarios for NLI models to better test logical reasoning capabilities. One big question remains—*have these techniques actually eliminated relation leakage biases?*

In this chapter, we demonstrate the following:

New (as of 2023) NLI datasets continued to exhibit single sentence relation leakage. We compare SSC performance for 10 NLI datasets (including those previously assessed by [66]) using a simple transformer baseline, finding that **single sentence relation leakage remains a severe problem**.

NLI models still use the leakage features to cheat. We analyze the datasets using *output decision agreement* and *input token importance* statistics between models trained in the SSC and PSC to demonstrate this.

Automated leakage feature detection is feasible. We introduce a novel model-based metric and dataset analysis tool, the *Progressive Evaluation of Cluster Outliers* (PECO) (chapter 3), for examining the degree of single sentence relation leakage and eliminating it in future datasets[1].

[1]Code at `github.com/michaelsaxon/DatasetAnalysis`, and an animated demo is available at `saxon.me/peco`.

			Accuracy (% PSC)			Hparams	
Dataset	SOTA Model	Replication Model	SOTA	Ours	Maj.	LR	Batch
SICK	NeuralLog [90]	`roberta-large`	90.3	87.8	56.0	1e–5	18
SNLI	EFL [91]	`roberta-large`	93.1	90.6	33.8	5e–6	128
MNLI-u	T5 [92]	`roberta-large`	92.0	88.7	35.6	5e–6	12
MNLI-b			91.7	88.3	36.5	5e–6	12
XNLI	ByT5 [93]	`xlm-roberta`	83.7	73.7	33.3	1e–5	16
FEVER	KILT [94]	`roberta-large`	86.3	74.7	33.3	1e–5	8
ANLI	InfoBERT [95]	`roberta-large`	58.3	54.0	33.5	1e–6	18
↪A1			75.5	61.9	33.4	1e–6	18
↪A2	ALBERT [96]		58.6	50.0	33.4	1e–6	18
↪A3			53.4	49.8	33.5	1e–6	18
OCNLI	`RoBERTa-wwm-ext-l` [97]	`bert-base-chinese`	78.2	71.8	36.8	1e–5	128
CAugNLI	— (Ours)	`roberta-large`	84.7	84.7	33.9	5e–6	64
$\text{SNLI}_{\text{debiased}}$		`roberta-large`	95.6	95.6	35.5	1e–5	64
$\text{MNLI}_{\text{debiased}}$		`roberta-large`	96.9	96.9	36.2	1e–5	16

Table 3.2: For each dataset we analyze, the current state of the art (SOTA) model, base pretrained LM we use for replication, and accuracy for SOTA, our replication (Ours), and majority label-only (Maj.) classification in the standard paired-sentence NLI classification condition (PSC) with selected best-performing hyperparameters.

3.1 Quantifying NLI Dataset Bias

An ideal NLI benchmark is neither "saturated" nor biased. Saturated benchmarks are datasets for which current approaches already achieve high accuracy. They are "solved" and have limited utility in tracking future progress [89, 77]. We refer to as *biased* any NLI benchmarks that exhibit significant single-sentence relation leakage through high achievable SSC accuracy in at least one single sentence condition.

We analyze 10 datasets containing a total of 14 test or validation sets in terms of *biasedness* and *saturatedness*, across 17 SSC conditions (premise-only (s_1) or hypothesis-only (s_2)). table 3.1 provides an overview of this information along with statistics such as train/dev/test set size, and which sentence population is potentially unbalanced.

Each dataset D is composed of (s_1, s_2, r) tuples. We use the notation $(X_i, Y_i) \leftarrow D$ to describe the samples. Depending on whether the training condition is PSC or SSC, each X_i can be (s_{1i}, s_{2i}), s_{1i}, or s_{2i}. Models trained in condition c are referred to as f_c.

3.1.1 Saturation and Bias Scoring

We assess accuracy on the test set (or validation set when no labeled test set is available) for each dataset in paired- and single-sentence conditions.

Saturation (Accuracy): We report state-of-the-art (SOTA) model performance results in the PSC,

$$\mathrm{A}_{\mathrm{SOTA}}(D) = P(f_{\mathrm{SOTA}}(X_{\mathrm{test}}) = Y_{\mathrm{test}}); \quad X, Y \in D \tag{3.1}$$

For our model-level and sample-level comparative analysis between the PSC and SSC we train our own transformer-based models using a simple procedure (Sec. 3.1.5) to assess *replication* accuracy,

$$\mathrm{A}_{\mathrm{PSC}}(D) = P(f_{\mathrm{PSC}}(X_{\mathrm{test}}) = Y_{\mathrm{test}}); \quad X, Y \in D \tag{3.2}$$

SSC Accuracy: For each elicited population SSC we train a model f_{SSC} according to the procedure described in §3.1.5, to assess $\mathrm{A}_{\mathrm{SSC}}$:

$$\mathrm{A}_{\mathrm{SSC}}(D) = P(f_{\mathrm{SSC}}(X_{\mathrm{test}}) = Y_{\mathrm{test}}); \quad X, Y \in D \tag{3.3}$$

Table 3.2 shows current SOTA models and results for the 10 datasets, as well as our PSC model performance and the relevant training hyperparameters (more detail in §3.1.5). Figure 3.2 shows SSC accuracy against replication (PSC) accuracy for each dataset. Datasets that exhibit higher SSC accuracy have worse single sentence relation leakage, and are thereby questionable in their ability to capture reasoning abilities. Ideally, an **optimal benchmark** for NLI will both have low maximum SSC accuracy and low maximum PSC accuracy (room for future model growth).

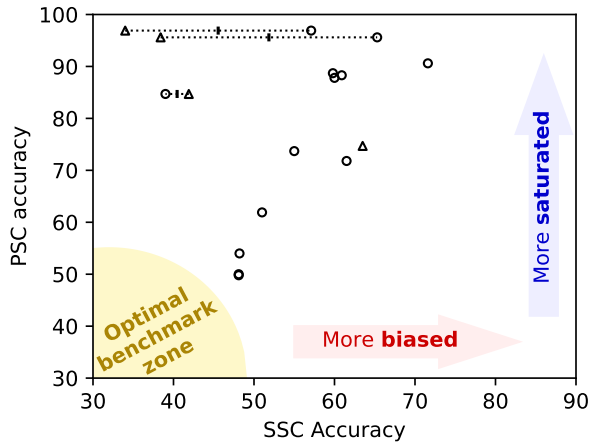


Figure 3.2: Where the existing NLI datasets fall on the SSC accuracy/PSC accuracy axis. An ideal benchmark isn't saturated (lower SOTA accuracy) and also is unbiased (low SSC accuracy). △ markers represent premise-only SSC results and ∘ markers represent hypothesis-only SSCs. For datasets exhibiting leakage in both conditions, the two points are connected by a dotted line with a | marker for their average SSC accuracy.

These absolute measures of dataset bias and saturation are useful targets for future optimal benchmarks, it is important to understand how these measures interact with each other.

3.1.2 Relative Dataset Bias Scoring

We assess two relative dataset bias scores.

SSC Improvement over Chance: We subtract the SSC test accuracy from the accuracy achieved by a "guess majority label" strategy following [66]:

$$\Delta_{\text{maj}} = P(f_{\text{SSC}}(X_{\text{test}}) = Y_{\text{test}}) - P(Y_{\text{maj}} = Y_{\text{test}}) \tag{3.4}$$

This metric gives an insight into single sentence relation leakage that compensates for datasets (such as SICK) with an uneven base label distribution.

Dataset	Cond.	SSC	$\%R_R$	Δ_{maj}
SICK	s_2	60.0	68.3	4.0
SNLI	s_2	71.6	79.0	37.8
MNLI-b	s_2	59.8	67.4	24.2
MNLI-u	|	60.9	69.0	24.4
XNLI	s_2	55.0	74.6	21.7
FEVER	s_1	63.5	85.0	30.2
ANLI	s_2	48.2	89.3	14.7
↪A1	|	67.5	82.4	17.6
↪A2	|	82.1	96.2	14.7
↪A3	|	90.1	96.6	14.6
OCNLI	s_2	61.5	85.7	24.7
CAugNLI	s_1	41.9	49.5	8.0
|	s_2	39.0	46.0	5.1
$SNLI_{debiased}$	s_1	45.3	45.3	2.9
|	s_2	65.3	68.3	29.8
$MNLI_{debiased}$	s_1	34.0	35.6	-2.2
|	s_2	57.1	58.9	20.9

Table 3.3: For each NLI dataset and potential leakage-exhibiting single-sentence condition (Cond.) we report *test* accuracy in the single sentence condition (SSC), and three derived metrics from Sec. 3.1.1. SOTA test accuracy recovery ($\%R_S$), replication test recovery ($\%R_R$), and biased condition improvement over the chance majority guessing strategy (Δ_{maj}).

SSC-PSC Accuracy Recovered: Accuracy achieved by SSC model over PSC:

$$\%R_R = \frac{P(f_{SSC}(X_{test}) = Y_{test})}{P(f_{PSC}(X_{test}) = Y_{test})} \tag{3.5}$$

This metric captures how similarly the single sentence and normal condition models perform.

3.1.3 Dataset details

SICK Sentences Involving Compositional Knowledge [78] was produced by instructing annotators to label existing sourced pairs from 8K ImageFlickr data set [98] and SemEval 2012 STS MSR-Video Description data set [99]. The dataset is in

English. Each sentence pair was annotated for relatedness and entailment by means of crowdsourcing techniques.

SNLI The Stanford NLI dataset [79] contains content from the Flickr 30k Corpus [98], VisualGenome corpus [100] and [87]. The corpus is in English. The dataset is collected through human-written English sentence pairs.

MNLI The Multi-genre NLI Corpus [80] was modeled on the SNLI corpus [79] but it differs in the range of genres of spoken and written English text supporting cross-genre evaluation.

XNLI The Cross-Lingual NLI Corpus [81] consists of manually-translated dev and test samples from MNLI in 14 languages: French, Spanish, German, Greek, Belgian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu. It is interesting for analysis because on a high level the semantics of the data follow MNLI. The corpus is made to evaluate the inference in any language when only English data is presented at training time.

FEVER NLI-style FEVER [82] is an NLI reformulation of the FEVER claim verification dataset [101]. The original dataset was collected by eliciting annotators to write fact sentences that are *supported*, *refuted*, or *unverifiable* relative source passages drawn from Wikipedia. This is converted into an NLI task by treating the elicited sentences as premises and the source passages as NLI pairs with relations *entail*, *contradict*, or *neutral* respectively. This dataset is unique in that the premises were elicited from seed hypotheses, meaning it has a balanced hyp. distribution but potentially biased prem. distribution.

ANLI The adversarial NLI corpus [83] was collected through crowdworkers and the purpose of this dataset creation is to make the state-of-art results fail in this dataset. The sentences are selected from the Wekipedia and manually curated HotpotQA training set [102]. The language is in English. It contains three partitions of increasing complexity and size, which we refer to hereafter as A1, A2, and A3. Detailed data statistics are in Table 3.1.

OCNLI The Original Chinese NLI corpus was collected following MNLI-procedures but with strategies intended to produce challenging inference pairs [84]. No translation was employed in producing this data; the source premise sentences and elicited hypotheses are original.

CAugNLI [85] produced counterfactually augmented datasets for NLI and sentiment analysis using human annotators, instructing them to make minimal changes to the sentences beyond those necessary to change the label. It extends the work of [103] and [79]. They find that a BiLSTM classifier achieves negligible performance over chance when trained on hypothesis only. However, since their dataset includes elicited modified sentences in both the premise and hypothesis populations, there are opportunities for bias on both.

CAugNLI was produced by having human annotators minimally modify either the premise or hypothesis of 2,500 samples drawn randomly from SNLI so as to produce new samples with similar structure and word distributions but different meanings. These modifications are intended to reduce spurious correlations, in particular by roughly equalizing the distribution of relation labels with respect to word-level and semantic-level patterns in the elicited hypothesis sentences.

$\text{SNLI}_{\text{debiased}}$ and MNLI_{db} are augmentations of the SNLI and MNLI train sets produced by training GPT-2 [104] generators on them, and then generating samples which they check for accuracy using a pretrained RoBERTa NLI classifier, and then reject if they exhibit spurious correlations including samplewise hypothesis-only model classifiability [86]. To do this they first train static hypothesis-only SNLI and MNLI models, and reject all generated samples that can be successfully classified hypothesis-only by them. However, beyond this test under a static hypothesis-only distribution they do not attempt to assess if their generator models introduce *new leakage features* in the sentence distributions as a result of their accuracy filtering process. To test this we create test splits on the data (as they provide train sets only) which contain no sentence overlap with the train sets through random sampling.

3.1.4 Biased Model Results

Table 3.3 shows the extent of the single-sentence relation leakage problem across the 17 SSC tests on the 14 splits for the 10 NLI datasets. These results clearly show that *each dataset exhibits significant single-sentence relation leakage* for at least one condition. The comparison columns replication test recovery ($\%R_R$), and SSC improvement over the chance majority guessing strategy (Δ_{maj}) are all computed using the single sentence condition accuracy and the standard NLI two-sentence condition SOTA and replication accuracy values in table 3.2.

3.1.5 Model Training

Our training technique is simple and applied consistently to all datasets. We fine-tune three different language-specific pretrained transformer checkpoints[2] from

[2] `roberta-large` [105], `xlm-roberta` [106], `bert-base-chinese` [107]

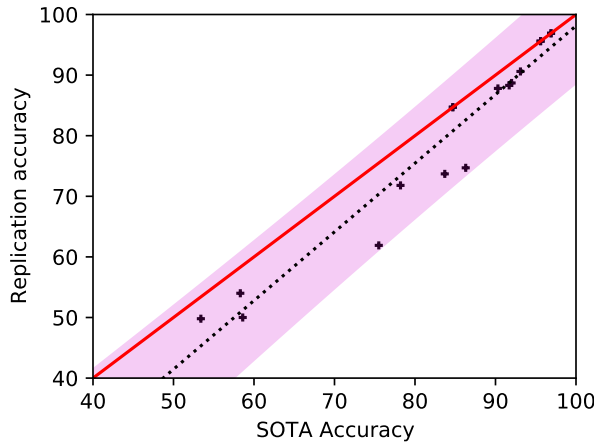


Figure 3.3: Test set accuracy for SOTA models vs our universal replication procedure models for each dataset in PSC, with a trendline (PCC=0.97) and $y = x$ line.

`HuggingFace` [108] using `Pytorch Lightning`. All models were trained on NVIDIA A-100 GPUs. All models are the HuggingFace `xForSequenceClassification` with `num_classes=3` and no other modifications. All models are trained using the Adam optimizer with cross entropy loss.

We find that this procedure produces broadly near-SOTA performance models, with a maximum relative accuracy difference of 8%, and a 92% Pearson's correlation coefficient (PCC) between SOTA and replication accuracy across the datasets (fig. 3.3). Our replications are a reasonable proxy to SOTA for comparative dataset analysis.

Replication Training Details

Hyperparameters For each dataset in the paired sentence condition (PSC), we select a batch size for maximum GPU utilization. We perform a grid search over lr in {5e–7, 1e–6, 5e–6, 1e–5, 5e–5}.

Single Dataset Fine-tuning We obtain separate fine-tuning checkpoints from the pretrained models for each dataset to enable clean analysis of one dataset at a time.

We do not accumulate fine-tuning passes across multiple datasets.

Single Sentence Condition Training

To train each dataset's corresponding **SSC** model(s), we use the same setup as the PSC model but follow [66]'s formulation of fine-tuning the chosen classification model on only the SSC sentence, *premise only* or *hypothesis only*.

FEVER exhibits bias in the premise distribution, and CAugNLI, $\text{SNLI}_{\text{debiased}}$, and $\text{MNLI}_{\text{debiased}}$ exhibit imbalance in both (table 3.1). For the datasets that have imbalanced distributions in both conditions, we separately train bias models for both hypothesis-only and premise-only.

3.2 Analyzing NLI Dataset Bias

In this section we introduce quantification techniques for more accurately characterizing the extent of these bias problems in the aforementioned NLI datasets, analyze how they interact with the observable bias itself, and develop tools for producing future NLI benchmarks that more closely resemble the ideal benchmark.

3.2.1 Sample-level Model Behavior

We are particularly interested in understanding the degree to which models trained in the PSC and SSCs "reason" similarly. For this section we use the notations $f(\boldsymbol{X}_{\text{test}})$, $\boldsymbol{Y}_{\text{test}}$ to denote the $(1 \times N)$ column vectors of model output decisions and labels for a test set of N samples, and a simple agreement function $Ag(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ as the ratio of elements that are identical across all $\boldsymbol{Y}_i$ to the vector size N. In other words,

$$Ag(f(\boldsymbol{X}_{\text{test}}), \boldsymbol{Y}_{\text{test}}) = P(Y_{\text{test}} = f(X_{\text{test}})) \tag{3.6}$$

SSC-PSC Agreement (NBA): The number of samples for which the SSC and PSC models agree over the total number of samples in the set:

$$\text{NBA} = \frac{Ag(f_{\text{SSC}}(\boldsymbol{X}_{\text{test}}), f_{\text{PSC}}(\boldsymbol{X}_{\text{test}}))}{|\boldsymbol{X}_{\text{test}}|} \tag{3.7}$$

SSC-PSC Recovery (NBR): The number of samples for which the SSC and PSC models agree, *and both classify correctly* over the total number of samples they agree on:

$$\text{NBR} = \frac{Ag(f_{\text{SSC}}(\boldsymbol{X}_{\text{test}}), f_{\text{PSC}}(\boldsymbol{X}_{\text{test}}), \boldsymbol{Y}_{\text{test}})}{Ag(f_{\text{SSC}}(\boldsymbol{X}_{\text{test}}), f_{\text{PSC}}(\boldsymbol{X}_{\text{test}}))} \tag{3.8}$$

Token Relevance Agreement (TRA): Do SSC and PSC models reason alike? For a sentence X with length n, we compute the gradient of the classification output posterior with respect to each token embedding $emb(w_j)$. We take the 2-norm of the each gradient vector and normalize it over the entire sequence to produce a normalized local explanation vector $m(f(X))$ [109]:

$$m(f, X) = \Big[\frac{|\nabla_{\text{emb}(w_j)}(f(\text{emb}(w_j)))|_2}{\sum_{i=1}^{n}(|\nabla_{\text{emb}(w_i)}(f(\text{emb}(w_i)))|_2)}\Big]_{j=1}^{n} \tag{3.9}$$

To compare "reasoning" similarity between the two models, we compute the samplewise input token relevance agreement can be computed using cosine similarity:

$$\text{TRA}(X_i) = \frac{m(f_{\text{PSC}}, X_i) \cdot m(f_{\text{SSC}}, X_i)}{||m(f_{\text{PSC}}, X_i)||\,||m(f_{\text{SSC}}, X_i)||} \tag{3.10}$$

As SSC and PSC inputs have different lengths, we pad the SSC importance vector for $m(f_{\text{SSC}}(X_i))$ with zeros to equalize the length of the local explanation map vectors. The dataset-level token relevance agreement is the average of samplewise TRA.

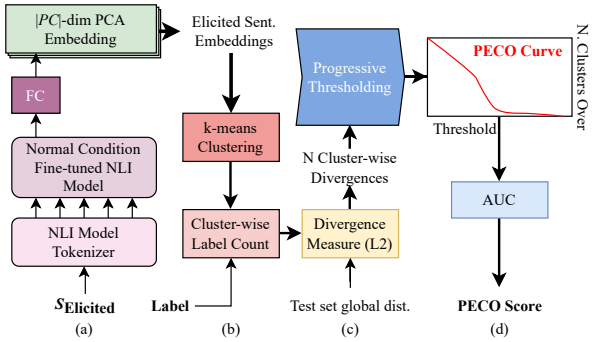


Figure 3.4: An overview of the approach to computing the PECO score from a collection of elicited population sentences s_{elicited} and their corresponding Labels r. When a fixed threshold is chosen, the Hypothesis embeddings can be dimensionality-reduced using T-SNE to produce plots like chapter 3.

3.2.2 Cluster-based Bias Evaluation

We are interested in investigating how the biased distributions of the elicited sentences in NLI datasets are captured in the learned representation spaces of models trained on them. In particular, we are interested in answering this question: **is elicited sentence label leakage captured semantically in regions of latent space**?

To answer this we produce dimensionality-reduced *elicited sentence embeddings* for the test set, using the PSC replication models, then fit a high-k k-means clustering to this collection of embeddings. This will allow us to analyze how the local distribution of labels varies over the elicited sentence embedding space. By comparing the KL-divergence of the label distribution within each cluster and the global label distribution, we can compute the **Progressive Evaluation of Cluster Outliers** (PECO) score (fig. 3.4).

Elicited Sentence Embeddings: To embed the elicited sentences *as they're learned by a model in the PSC*, we feed the elicited sentences s_e through the PSC replication fine-tuned NLI model encoder. We extract the latent codes produced at the output very last fully connected layer of the model before the linear classifier to collect latent codes for every s_e in the test set. We then embed these codes into their 30 principal

components to produce the embeddings (fig. 3.4 (a)).

Clustering: We fit a high-k (in this case, $k = 50$) k-means clustering over the distribution of elicited sentence embeddings to provide a set of local bins for analysis. For each cluster, we count the relation labels its samples contain, to produce a set of 50 cluster-label distributions (fig. 3.4 (b)).

Computing Cluster Divergences: For each cluster label distribution $p_i = P(Y|\text{cluster} = i)$, we assess the L2 divergence between it and the global label distribution p_G to produce divergence scores s_i:

$$s_i = \frac{1}{3}\sum_{j=1}^{3}(P(Y = j) - P(Y = j|\text{cluster} = i))^2 \tag{3.11}$$

This step is depicted in fig. 3.4 (c).

Progressive Evaluation: Finally, we compute the PECO score for this collection of cluster divergences as the area under the curve produced by counting the number clusters with divergence s_i over some threshold t for the range of s_i.

$$\text{PECO} = 100 \int_{\min(s)}^{\max(s)} \frac{\text{count}_i(s_i > t)}{k} dt \tag{3.12}$$

Generality of PECO: These same techniques could be applied to a wide variety of potential leakage features on the input to analyze a wide variety of correlation types. For example, input sentence words could be shuffled to test for word order invariance, or word classes could be specifically masked to test for spurious vocabulary correlations.

Dataset	Cond.	NBA	NBR	TRA	PECO
SICK	s_2	48.8	49.7	63.6	13.9
SNLI	s_2	70.0	71.8	63.9	14
MNLI-b	s_2	32.5	33.4	57.6	6.5
MNLI-u	|	47.5	48.6	59.3	7.5
XNLI	s_2	52.3	54.0	64.7	9
FEVER	s_1	39.4	38.0	50.7	14.3
ANLI	s_2	37.6	53.4	28.0	5.5
↪A1	|	53.8	57.9	54.0	15.3
↪A2	|	53.9	60.3	51.2	11.1
↪A3	|	52.1	58.5	55.7	15.9
OCNLI	s_2	69.0	74.1	78.5	17.9
CAugNLI	s_1	32.9	32.7	39.5	5.5
|	s_2	43.4	42.5	59.7	5.4
$\text{SNLI}_{\text{debiased}}$	s_1	32.0	31.8	41.5	8.2
|	s_2	60.9	61.4	58.4	8.3
$\text{MNLI}_{\text{debiased}}$	s_1	34.7	34.0	46.7	6.6
|	s_2	42.2	42.0	58.0	6.8

Table 3.4: Metrics comparing the behavior of our replication and single-sentence condition models on each dataset using the metrics introduced in Secs. 3.2.1, 3.2.2: Normal-Bias Agreement (NBA) and Recovery (NBR), Token Relevance Agreement (TRA) and the Progressive Evaluation of Cluster Outliers (PECO) score.

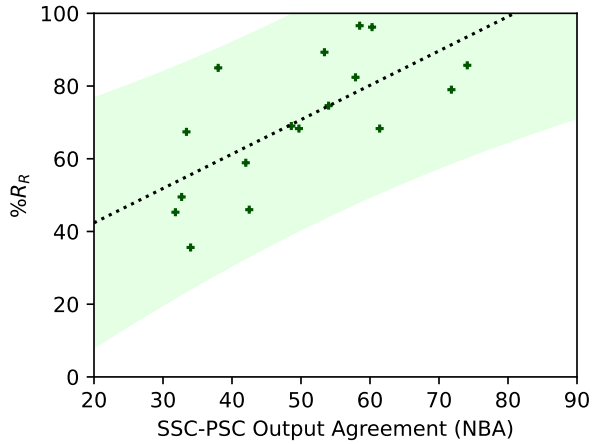


Figure 3.5: Model-wise output agreement vs Bias accuracy recovery (%R_R). As the replication PSC model and the SSC model agree more often for a given dataset, their performances in the two conditions converge.

		k=10						*k=25*					
		No PCA		$\|PC\| = 50$		$\|PC\| = 100$		No PCA		$\|PC\| = 50$		$\|PC\| = 100$	
Dataset	Cond.	L2	KLD	L2	KLD	L2	KLD	L2	KLD	L2	KLD	L2	KLD
SICK	s_2	6.0	0.085	6.0	0.075	6.0	0.080	6.8	0.084	6.0	0.080	6.0	0.078
SNLI	s_2	8.5	0.155	8.5	0.150	8.5	0.150	8.4	0.138	8.8	0.140	8.4	0.136
MNLI-b	s_2	5.0	0.055	5.0	0.055	5.0	0.055	5.2	0.064	5.4	0.064	5.2	0.060
MNLI-u	|	5.5	0.065	5.5	0.065	5.5	0.065	5.8	0.072	5.6	0.072	5.8	0.070
XNLI	s_2	5.0	0.065	5.0	0.065	5.0	0.065	5.2	0.068	5.2	0.068	5.2	0.068
FEVER	s_1	11.0	0.160	11.0	0.160	11.0	0.160	13.2	0.154	10.6	0.150	10.2	0.144
ANLI	s_2	15.0	0.095	14.5	0.070	14.5	0.070	31.0	0.176	8.8	0.058	8.8	0.058
↪A1	|	5.5	0.085	5.5	0.075	5.5	0.090	17.2	0.124	28.4	0.136	21.8	0.118
↪A2	|	5.5	0.065	5.5	0.070	5.5	0.070	9.4	0.086	9.2	0.080	9.4	0.084
↪A3	|	5.5	0.070	5.5	0.070	5.5	0.065	5.8	0.082	5.6	0.084	5.8	0.080
OCNLI	s_2	10.5	0.165	10.5	0.170	10.5	0.170	14.0	0.178	14.2	0.170	14.0	0.172
CAugNLI	s_1	5.0	0.050	5.0	0.055	5.0	0.050	5.0	0.058	5.2	0.060	8.6	0.080
|	s_2	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050
$\text{SNLI}_{\text{debiased}}$	s_1	6.5	0.100	6.5	0.100	6.5	0.100	6.2	0.090	6.4	0.094	6.6	0.096
|	s_2	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050
$\text{MNLI}_{\text{debiased}}$	s_1	5.5	0.065	5.5	0.065	5.0	0.065	5.8	0.072	5.8	0.078	5.8	0.074
|	s_2	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050	5.0	0.050

		k=50						*k=100*					
		No PCA		$\|PC\| = 50$		$\|PC\| = 100$		No PCA		$\|PC\| = 50$		$\|PC\| = 100$	
Dataset	Cond.	L2	KLD	L2	KLD	L2	KLD	L2	KLD	L2	KLD	L2	KLD
SICK	s_2	6.2	0.081	7.4	0.082	7.9	0.081	13.5	0.104	13.2	0.098	12.4	0.102
SNLI	s_2	11.3	0.141	9.8	0.137	9.8	0.141	12.5	0.148	10.0	0.143	10.4	0.136
MNLI-b	s_2	5.4	0.066	5.4	0.066	5.3	0.064	5.4	0.069	5.4	0.068	6.3	0.069
MNLI-u	|	7.4	0.073	7.4	0.076	5.9	0.078	8.5	0.087	9.3	0.087	8.6	0.086
XNLI	s_2	5.5	0.074	5.6	0.074	5.6	0.073	5.7	0.079	5.7	0.080	5.8	0.079
FEVER	s_1	14.6	0.154	14.7	0.152	12.7	0.149	16.7	0.159	17.2	0.157	17.8	0.160
ANLI	s_2	25.2	0.132	8.9	0.063	7.0	0.058	26.9	0.162	9.2	0.071	9.0	0.072
↪A1	|	37.8	0.180	31.0	0.168	36.1	0.179	68.2	0.354	56.8	0.301	67.9	0.341
↪A2	|	17.3	0.114	19.6	0.112	15.1	0.103	37.1	0.171	31.8	0.188	34.9	0.179
↪A3	|	13.4	0.108	13.8	0.106	9.8	0.094	26.8	0.147	30.5	0.162	27.8	0.152
OCNLI	s_2	18.8	0.190	14.9	0.180	14.5	0.176	25.2	0.193	26.5	0.207	21.4	0.186
CAugNLI	s_1	5.0	0.051	5.0	0.050	5.0	0.051	5.9	0.059	5.9	0.056	5.1	0.056
|	s_2	13.8	0.073	6.9	0.068	12.1	0.101	19.7	0.126	19.0	0.135	13.2	0.091
$\text{SNLI}_{\text{debiased}}$	s_1	5.0	0.050	5.0	0.051	5.0	0.050	5.0	0.051	5.0	0.051	5.0	0.051
|	s_2	5.9	0.080	6.3	0.088	6.1	0.085	6.0	0.085	6.9	0.090	6.2	0.086
$\text{MNLI}_{\text{debiased}}$	s_1	5.0	0.052	5.0	0.052	5.0	0.052	5.1	0.054	5.1	0.054	5.1	0.056
|	s_2	5.6	0.070	5.6	0.068	5.5	0.069	6.3	0.071	5.3	0.065	5.6	0.070

Table 3.5: PECO scores for various levels of principal component reduction ($|PC|$), various numbers of k-means clusters, using L2 and KLD distance. These numbers are collected from a separate reproduction to table 3.3, which uses different k for different datasets (reflecting their different sizes), and $|PC| = 30$. This table clearly illustrates that e.g., L2 PECO are pretty consistent across different numbers of principal components, for a given number of clusters k. We find that calibrating k based on the number of test-set samples for a given dataset is valuable for producing good characterization of the degree of bias. To improve legibility, we kept PECO_{KLD} in the $[0, 1]$ range.

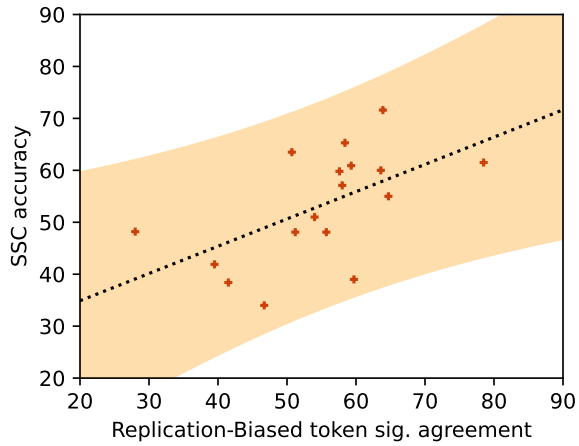


Figure 3.6: Token releveance agreement (TRA) vs SSC accuracy. When the PSC model and SSC models "reason" more similarly for a dataset, the relation leakage bias exhibited in that dataset tends to be higher.

3.3 Results

As discussed above, ideal NLI benchmarks are neither *saturated* nor *biased*. Unfortunately, as table 3.3 demonstrates, none of the NLI datasets tested thus far satisfy this condition. This is more clearly illustrated in fig. 3.2. Two questions remain: *to what extent do current models cheat* and *how can we make less biased, less saturated datasets?* table 3.4 contains experimental results intended to answer these two questions. The "agreement metrics" as introduced in Sec. 3.2.1, SSC-PSC Agreement, SSC-PSC Recovery, and Token Relevance Agreement are provided in table 3.4.

3.3.1 Result-Metric Correlations

We find that SSC-PSC model output agreement (NBA) and recovery rate $\%R_R$ are **correlated with a PCC of 0.69** (fig. 3.5). Datasets where the SSC and PSC models predict more similarly have closer the SSC and PSC performance results are for said datasets. While this result is surprising, ANLI R3 is an interesting outlier (§3.4).

We find that TRA and SSC accuracy are also **positively correlated with a PCC**

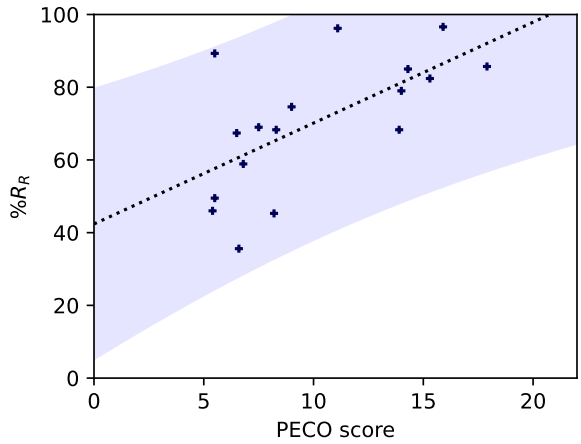


Figure 3.7: Test-set PECO score vs Bias accuracy recovery (%R_R). This result suggests that interventions that produce lower PECO score datasets tend to yield datasets that exhibit less relation leakage in the SSC.

of 0.57 (fig. 3.6). This result demonstrates that for a single dataset, similar reasoning patterns for the single sentence condition and standard sentence pair condition is strongly correlated to single-sentence relation leakage. In other words, **standard condition NLI models trained on biased (high leakage) datasets tend to cheat**. Thus, models indeed rely on annotation artifacts in NLI datasets to achieve high accuracy, and demonstrates that this continues to be a problem in newer NLI datasets, in spite of mitigation attempts. How can we use this knowledge to build better benchmarks?

fig. 3.7 depicts the relationship between PECO score and bias recovery (%R_R). We find the two are **positively correlated with a PCC of 0.64**. This result is fairly intuitive: the more uneven the distribution of labels is in the single-sentence latent spaces (and thus, the higher the area under the PECO curve), the more SSC performance approaches the standard PSC condition performance for a given NLI dataset. This suggests PECO-reducing interventions may be able to target debiasing efforts.

3.4 Discussion

Some examples are only correctly classified in the single-sentence condition. A common assumption to discussions of cheating features in machine learning is that they play a role in inflated classification accuracy when present. However, ANLI R3 provides an interesting counterexample. For this dataset, the hypothesis-only model achieves a SSC accuracy of 48.1%, and the PSC model achieves 49.8% (a $\%R_R$ of 90%), and SOTA achieves 53%. Despite this score similarity, the samples which the two conditions are able to actually classify correctly vary surprisingly. With an NBR of 58.5%, only $\approx$ 27% of test samples are correctly classified by both the single and two sentence condition models. This means that around **21% of samples in A3 test are *only* correctly classified by the single-sentence model**.

Perhaps unsurprisingly, ANLI exhibits the lowest TRA out of all datasets tested, indicating that it is somewhat of an outlier in having the SSC and PSC condition models reason differently on it.

XNLI demonstrates the cross-lingual and semantic nature of single-sentence leakage. While previous work has focused on finding words, phrases, patterns, and heuristics in the surface form of the data, our study of XNLI provides an interesting opportunity to investigate the potential for the influence of underlying semantics as a leakage feature. XNLI consists exclusively of 14 language test and val sets, manually translated from MNLI examples. Our XNLI PSC and SSC models are thus trained on MNLI alone, using the multilingual `xlm-roberta` checkpoint.

This produces a natural experiment wherein surface form biases present in the training data are completely eradicated in the test set (as only the 14 non-English languages table 3.1 are present), while the underlying meanings encoded by those

words remain. In table 3.3 we indeed find that XNLI and MNLI exhibit very similar result comparisons. The models on both datasets have a $\Delta_{\text{max}} \approx 20\%$ and $\%R_R \approx 65$. These leakage feature results, being robust to manual translation into 14 different languages, seem to indicate that there is a strong fundamental semantic component to the human biases driving the elicited sentence relation leakage.

Relation leakage remains Unsolved. Elicited sentence relation leakage is a problem for all evaluated NLI datasets, including the new ones intended to fix it. Recent datasets such as XNLI, FEVER, and OCNLI, exhibit high absolute SSC performance over majority ($\Delta_{\text{maj}} > 20$).

Although ANLI and CAugNLI are improvements over the others in terms of Δ_{maj}, with CAugNLI shining particularly in this regard, none eliminate the relation leakage problem entirely, as even CAugNLI still has $\Delta_{\text{maj}} = 8.0$, an 8% performance over chance in the single sentence condition.

$\text{SNLI}_{\text{debiased}}$ and $\text{MNLI}_{\text{debiased}}$, despite their intended purpose, still contain significant amounts of SS label leakage (29.8% and 20.9% over chance). This might be because while their production [86] does eliminate bias *originally present in SNLI and MNLI*, it fails to prevent the introduction of *new bias* in the data generation pipeline.

Cluster approaches are promising for future debiasing efforts. chapter 3 shows how the PECO-derived cluster-bias T-SNE plots can be used directly to visualize, analyze, and "debug" biased datasets. In the plot, SNLI clearly has considerably more high-biased clusters taking up a considerable portion of the latent space as compared to CAugNLI, for the bias threshold of 0.2.

An intervention could be performed on identified bias regions in the distribution by having human annotators create new premise sentences from the given hypotheses,

thereby forcing the PECO-based bias metrics to reduce. This idea is further backed up by the PCC of 35.8 that we find between PECO and $\%R_R$, suggesting that producing datasets of lower PECO score will naturally lead to lower recovered performance in the SSC, and thus less elicited sentence relation leakage.

3.4.1 Related Work

Understanding Bias in NLI [110] demonstrated that counterfactual augmentation alone cannot debias NLI. Multi-task learning can improve model robustness to fitting spurious features [111], but because the underlying **benchmarks are biased**, progress on the desired reasoning capability is questionable [66]. [112] strengthen the finding that annotator bias is a key driver of this poor generalization performance, showing that NLI models can struggle to even generalize across disjoint sets of annotators on the same task.

Simple word- and n-gram level approaches have proven surprisingly capable in a-priori characterizations of dataset difficulty [113] and producing difficult test sets [114] in diverse language domains such as SLU. [115] show how such purely frequentist approaches can identify word-level spurious correlations with respect to label class which drive in-part the shortcut features for classes of "competency problems" such as NLI.

Mitigating Bias in NLI [116] demonstrate an approach to train NLI models robustly against some of these biases, using [87]'s hard test set. [67] utilize simple heuristics like lexical overlap to produce the synthetic debiased HANS NLI dataset to test generalization. This dataset has been used to evaluate techniques including predicate-argument- [117] and syntactic transformation-based [118] augmentations. [119] leverage a bag-of-words approach to debias datasets along lexical features. However, these approaches

have yet to improve generalization in comprehensive replication studies has thus far [120]. Meanwhile [121] propose a fully unsupervised data collection pipeline for NLI, in order to sidestep the problem of human biases entirely.

Approaches like [85] and [86] are very promising for producing data that reduces bias on a samplewise, but not populationwise level. Using our semantic, model-driven local bias finding strategies, future interventions can lead to the large scale production of debiased NLI datasets and a new generation of higher quality benchmarks for language understanding. [122] perform such a targeted augmentation approach using the *dataset cartography* sample characterization scheme from [123] to produce WANLI, an NLI dataset that allows for improved performance on the aforementioned challenging test sets.

3.4.2 PECO vs Dataset Cartography

To determine if PECO-driven dataset augmentation is redundant given the recent release of WANLI, we seek to determine if usable samplewise information for targeting interventions (e.g., presence in a "bias" cluster) is captured during PECO analysis and is redundant to the relevant samplewise characterization produced in dataset cartography.

To do this, we collect the samplewise *confidence* feature [123] during training of the PSC model for each validation set sample in SNLI. We then assign each validation sample to its corresponding PECO cluster (out of the 50) and produce two histograms of the confidence feature, one for "biased" clusters ($\text{PECO}_{L2} > 0.25$) and one for the other clusters. fig. 3.8 shows the results of this experiment. We find that out of the 10k validation set examples, roughly 2 are assigned to an "unbiased" cluster for every 1 assigned to "biased," roughly evenly across all confidence bins.

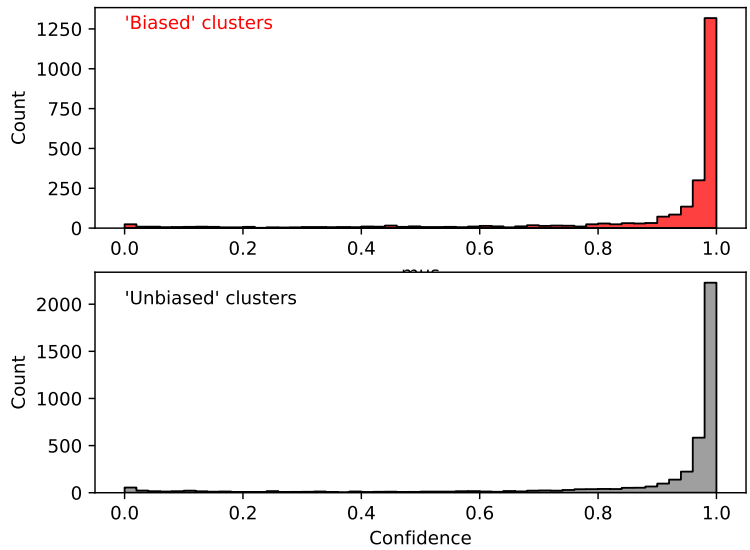


Figure 3.8: Histograms for the samplewise *confidence* feature from dataset cartography in SNLI, for high- (biased) and low-PECO$_{L2}$ clusters.

This result suggests that PECO clusterwise "biasedness" is orthogonal to the samplewise ease of learnability captured by the dataset cartography confidence feature. In other words, we find that **some samples are easy to learn (high confidence) because they are simple, while other samples are easy to learn because the model is using cheating features.** PECO-like analyses will be instrumental in guiding future efforts to eliminate shortcut features in natural language datasets.

3.5 Conclusion

In the half decade since this NLI cheating feature phenomenon was discovered [66], single sentence relation leakage bias continued to be difficult. Efforts to debias NLI have led to datasets that merely exhibit different kinds of bias than those shown before, or less saturated benchmarks that continue to exhibit cheating features. Future work must prioritize reducing observable bias directly using a model-driven approach.

Part II

Measurement challenges in multimodal systems

Chapter 4

Text-to-image faithfulness

With advances in the quality of text-to-image (T2I) models has come interest in benchmarking their *prompt faithfulness*—the semantic coherence of generated images to the prompts they were conditioned on. A variety of T2I faithfulness metrics have been proposed, leveraging advances in cross-modal embeddings and vision-language models. However, these metrics are not rigorously compared and benchmarked, instead presented with correlation to human Likert scores over a set of easy-to-discriminate images against seemingly weak baselines.

In this chapter, I present a treatment of this issue based on principles put forth in **Chapter 2**. Stemming from a precise characterization of *what faithfulness means* in generated images, we create a new resource—`T2IScoreScore`—to model it and deliver surprising findings about the relative efficacy of T2I faithfulness metrics.

T2I models are improving at a breakneck pace in terms of the coherence of generated images to their conditioning prompts [124, 125, 126, 127]. Despite this, persistent challenges in achieving image-prompt faithfulness [128, 129] remain—particularly in freely available models that don't sit behind proprietary APIs. Indeed, many techniques to improve T2I models have been proposed of late, aiming to reduce

hallucination [130, 131], duplication [132], composition errors [131, 133], and missing objects [134, 135]. However, there is no consensus on how to best compare these many models and methods, so it is hard to objectively track T2I progress [136, 137].

Recent work has proposed a litany of automated *image-prompt coherence metrics* which rate the *faithfulness* of generated images: the degree to which a they satisfy the implicit requirements set forth in the generating prompt [138, 139, 140, 141]. These proposed metrics vary considerably in design; as rating how well an image matches to its prompt is a nontrivial multimodal challenge [137, 142, 143].

This variety itself presents a *meta-evaluation problem*: there is no **consensus on how these faithfulness metrics ought to be compared**, and consequently each new metric is validated on its own ad-hoc test set against prior baselines. Typically these *self-evaluations* consist of a set of prompt-image pairs with accompanying human annotations (usually simple Likert scores [142, 139]), and metrics are judged on their correlation to these human judgements [143]. Such self-evaluation is not ideal; authors may unwittingly tilt the scales by using evaluation examples which cater to the particular strengths of their proposed method, and variance of metric performance between different evaluation sets (containing different images and prompt semantics [69, 113]) is high [144]. Additionally, relying on correlation to human judgements of small sets of images across different prompts is highly subjective [145, 146] and prone to including judgements of quality and style that are orthogonal to prompt coherence. We need a consistent and objective meta-evaluation.

To this end we propose `T2IScoreScore` (`TS2`), a *benchmark and set of meta-metrics* for evaluating T2I faithfulness metrics. While it contains a similar number of images to previously proposed coherence metric evaluation sets, it contains fewer prompts. This high image-to-prompt ratio allows us to organize the images along *semantic error graphs*, or SEGs (fig. 4.1), where each edge corresponds to a specific error with respect

to the prompt that a child image set possesses but its parent images do not. These semantic error graphs permit objective scoring of a metric by answering:

1. Can a metric correctly **order** increasingly wrong images against their generating prompt?

2. Can a metric reliably **separate** sets of images that differ by a specific semantic error?

3. Does the metric **confidently** separate the image sets within its dynamic range?

We adapt existing statistical tests [147, 148] to the SEG setting to answer these questions for a broad set of T2I faithfulness metrics. To our surprise, despite their inferior performance in correlating to human preferences against complicated vision-language model (VLM)-based metrics [139, 129, 141, 140], simple embedding-correlation methods like CLIPScore [138] are actually quite performant on our meta-metrics, and **Pareto-optimal** with respect to compute cost (§5.3). In this chapter, I will:

- Formalize the task of objectively assessing T2I prompt coherence metrics by their ability to correctly order and separate image populations within semantic error graphs (SEGs). (§4.2)

- Present `T2IScoreScore` (`TS2`), our evaluation for this task: a carefully-curated benchmark dataset of SEGs each containing between 4 and 76 images, permitting 93,000 total pairwise image comparisons and meta-metrics for ordering and separation in SEGs. (§4.2, §4.3)

- Evaluate a broad and representative set of T2I faithfulness benchmarks using `TS2`, demonstrate that it identifies novel failure cases, and motivate future work on improved metrics. (§5.2 §5.3, §4.6)

4.1 Background

Most evaluations in the T2I space test the quality of generating models based on *a fixed faithfulness metric's scores over a fixed benchmark set of prompts.* Often these reference prompt sets are designed for testing a single specific capability. DrawBench [127], T2I-CompBench [149], and ABC-6k [131] focus on attributes like compositionality, cardinality, and spatial relations, in strictly text-guided image generation, while ImagenHub [137] tests them in a broader set of settings like image editing and subject-driven synthesis. Other evaluation dimensions such as multilinguality [150, 151] and stereotype bias [152] have also been explored. These prompts are usually sourced from some combination of existing natural image captioning resources [153, 154, 155] and sets of in-the-wild conditioning prompts produced by real users [156, 157]. These benchmarks assess image quality either by direct human analysis or automated metrics [137] including those we analyze in this work. Often the goal of these benchmarks is to analyze how well a model generates images that comport with human preferences [158, 137], and directly elicit opinions from users through a web interface for this purpose.

To meta-evaluate faithfulness metrics *a benchmark containing a fixed set of images and prompts is required*. However, all existing benchmarks for this purpose suffer from two key limitations:

1. Evaluating on noisy *human preference* rather than *explicitly labeled objective differences*

2. *Low image-to-prompt ratio* limiting evaluation of discriminatory power over similar images

Captioning benchmarks [153, 154, 155] are poor candidates for faithfulness metric

Dataset	# Images Total	# Images T2I-Gen.	Img. per Prompt Avg	Img. per Prompt Per equiv. pref	# Img Comparisons Total	# Img Comparisons T2I Errors	Ad-hoc
Benchmarks for captioning models.							
Flickr8k [153]	8k	0	0.2	0.2	8k	0	–
Flickr30k [154]	31k	0	0.2	0.2	31k	0	–
MSCOCO Captions [155]	330k	0	0.22	0.22	330k	0	–
Benchmarks for image retrieval/matching models. (Could be used for T2I metric evaluation)							
SeeTRUE [159]	31k	0	1	1	31k	0	✓
Pick-a-Pic [158]	500k	500k	2[1]	1	7M	0	✓
Benchmarks for T2I faithfulness metrics.							
TIFA v1.0 [139]	800	800	5	1	4k	0	✓
DSG-1k [139]	1k	1k	1	1	1k	0	✓
`T2IScoreScore`	*2.8k*	*2690*	**17**	**3.4**	*93k*	**3.1k**	**–**

Table 4.1: Comparison of benchmark datasets that can be used to evaluate T2I faithfulness. *Per equiv pref* means the average number of images for each prompt that are assigned the same preference or correctness score. **Bold** numbers are best overall, *italic* are best of the T2I metric benchmarks.

evaluation as single images are paired with multiple prompt candidates rather than vice versa. Image matching and entailment benchmarks such as SeeTRUE [159] and Pick-a-Pic [158] are also limited by a low ratio.

The few extant deliberately-designed faithfulness evaluation sets are limited by both factors. TIFA v1.0 [139] and DSG-1k [129] were proposed *ad-hoc* to demonstrate the utility of their accompanying metrics by relating the scores assigned by the metric to human preferences. These are done over small sets of images (800 & 1000 respectively) with slightly more images-per-prompt (5 & 1).

The two limitations of these prior evals are linked. A reliance on human preference correlations is a natural consequence of having few and poorly organized images to compare to each prompt. A lack of meta-metrics designed for evaluating structured aspects other than human preference correlation means limited utility in collecting larger, structured sets of images for each prompt. By providing both *meta-metrics* ***and*** *a structured eval set* `T2IScoreScore` overcomes these limitations (table 4.1).

[1] Although this benchmark has 14 pairs of images per prompt, each pair is separately annotated. With no way to compare between pairs, the effective number of images per prompt is 2, with many repeated prompts.

4.2 `T2IScoreScore` meta-metrics

We introduce three measures of ordering and separation by a given metric within semantic error graphs (SEGs). We define SEG S as prompt P and a directed acyclic graph of nodes n_i containing one or more images I_j sharing the same errors wrt. the prompt. We label each node by its error count and type (eg, [`0`, `1a`, `1b`] has 1 node with 0 errors, and 2 nodes with 1 error each of different type).

A good prompt coherence metric will correctly rank images along each walk of increasing error counts within a SEG, and separate the scores assigned to images in successive nodes. Our metrics assess this by evaluating each walk separately. For ease of notation, we refer to each SEG as a set of walks $W \in S$ over nodes of increasing error count (eg, (0, 1a, 2a), (0, 1a, 2b), etc), where each walk is the in-order set of all (image, prompt, num. error) triples $(I, P, N) \in W$. For example, the first walk in fig. 4.1 is [(`0-0.jpg`, P, 0), (`0-1.jpg`, P, 0), (`1-0.jpg`, P, 1),(`2-0.jpg`, P, 2), ...].

We introduce measures of metric m for SEG S: $\texttt{rank}_m(S)$, $\texttt{sep}_m(S)$ & $\texttt{delta}_m(S)$, assessed over every walk $W \in S$ in all SEGs in the `TS2` dataset to score a metric. They're defined as:

4.2.1 Ordering score over walks: $\texttt{rank}_m$

We use Spearman's rank correlation coefficient ρ [147] between image-level error count and metric-assigned score over every walk on a SEG to assess how a metric's faithfulness score aligns to our objective structural error counts. Spearman's ρ is the PCC of the rank order of variables X, Y:

$$\rho(X,Y) = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}; \quad R(X) = \Big\{ \sum_{x_i \in X} \mathbb{1}(x_i < x) \;\Big|\; x_i \in X \Big\} \tag{4.1}$$

Thus, in our case the SEG-level rank order score $\texttt{rank}_m(S)$ for scoring model m is defined as:

$$\texttt{rank}_m(S) = \frac{1}{|S|} \sum_{W \in S} r_s(\{m(I,P)|(I,P,N) \in W\}, \{N|(I,P,N) \in W\}) \tag{4.2}$$

One limitation of Spearman's ρ for characterizing scores is that it is undefined if one set $R(U)$ exclusively contains identical elements, as $\sigma_{R(U)} = 0$. For tractability in these scenarios we define $\rho(\cdot, R(U)) := 0$. If a metric assigns identical scores to all examples across different error levels, it presents no discernible relationship between error severity and score for that image set.

4.2.2 Statistical separation of error populations score: $\texttt{sep}_m$

We assess the two-sample Kolmogorov–Smirnov statistic [148] pairwise between the populations of metric m's scores assigned to each sample between two error nodes n_i and n_j as populations. The Kolmogorov–Smirnov statistic is a non-parametric measure of the separation between two distributions [160, 161], defined as the maximum vertical difference between their empirical cumulative distribution functions $F_X(s)$:

$$D_{KS}(X,Y) = \sup_{x \in R_m} |F_X(x) - F_Y(x)| \tag{4.3}$$

Where $F_X(x)$ is proportion of samples in population X for which the metric-assigned score $m(i) \leq x$. We compute D_{KS} for every pair of adjacent error nodes in each tree walk W[2], and report the average over all of these as the SEG separation score $\texttt{sep}_m$(S):

[2]We use individual nodes $n_i \in S$ containing pairs of prompt, image $(P, I) \in n_i$ to make this equation easier to read.

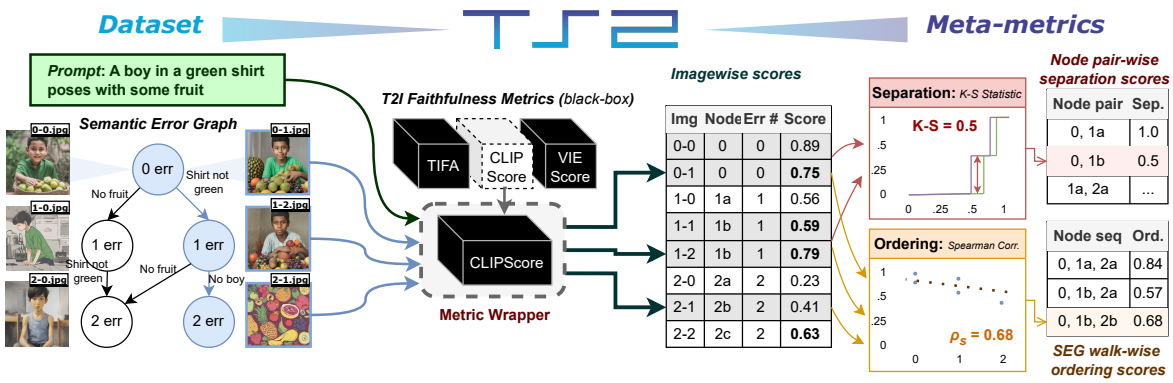


Figure 4.1: Overview of `T2IScoreScore`. T2I evaluation metrics are scored based on their ability to correctly organize images in a *semantic error graph* (SEG) relative to their generating prompt, checking ordering (Spearman's ρ) and separation of nodes (Kolmogorov–Smirnov statistic).

$$\texttt{sep}_m(S) = \frac{1}{|S|} \sum_{n_i \in S} D_{KS}(\{m(P,I)|(P,I) \in n_i\}, \{m(P,I)|(P,I) \in n_{i+1}\}) \tag{4.4}$$

4.2.3 Separation of nodes within dynamic range: $\texttt{delta}_m$

While $\texttt{sep}_m$ nonparametrically estimates whether pairs of nodes are drawn from different distributions (and thereby distinguished), it provides no information about the distance by which they are separated within the metric's dynamic range. This measure gives an alternative look at separation between nodes: the more separation a metric provides between nodes, the less severe slight variations in assigned score will be to ignoring errors in generated images. We assess it as:

$$\texttt{delta}_m(S) = \frac{1}{|S|\sigma_{m(\forall S)}} \sum_{N_i \in W} \text{avg}(\{m(P,I)|(P,I) \in N_i\}) - \text{avg}(\{m(P,I)|(P,I) \in N_{i+1}\}) \tag{4.5}$$

Where $\sigma_{m(\forall S)}$ is the standard deviation of scores from metric m on all images in all SEGs in `TS2`. Our score is the average distance between the mean metric score

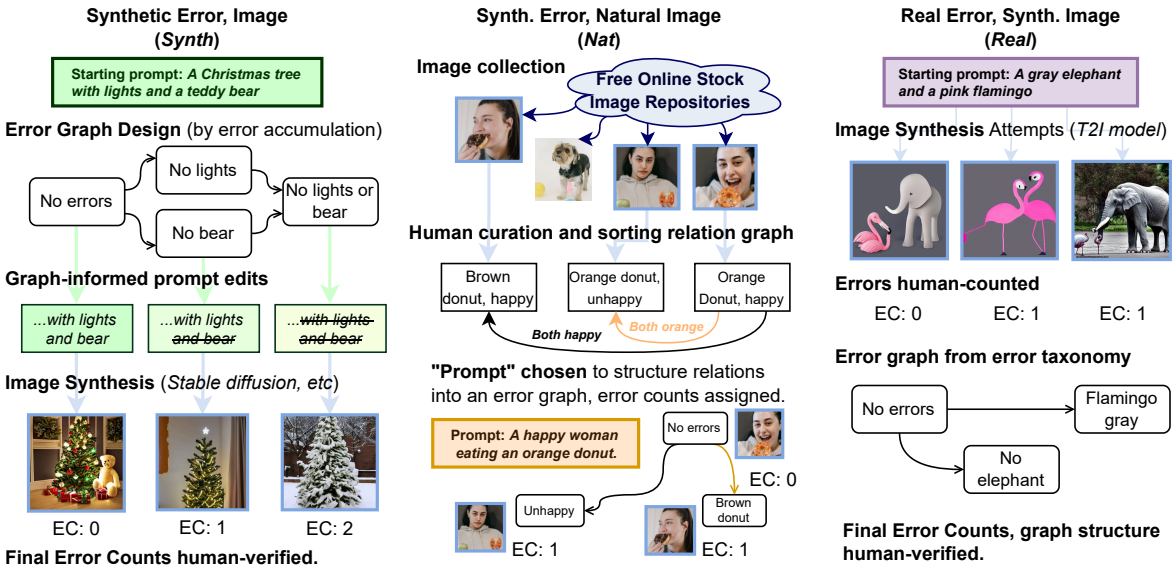


Figure 4.2: The three *semantic error graph* production procedures. **Synth.** (images generated from multiple prompts written to populate a SEG), **Nat.** (natural images populate a SEG), and **Real** (real errors from image generation attempts from one prompt populate a SEG).

of all adjacent nodes in all SEGs, rescaled by the standard deviation of the score to normalize against the metric's dynamic range.

4.3 The T2IScoreScore Dataset

We now turn to describing the TS2 dataset collection process. We use three different semantic error graph collection procedures to produce a diverse set of SEGs. Each contains one prompt and between 4 and 76 images assigned to error nodes. Each node usually contains more than one image, though for simplicity in presentation we only show one image assigned to each node in this section.

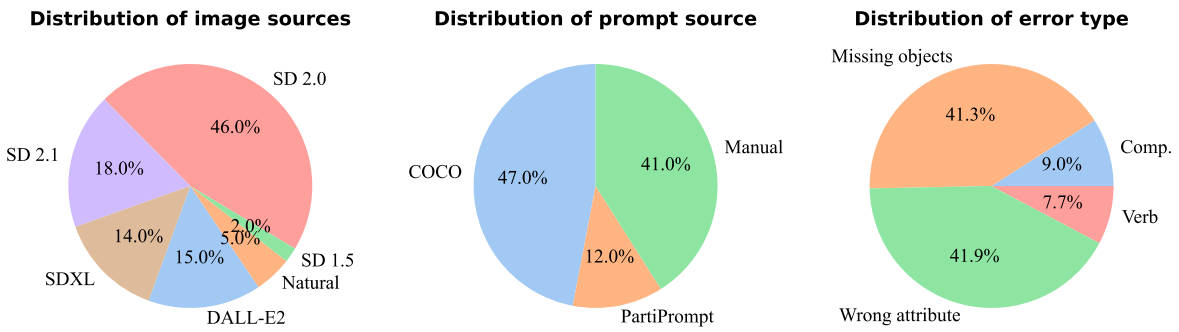
4.3.1 Dataset Collection Procedure

Figure 4.2 depicts the three procedures by which we produce and populate SEGs with images: **synth**etic images from a synthetic graph (*Synth*), graph from **nat**ural images (*Nat*), and graph from **real** errors of synthetic images (*Real*), differentiated by prompt source, image source, and the order of production.

Synth. Synthetic SEGs are produced "graph first." From an initial prompt we list all entities and properties it contains, then ablate them to produce an error graph. We then manually write prompts describing each node, generate their images, and manually check image-node faithfulness. For example, in the left panel of fig. 4.2, the initial prompt "a Christmas tree with lights and a teddy bear" is converted to error prompts such as "a Christmas tree with lights."

Nat. The natural error trees exclusively contain real images sourced from the free stock image repository Pexels. We generate SEGs in "image, graph, prompt" order. We source sets of natural images that share objects and models. We organize them by relation graphs describing how the images differ by objects, actions, attributes, and composition. We then select a head node in this relation graph and write a "prompt" describing this head node (eg., fig. 4.2 center panel). We produced SEGs of natural images to assess whether distributional differences between synthetic images and real images might lead to measurable impacts on performance for the faithfulness metrics that typically use base models pretrained exclusively on natural images [162].

Real. The real error, synthetic image SEGs are produced following a "prompt, image, graph" order. From a seed prompt (both manually written and sourced from COCO or PartiPrompts) we simply generate a large set of images using a T2I model, then



(a) Image Source (b) Prompt Source (c) Error node (SEG edge) type

Figure 4.3: Overview of the distribution of sample types in TS2: (a) Where source images came from: 5% of images in the benchmark are real photographs from Pexels, while the remainder were generated by Stable Diffusion (SD) or DALL-E variants. (b) Source of the eliciting prompt; either existing resources or us (Manual). (c) Distribution of error types edges in all SEGs.

annotate the errors in each generated image. These error-labeled images are then organized into a final error graph for the SEG. This procedure is documented in the right panel of fig. 4.2.

4.3.2 Dataset structure, size, and validity

Each of the 165 SEGs in TS2 is manually checked by three human annotators. The head node of each SEG contains at least one image that has been assessed by the annotators to contain no errors of verbal information (eg, entity in the image isn't performing the described action), compositionality (eg, object described as "on top" but is beneath object), missing objects, or incorrect object attributes.

Each edge on the SEG represents an error of one of the aforementioned types. Each node is labeled with the number of edges along its shortest path back to node 0, representing its error count (fig. 4.1). Each node contains at least one image which is erroneous according to the described errors.

The synthetic images are generated with several T2I models—including DALL-E 2 [125], Stable Diffusion 1.5 [126], 2.0, 2.1, and SDXL [163]. We use MS-COCO [164], PartiPrompt [156], and manually written prompts in head nodes for the Synth and Real subsets. Image sources, prompt sources, and error type statistics are documented in fig. 4.3.

The total number of images and prompts is roughly in line with previous benchmarks that have been used to verify methods such as TIFA [139] and DSG [129] in their own papers, and permits significantly more image comparisons-per-prompt than any prior benchmark (table 4.1), which we submit is the primary type of comparison required to verify that prompt-faithfulness assessments work.

4.4 Experiments

Using TS2 we evaluated three classes of T2I evaluation metrics: *embedding-correlation* (comparing embeddings of prompt and image), *QG/A* (using VQA to check if requirement questions generated from the prompt are satisfied) and *caption-based* (comparing captions extracted from the generated images to the prompt). We evaluate these metrics with multiple backend VLMs (§4.4.3).

For each metric, we score every image against its SEG's prompt. We report the results of our Ordering and Separation metrics across all SEGs, as well as for our the Synth, Nat, and Real SEG subsets.

4.4.1 Embedding-correlation Metrics

CLIPScore [138] is a popular prompt faithfulness metric based on simple text-image similarity. CLIPScore is computed as the cosine similarity of the L_2-normalized

CLIP-assessed [162] image and text embeddings.

ALIGNScore (not to be confused with the text-only *AlignScore* [165]) is a variant embedding-based similarity score we produced using the ALIGN [166] embedding model rather than CLIP to embed the prompt and image. Other than model it is equivalent to CLIPScore.

CLIPScore and ALIGNScore are computed using positive cosine similarity between text and image features, from feature extractor model $\mathcal{M}$ as:

$$\texttt{clip-s}(p, i) = \max(\cos(\mathcal{M}_{\mathrm{I}}(i), \mathcal{M}_{\mathrm{T}}(p)), 0) \tag{4.6}$$

4.4.2 Question Generation & Answering (QG/A) Metrics

Question Generation & Answering Metrics use an LM $\mathcal{M}_{QG}$ to produce a set of requirement question/answer pairs $(q, a) \in Q$ from prompt p, and then use a vision-language model $\mathcal{M}_{VL}$ to check each requirements against the image, reporting satisfaction rate as the image's faithfulness score. QG/A metrics vary by how questions are generated and relate to each other.

$$\texttt{tifa-s}(p, i) = \frac{1}{|Q|} \sum_{(q,a) \in Q} \mathbb{1}(\mathcal{M}_B(\mathcal{M}_{VL}(i, q)) = a) \tag{4.7}$$

TIFA [139] prompts an LM (GPT-3) to generate a set of multiple choice and yes-no questions and their expected answers relative to the prompt. Then a vision language model $\mathcal{M}_{VL}$ produces "free-form" answers to each question, which are converted into multiple choice answers a' using an SBERT model. The TIFA score for a given image is then the rate of correct answers.

DSG (*Davidsonian scene graph*) [129] shares the QA structure of TIFA, but generates a set of requirement questions which are non-overlapping, have exclusively yes/no

answers, and sit on a directed acyclic graph such that a question is only satisfied if it and all its parent questions are answered `yes`.

4.4.3 Backend VLMs

All QG/A metrics rely on the use of a generative vision-language model (VLM) either for performing question answering ($\mathcal{M}_{VL}$ in §4.4.2) or captioning ($\mathcal{M}_C$ in §4.4.4). Thanks to the simple decomposable framework of the QG/A methodologies, we were able to efficiently test the performance of both TIFA and DSG using several VLMs as visual question-answering backends $\mathcal{M}_{VL}$. We used mPLUG, LLaVA, BLIP, InstructBLIP, Fuyu, and GPT-4V as VLM backends for the QG/A metrics.

mPLUG is a class of vision-language models that use skip connections between visual encoder embedding layers between cross-modal attention blocks in the transformer stack [167]. We use the mPLUG-OWL [168] 7b checkpoint which uses LLaMA 7b [169] as the pretrained text encoder.

LLaVA is a fine-tune of Vicuna [170] (decoder-only transformer model) that uses a learned MLP "vision-language connector" layer to map a single input image's CLIP encodings into a shared embedding space [171, 172]. We use LLaVA 1.5 13b. Because LLaVA was instruction fine-tuned for chat applications, we experiment with a variant system prompt that requests *concise* answers from the system. We mark this alternate option LLaVa 1.5 (alt) in plots and figures.

BLIP is a jointly-trained self-attention ViT trained with cross attention to multiple transformer encoder and decoder pipelines with different tasks [173]. We use the BLIP encoder/causal LM decoder combination as a transformer encoder-decoder model to produce VQA answers from the `blip-vqa-base` checkpoint.

InstructBLIP extends BLIP by including an instruction fine-tuned "Q-Former"

that selects salient instruction-related visual features from a frozen ViT for input to a frozen LLM that answers the query conditioned on the selected features [174]. We use `instructblip-flan-t5-xl`.

Fuyu is a decoder-only VLM that splits an input image into a sequence of patches that are separately projected directly into the transformer embedding space, which jointly learns ViT and LM behaviors [175]. We use `Fuyu-8b`.

GPT4-V is the largest (as of June 2024) VLM provided by OpenAI [13].

4.4.4 Caption-comparison Metrics

LLMScore [140] captures the fine-grained similarity between the image and text with rationales by leveraging the visual details understanding capability from vision experts and the reasoning capability of LLMs. The visual information is parsed in hierarchical scene descriptions with global and local captions. Then the text-only LLM (we use GPT3) will compare the multi-granularity visual descriptions with the input text prompt to give a score according to the evaluation guideline prompt.

VIEScore [141] rates aspects of semantic consistency (SC) and perceptual quality (PQ) ultimately providing a rating score on a scale of 0 and 10. We use 0-shot LLaVA-1.5 as the backbone MLLM to evaluate how successfully the image follows the text-to-image prompt.

4.5 Results

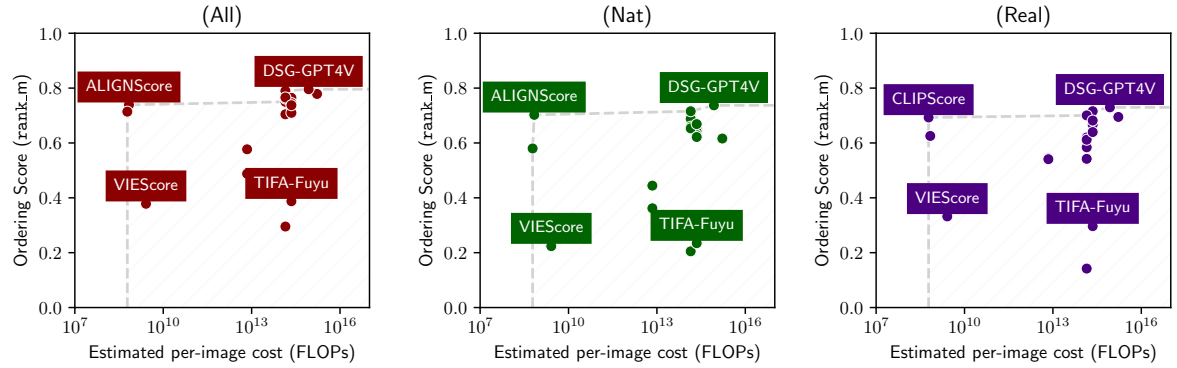
Table 4.2 shows the results for the **Ordering** ($\texttt{rank}_m$) and **Separation** ($\texttt{sep}_m$, $\texttt{delta}_m$) scores for each metric we assessed, on average for all SEGs (Avg), and the three SEG subsets, as well as an approximate FLOP cost per run for each metric (§4.5.1).

		CLIPScore	ALIGNScore	mPLUG	LLaVA 1.5	LLaVA 1.5 (alt)	InstructBLIP	BLIP1	Fuyu	GPT4-V	mPLUG	LLaVA 1.5	LLaVA 1.5 (alt)	InstructBLIP	BLIP1	Fuyu	GPT4-V	LLMScore EC	LLMScore Over	VIEScore
		Emb-based		TIFA							DSG							Caption-based		
$\texttt{rank}_m$	**Avg**	71.4	73.9	71.0	74.5	74.4	76.5	73.8	38.7	77.9	70.4	76.2	75.0	79.0	76.6	29.5	**79.6**	48.8	57.7	37.8
	Synth	75.0	77.6	72.6	79.2	79.2	80.2	78.8	44.5	83.6	74.6	80.1	81.6	**85.1**	81.6	35.4	82.6	50.2	61.6	42.5
	Nat	58.0	70.2	66.9	62.8	64.0	65.1	62.2	23.5	61.6	65.3	65.9	68.8	70.7	71.6	20.5	**73.7**	36.2	44.4	22.4
	Real	69.3	62.6	68.2	66.7	64.5	71.6	64.0	29.7	69.5	58.4	70.0	54.2	62.0	61.2	14.2	**73.0**	54.4	54.1	33.2
$\texttt{sep}_m$	**Avg**	90.7	**92.8**	80.6	82.5	81.9	85.0	81.8	67.2	83.2	78.4	83.1	80.3	84.2	80.8	63.6	84.2	73.6	73.5	51.8
	Synth	90.5	**94.1**	80.6	85.5	85.2	86.7	84.1	67.3	86.2	80.9	85.7	83.9	87.8	84.9	65.8	86.6	71.1	72.8	53.7
	Nat	91.5	**92.6**	84.2	75.1	75.6	82.8	77.9	75.7	80.5	71.2	80.9	76.7	81.8	73.3	68.6	81.7	80.5	76.7	44.5
	Real	**90.3**	87.9	77.4	76.8	74.4	80.5	76.4	59.3	73.7	75.1	74.5	69.4	71.9	70.8	50.3	76.6	77.3	73.6	50.7
$\texttt{delta}_m$	**Avg**	89.7	95.6	92.4	97.5	97.1	99.8	94.0	43.9	92.3	94.2	104.7	102.	**110.**	103.7	37.7	110.2	66.9	56.3	45.9
	Synth	89.9	95.6	92.5	97.6	97.3	99.9	93.9	44.6	92.5	94.4	104.7	102.	**110.**	103.6	37.8	110.1	67.2	56.7	46.7
	Nat	92.5	98.6	94.4	98.7	98.8	101.1	95.5	46.1	93.9	96.0	105.5	103.	**111.**	104.6	39.3	111.5	66.2	56.4	46.6
	Real	95.1	100.8	96.0	100.5	100.4	102.6	96.9	46.9	95.9	97.3	106.8	104.	**113.**	105.9	39.4	**113.0**	65.0	56.0	46.3
FLOP/run		**604!**	688M	224T	224T	224T	224T	224T	224T	1.66P	140T	140T	140T	140T	140T	140T	860T	7.01	7.01	2.6T

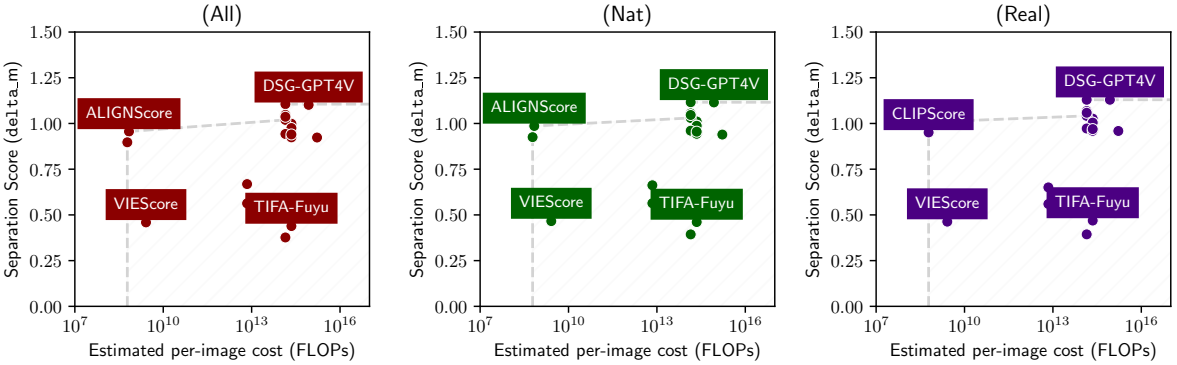
Table 4.2: Spearman ordering score $\texttt{rank}_m$, and Kolmogorov–Smirnov separation score $\texttt{sep}_m$, average dynamic range delta $\texttt{delta}_m$, (all reported as % for readability) and estimated FLOPs to score an image for each model. Best **bold**, within 2% of best underlined, top four colored by type (emb-based, TIFA, DSG, caption-based). See §4.5.1 for information on how we estimate compute costs.

We found that the Synth set consisting of hand-designed (and probably more obvious) errors was the easiest subset for all metrics to correctly order. The average $\texttt{rank}_m$ score for Synth across all metrics was 70%, for Nat 55% and for Real 56%. However, different subsets were hard for different classes of metrics. For the QG/A metrics, Real was hardest, while Nat was harder for the other classes.

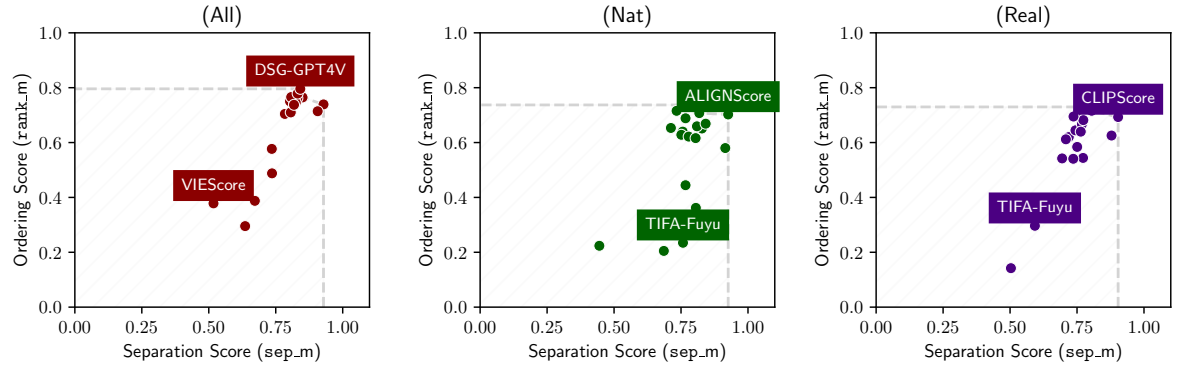
As the embedding-correlation metrics came first, TIFA [139], DSG [129], LLMScore [140], and VIEScore [141] all compare themselves against a CLIPScore baseline [138]. Despite the superiority these metrics supposedly held on their respective ad-hoc evaluations, *the computationally cheaper CLIPScore and ALIGNScore are Pareto-optimal in most cases* (fig. 4.4), sharing the optimality frontier with DSG or TIFA with GPT-4V, methods that are $\approx$ 6 orders of magnitude more computationally expensive.



(a) Ordering ($\texttt{rank}_m$) vs estimated cost/image (FLOPs), all metrics



(b) Separation ($\texttt{delta}_m$) vs estimated cost/image (FLOPs)



(c) Ordering ($\texttt{rank}_m$) vs Separation ($\texttt{sep}_m$)

Figure 4.4: Plots of ordering and separation scores against estimated per-image metric evaluation costs in FLOPs and each other. For all analyses, the Pareto optimal metrics are DSG and TIFA with GPT-4, and the vastly less expensive embedding-correlation ALIGNScore and CLIPScore.

4.5.1 Compute cost estimates

Estimating FLOPs per inference pass for each model. We use OpenAI's estimate [176] of $\approx 2N$ OPs per forward pass for a large transformer model, where N is the total number of parameters.

Obtaining parameter count estimates for closed models. TIFA, DSG, and LLM-Score use some combination of GPT3 and GPT4, whose exact parameter counts and FLOP/inference costs haven't been publicly disclosed. We use estimates from Semi-Analysis to get an approximate FLOP cost. While these numbers are likely imperfect, their orders of magnitude are as accurate as we can get.

Obtaining metric FLOP cost per single image eval. Given FLOP/forward pass estimates for each model, our estimates of the FLOP cost to evaluate an image is a function of the number of model calls and estimated tokens per call. The embedding-correlation metrics require 2 calls to an embedding model, matmul and sum operations to get the cosine similarity. TIFA requires on average 8 questions, with GPT-3 question-generating calls of average 40 tokens, and VQA model calls of average length 20 tokens. For DSG these numbers are 5, 40, 15. The costs of calling the freeform-to-multiple choice model are negligible. We estimate that LLMScore and VIEScore both require approximately 50 tokens of LLM or VLM compute to score an image. LLMScore's use of mPLUG to caption is negligible alongside the cost of running GPT-3.

Total compute cost of our study. In total, we estimate our study took 9.89×10^18 FLOPs, mostly through OpenAI's service, but also on lab-owned NVIDIA Titan-X and A-100 GPUs.

4.6 Discussion

The headline takeaway from our findings is that, contrary to claims of superiority leveled in their own papers, for all the QG/A and caption-comparison metrics [139, 140, 129, 141], *the cheap embedding-correlation metrics such as CLIPScore are sufficient or even preferable* at capturing objective semantic errors relative to fixed prompts. We view this capacity to accurately discriminate similar images relative to a prompt as the core feature a good prompt faithfulness metric must possess.

`T2IScoreScore` is effectively evaluating T2I metrics as relative *score regressors*—functions that are predicting a specific score for an image. However, there are additional desirable elements to a Human aesthetic preferences are—by design—ignored in `TS2` meta-evaluation. Though the QG/A and caption-comparison metrics fail to meaningfully outperform the cheap embedding-correlation metrics, they may have advantages that are not captured by `TS2`, such as in modeling human aesthetic preferences. Paired with a standalone benchmark of human aesthetic preferences over error-free images, a metric's error assessment and aesthetic fidelity could be measured independently.

As the first objective evaluation of faithfulness metrics based on structural semantic errors, `TS2` enables more fine-grained measurement of metric desiderata, leading researchers to build better metrics, and empowering developers to make trade off-informed metric choices.

4.6.1 Pareto frontiers with compute cost

Why do we care about compute cost? When evaluating T2I models at release, compute costs are not very important—an evaluation only has to be run on a small set of images from benchmark prompts.

However, during training or in online monitoring, compute cost for faithfulness metrics becomes quite important. Faithfulness metrics could be used as reward signals while training a T2I model (or prompt generator), or called repeatedly during validation passes. Faithfulness metrics could be deployed in applications to guide an online prompt refinement system, to trigger a second call in user-facing applications, or to analyze prompt corpora to surface challenging examples for further training or analysis. In all of these settings, a performant, low-cost model such as CLIPScore is valuable. TS2 demonstrates that the performance premium for the expensive metrics is quite small.

4.6.2 Considering error graphs enables objective evaluation

Previous evaluations of coherence metrics have evaluated metrics as human preference score regressors over single images, or over image pairs. The challenge with such an evaluation is that *human preferences are not objective*—especially when provided by a small pool of annotators, correlation to such scores is an unclear signal.

However, by instead evaluating walks over error counts in SEGs, the TS2 captures a more objective notion of correctness, by ignoring the subjective relationships between pairs of unconnected nodes.

For example, consider the SEG presented in fig. 4.1, where two different single-error nodes are shown. Given the prompt "a bot in a green shirt poses with some fruit," one of these nodes contains images without fruit, and the other contains boys wearing a blue shirt, rather than green. Which of these types of images are actually worse with respect to the prompt? This is a subjective decision—some annotators may find the missing fruits more important than the incorrectly colored shirt. TS2 ignores this distinction, as no nodes of equivalent error counts are connected in any SEG.

While the difference between those nodes is subjective, the difference between both of these nodes and their shared child node—one where fruits are missing **and** the shirt is incorrectly colored—is objective.

4.6.3 Human baselines and metric ignorance of ranking task

It is also important to note that metrics under test are not aware of the implicit ranking task in TS2, as they are evaluated as score regressors. Objective human annotation was only possible *because the annotators were aware that relative ranking was the goal*.

If human performance were judged on the task of simple Likert scoring of image-prompt accuracy without instructions, humans may not significantly outperform the metrics. However, if the human annotators were instructed to count the number of errors, we suspect they would perform quite well, even without the other images for comparison over which the ranking task is performed.

Though we provide no human baseline, we do not think this is a significant weakness—human performance on the inherently synthetic task of image quality scoring is not as important as performance on ranking along objective errors.

4.6.4 Systematic advantages for some metrics

One disadvantage of using Spearman's ρ is that it "expects" ties to be the same in both distributions. For example, if a set of images has error count $(0, 1, 1, 2)$, the ordering $(1, 0.5, 0.5, 0)$ will have a perfect $\rho = 1$, while the ordering $(1, 0.51, 0.49, 0)$ will be penalized, despite it also presenting a correct ordering. This means that **our Ordering score** $\texttt{rank}_m$ **systematically punishes the embedding-based metrics relative to the VLM-based ones**, as the embedding-correlation metrics CLIPScore and ALIGNScore

can take continuous values, whie TIFA, DSG, LLMScore, and VIEScore have a discrete range. In light of this systematic disadvantage for embedding-correlation metrics, it is even more striking that CLIP/ALIGNScore still are so performant and on the optimality frontier.

4.6.5 Near-perfect performance on $\texttt{rank}_m$ and $\texttt{sep}_m$ possible

Although the scores of many models on our metric are high, this meta-evaluation is far from "solved." In principle it should be possible to get much closer to 100 on average for both meta-metrics than we find. We view use of `TS2` as a necessary secondary evaluation for any new proposed T2I faithfulness metric; if it has high correlation to subjective human judgements but does not perform well on `T2IScoreScore`, skepticism might be warranted.

4.6.6 Impact of future VLM advances

Ultimately, all image coherence evaluation metrics stand to improve from further advances in general VLM quality. As a considerably more performant model than LLaVA, mPLUG, etc, it was unsurprising that GPT4-V worked much better as a backbone for TIFA and DSG than the aforementioned. However, there do appear to be diminishing returns, as the order of magnitude going from mPLUG- to GPT4-V-based evaluation yielded a sub-1% improvement in $\texttt{rank}_m$ performance on the most difficult and construct-valid Real set. Better constraint-generating processes may be required to push VLM-based evaluation metrics further.

4.7 Conclusion

We introduced `T2IScoreScore`, a first-of-its kind objective evaluation for text-to-image faithfulness metrics that utilizes a high image-to-prompt ratio to organize its reference images along semantic error graphs, through which a faithfulness metric can be assessed by our novel graph-based meta-metrics. Our study reveals a surprising finding: more expensive and recent "state of the art" VLM-metrics actually only have modest gains in performance over simpler and cheaper embedding-based metrics at best. Indeed, these cheap metrics such as CLIPScore and ALIGNScore are actually Pareto optimal along with the vastly more expensive and slightly more performant GPT-4V-based QG/A metrics, even when strictly comparing ordering and separation capabilities (leaving compute cost aside). This underscores the necessity for a more nuanced approach to benchmarking and developing metrics capable of capturing the subtle semantic nuances between prompts and generated images.

Careless design of evaluative targets was the fundamental problem in prior T2I faithfulness evaluations. Pairwise human preference over unrelated images *does not model faithfulness*. In `T2IScoreScore` we provided clear counterfactual differences between images that represent concrete errors relative to the target prompt: *these are the correct way to model faithfulness*. In **Chapter 5** I will present additional work critically examining existing evaluation practices in the domain of long-context vision-language modeling. It turns out that the poorly considered evaluative targets is a pervasive problem in AI research.

Chapter 5

Long-context vision-language models

Many multimodal domains beyond T2I faithfulness (**Chapter 4**) suffer from deep evaluative target issues. Long-context vision-language modeling evaluation is another task which relied on problematic, simplistic evaluations: needle-in-a-haystack retrieval.

Contrary to claims of near-perfect retrieval capabilities of models such as Gemini 1.5 across long contexts, we found that for tasks requiring reasoning to inform extraction from long inputs, a diverse set of vision-language models (VLMs) rapidly lose performance as the visual context length grows. This test assesses how well VLMs can ignore irrelevant information when answering queries—a task that is quite easy for language models (LMs) in the text domain—demonstrating that current state-of-the-art VLMs lack this essential capability for many long-context applications.

As open-weight long-context LMs have been around for the last few years [177], researchers have developed robust evaluations to compare their capabilities against those of proprietary LMs [178, 179]. However, until very recently [180], the small set of long-context VLMs supporting multi-image input have been inaccessible to the public [13], so evaluation practices for long-context VLMs [181, 182, 174] are in their infancy [183]. Most reasoning-based VLM evals have analogues in text-domain LM tasks. For

example, image captioning [167, 184, 185] can be likened to text summarization, and visual question answering (VQA) [186, 187, 188] to textual question answering (QA).

Text-based long-context tasks such as long-document question-answering (QA) [189], summarization [190], and retrieval-augmented generation (RAG) [191, 192] offer analogues to long-context reasoning in VLMs. Many of these tasks (e.g., QA, RAG) require natural language extractive reasoning, as discussed above.

Few VLM long-context evals measure extractive reasoning. Existing VQA benchmarks involving *distractor information* (thereby extraction) do not focus on long contexts, and long-context VQA benchmarks do not involve distractors. MultipanelVQA [193] includes distractor information, but only within single input images. Some VLM evaluations focused on hallucinations indirectly capture a notion of distraction [194, 195] but do not explicitly measure it alone. MILEBench [196] evaluates VQA in long contexts, but only on non-distractor tasks such as video summarization or difference detection, where all inputs are relevant.

Needle-in-a-haystack evaluation tasks (asking models to recover hidden passphrases at various positions) do require long-context extraction, but reasoning is not involved. Gemini 1.5 [178] extended this task to video comprehension by hiding a simple passphrase in many positions within a video. [197] present a static benchmark concurrently to ours centered on multimodal needle-in-a-haystack.

We aim to bridge gaps in multi-image VLM evaluation by drawing analogies to established long-context LM tasks. Long-document QA [198] is an easy-to-evaluate test that directly showcases model performance on a useful application, and hints at its potential in retrieval-augmented generation (RAG) more broadly. Central to these evaluations is the model's ability **to recognize which elements in a lengthy input are necessary for answering a query and which are not**, a skill we dub *extractive reasoning*. Do long-context VLMs also exhibit these extractive reasoning capabilities?

Image (X)	Question (Q)	Answer (A)
Randomly chosen from dataset $\mathcal{D}$	What cartoon character is on this cake?	Winnie the Pooh

1 image context 4 image *composed* context 9 image *interleaved* context

Figure 5.1: Example of Image (X) corresponding to question-answer pair (Q, A) under increasing visual context lengths in the *composed* and *interleaved* settings. *The green box for illustration only, not included in model inputs.*

While current VLM benchmarks align with long-document summarization tasks, such as whole-video question answering [199] or summarization [200], no extant benchmarks capture a VLM's ability to **filter out irrelevant images in a long context to reason or answer a query**—essentially, perform extractive reasoning over images.

Measuring extractive reasoning over image sequences is crucial. Just as extractive textual reasoning facilitates text-domain RAG, multi-modal RAG demands that VLMs efficiently reason over and extract information from documents featuring multiple images. Similarly, video QA necessitates that models focus solely on frames containing relevant information, much like long-document QA.

We introduce LoCoVQA, a benchmark generator for **Lo**ng **Co**ntext Visual Question Answering (VQA) **with distractors**. LoCoVQA enables the creation of long-context image understanding evaluations from any image comprehension dataset by presenting a sequence with the question-relevant image alongside a configurable set of visual distractors. This allows us to accurately assess how effectively VLMs extract **only the pertinent information to a query within a cluttered context**.

(a) 1 digit: [2] (b) 4 digits: [6, 6, 1, 0] (c) 8 digits: [9, 5, 2, 5, 3, 7, 5, 6]

Figure 5.2: Each subfigure represents a variable number of hidden MNIST digits in a 9 image composed context.

We find that even top proprietary VLMs struggle with this capability, even over short visual contexts, likely due to fundamental deficiencies in their training objectives. By unveiling this LoCoVQA identifies a crucial area for performance improvement in future VLMs. In this chapter, I will present work where we:

- Evaluate long-context extractive reasoning across a wide array of VLMs using LoCoVQA.
- Find that all existing VLMs exhibit significant fragility with increasing distractor context sizes.
- Use LoCoVQA to create a *visual needle in a haystack* test, revealing substantial positional biases in SOTA VLMs in extractive reasoning.

5.1 LoCoVQA Generation Method

Samples synthesized through LoCoVQA contain one or more *content* images X corresponding to a question and answer pair (Q, A). Content images can be sampled from various image comprehension benchmark, such as OK-VQA [201], MMStar [202], MNIST [203], and others. Alongside the content image(s), each sample includes up to

35 *distractor images*, which are either sampled in-distribution from the content image set (ensuring no content image collisions to prevent ambiguity, as described in §5.1.1) or out-of-distribution from other image sets such as MS COCO [164].

Samples can be arranged as **interleaved** image sequences for VLMs that accept multi-image inputs or as **composite** images arranged in a grid, as depicted in fig. 5.1. For all models capable of ingesting interleaved sequences, we evaluate both interleaved and composite examples.

Visual context refers to the visual elements within an image or sequence of images that are relevant to answering a question. This includes the *content* image and distractor images. The challenge is to identify and focus on the pertinent details while ignoring irrelevant information. We can sample images both in-distribution to create challenging visual contexts or out-of-distribution to isolate extraction capabilities in the simplest setting.

As a dynamic generator LoCoVQA can convert any vision-language dataset into a long-context benchmark [77]. We describe the three datasets we used below.

5.1.1 Single-image Reasoning Tasks

First, we discuss the visual reasoning benchmarks which we expand into long-context samples. For most question answering and visual reasoning tasks, this is the only LoCoVQA expansion that makes sense: few VQA samples can be combined such that a plausible new QA pair requires information from multiple images. Since most interleaved models we evaluate support 36 or fewer images as sequential inputs without modification, we do not evaluate any models with context lengths beyond 36. However, LoCoVQA scales to any size. For the single-context reasoning tasks, we exclusively sample the distractors in-distribution.

OK-VQA

OK-VQA [201] is a single-image visual question answering dataset containing 5,072 question-answer-image triples. It requires external knowledge to reason beyond the image. LoCoVQA generates in-distribution long-context OK-VQA samples, ensuring that no content images have concept collisions that may complicate evaluation. For instance, the question about a character on top of a cake, as shown in fig. 5.1, is sampled from OK-VQA.

Since OK-VQA is an open-domain, free-form answer dataset, we score the samples using three metrics: exact match (full score if the model's response contains any ground truth answer as a substring), and continuous text generation metrics BERTScore [204] and ROUGE_L [205] between candidates and references.

MMStar

MMStar [202] is a multi-domain visual question answering dataset combining examples from various existing datasets: 424 questions from SEEDBench [206], 366 questions from MMBench [207], 100 questions from MMMU [208], 345 questions from MathVista [209], 69 examples from ScienceQA [210], and 196 examples from AI2D [211]. MMStar contains 1,500 high-quality multiple-choice questions that *require visual information from the images* to answer, a filtering step not initially performed on the source datasets. For example, over 50% of ScienceQA questions can be solved by a text-only LLM [202]. Similar to OK-VQA, we generate LoCoVQA samples for MMStar using the collision filtering technique to produce pseudo-documents composed of multiple example images.

Filtering Collisions in LoCoVQA

To address the problem of *content-distractor collisions*—where multiple similar in-distribution images in the visual context make the QA pair ambiguous—we implement a robust LM-based filtering method. For each visual context image, we prompt GPT-4 to list the top five entities; if there is overlap, we consider the question potentially ambiguous. To validate this approach, we manually assessed a subset of LoCoVQA generations and found it to be consistent, with no such collisions.

5.1.2 Multi-image Reasoning Tasks

In §5.1.1, we explored tests designed to evaluate whether VLMs can extract a single relevant image from a sequence to answer a query, thereby probing their long-context reasoning capabilities. Extending this to test how well VLMs can extract information from a multi-image sequence is a natural progression. However, VQA examples cannot be easily combined in a way that requires multiple images to answer a single question. Therefore, we turn to constructing "sequential VQA" sets using synthetic tasks. Optical character recognition (OCR) is a straightforward task to convert into multi-image question answering by including multiple OCR examples as interleaved images and asking the VLM to list all the text.

MNIST

We use MNIST [203] as it is a canonical dataset for OCR. For a desired visual context length, we sample between 1 and 8 randomly-colored digits from the MNIST training set of around 60K images, resizing them to between 1/6 and 1/2 of the maximum height of other context images. The remaining distractor images are randomly sampled from a subset of 5K high-quality MS COCO [164] validation

Vision-Language Model	Context	Base LM	↓sample
Single-Image Input VLMs			
Moondream2-1.6B	2K	Phi-1.5	✓
LLaVA-1.5-7B	4K	Vicuna	✓
LLaVA-1.6-7B	8K	Mistral	✓
PaliGemma-3B	8K	Gemma	✓
Multi-Image Input VLMs			
Mantis-Bakllava-7B	8K	Mistral	✓
Mantis-Idefics2-8B	8K	Mistral	✓
Gemini 1.0 Pro Vision	32K	Gemini	-
GPT-4 Vision	128K	GPT-4	-
Gemini 1.5 Flash	1M	Gemini	-

Table 5.1: Overview of open-source and proprietary vision-language models (VLMs). *Downsampling* models (↓**sample**) rescale input images in the single-image setting rather than take as many ViT inputs as necessary.

images. The VLM is then prompted to *list all handwritten digits present in the sequence*.

By varying the number of digits in the sequence, we can dynamically adjust the difficulty level of the multi-image distractor OCR task.

5.2 Experiments

We evaluate the performance of nine current vision-language models on our LoCoVQA-generated benchmarks. The open-weight models tested are Moondream2 [212], LLaVA-1.5 [172], LLaVA-1.6 [213], PaliGemma-3b [214], and two Mantis variants [180]. Both Mantis variants are VLMs further tuned on interleaved multi-image instruction following. The three proprietary models we evaluate are GPT-4V [13], Gemini 1.0 Pro Vision [59], and Gemini 1.5 Flash [178]. table 5.1 showcases all the models along with the model context sizes and image downsampling.

Interleaved image support. All closed-source models and Mantis support multi-image interleaved inputs, while the others only support single images. For multi-image models, we evaluate both the **composed (cmp)** & **interleaved (int)** settings. For single-image models, we only test the **composed (cmp)** setting.

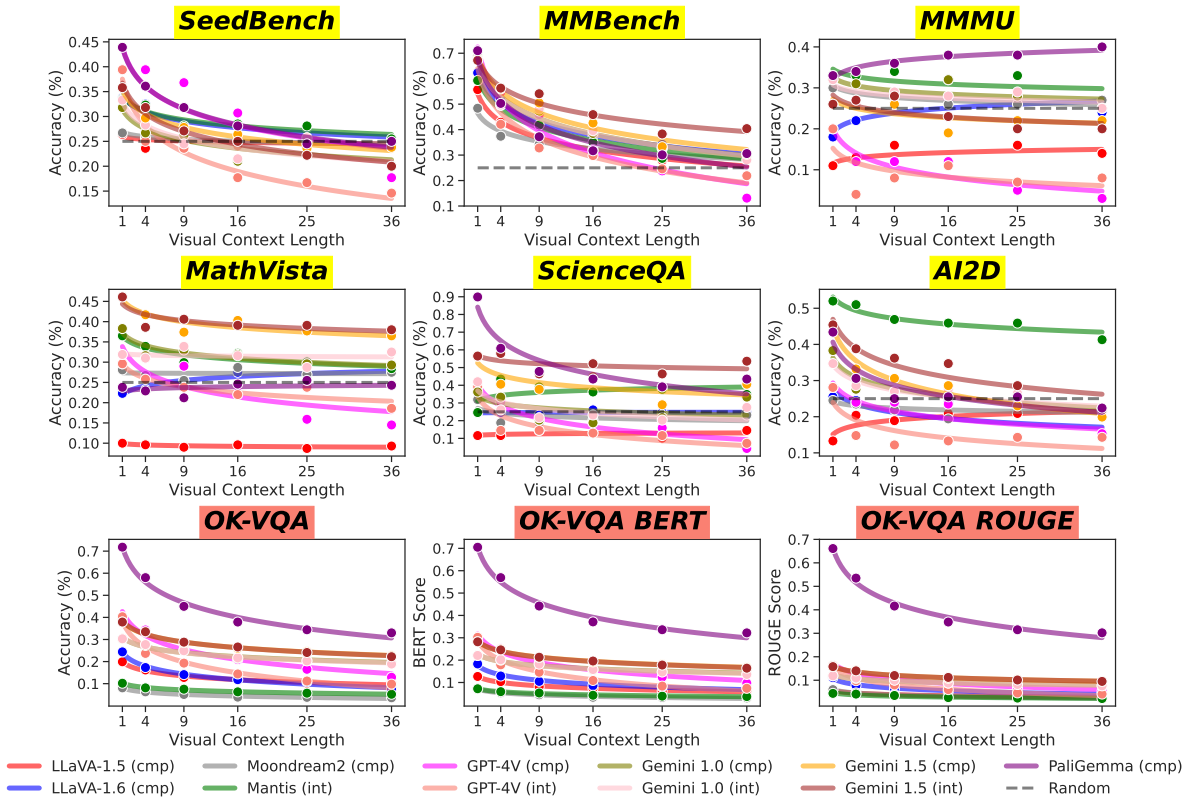


Figure 5.3: VLM Performance on MMStar and OK-VQA. Note the clearly declining logarithmic fit trends for many of the models. The (model, task) pairs for which these trends do not hold by and large are *below the random baseline.*

Downsampling images. Some models accept images of arbitrary resolution, while others automatically downscale inputs. This difference is crucial, especially in the **cmp** setting, as increasing the visual context can lead to information loss. For the downsampling models that support both **cmp** and **int** settings, we assess both. Any performance differences between these settings would highlight the impact of downsampling on the **cmp** setting.

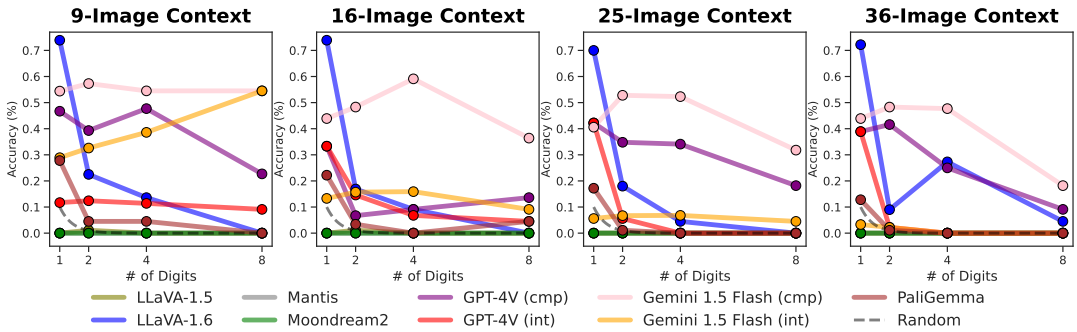


Figure 5.4: VLM Performance on the MNIST-Digits transcription task as a function of # of digits to transcribe. This demonstrates the relationship between "task difficulty" and performance, at four context sizes.

5.3 Results

Figure 5.3 illustrates how model performance (across 10 experiments, in both composed and interleaved settings) changes with increasing visual context lengths on single-image LoCoVQA tasks. The first two rows display results from the MMStar dataset, which contains filtered-image-necessary examples from six other datasets [202], with titles in yellow and random guess thresholds indicated by dotted black lines. The bottom row presents OK-VQA scores using three scoring metrics, with titles in salmon.

Across all models on OK-VQA and the majority on MMStar subsets for SeedBench, MMBench, ScienceQA, and AI2D, we observe *striking logarithmic decay trends* in model performance with increasing visual context length. In general, the closed-source models outperform their open-weight counterparts, especially on multiple-choice tasks. Among the open-weight models, PaliGemma performs the best, likely due to its substantially more extensive training on vision-language tasks compared to its counterparts like LLava or Mantis.

On several multiple-choice tasks, Mantis (the interleaved open-weight model) outperforms the other open-weight models that rely solely on composite inputs. This

is likely due to the image subsampling required by several of the other open-weight models negatively impacting performance when using composite inputs. However, on the OK-VQA task, Mantis and Moondream are the least performant, even at low context lengths. These models were likely not trained on visual question-answering tasks to the same extent as LLaVA variants during their instruction tuning steps.

The most noteworthy point is this: **the logarithmic decay trend holds equally well in interleaved and composed settings**. This indicates that the performance-visual context length trend is fundamental and **cannot be attributed solely to downsampling effects in the composite setting**. Furthermore, for the models tested in both conditions (GPT-4V and Gemini), the same trends are observed in both interleaved and composite settings. Sub-chance performance on the multiple-choice question-answering datasets, particularly for the closed-weight models, occur due to high refusal rates. This may illustrate how "alignment hampering" can hinder performance.

Performance on Multi-image Tasks. Figure 5.4 presents model performance on the MNIST transcription task by number of digits to transcribe. **Note the x-axis in this figure is changed; it is checking for a fixed context length the relationship to number of digits.** While there is a trend of decreasing performance with an increasing number of images, it does not follow a simple pattern like the context-length trend. For example, Gemini 1.5 experiences minimal performance degradation even as the target digit length increases.

Characterizing difficulty as a function of digit length is complex—as digit counts increase in a fixed context window, *the the output label search space grows*, while the *ratio of relevant images to irrelevant images increases*. This may explain why some models have *consistent* or even *increasing* performance as the number of digits increases. Analyzing why different models handle these axes of difficulty differently is an interesting future

direction.

However, akin to the single-image tasks, increasing the overall *visual context length* (seen in same-color, x value points across plots), makes the task more difficult, albeit without as clear a correlation.

5.4 Discussion

Why do we observe such striking logarithmic decay in performance with increasing visual context length? The fact that the trend is consistent in both interleaved and composite settings that downsampling effects in the latter are not the primary cause of performance decay. There may be an information-theoretic interpretation of this behavior: an increasing signal-to-noise ratio. As the visual context length increases, the amount of signal (relevant information) remains constant, while the amount of noise (distractor information) increases.

However, this signal-to-noise ratio problem also exists in text-only tasks, yet long-context LLMs perform well on needle in a haystack and long-context QA and summarization tasks, underlying their performance in retrieval-augmented generation. When it comes to visual contextual tasks that involve distractor information, VLMs struggle to perform even on the easiest long-context tasks, such as transcribing MNIST characters. They are even less capable on more complex, real-world tasks like distractor VQA. The core issue appears to be that both short-context and long-context VLMs are not trained on tasks requiring attention over multiple context images to retrieve information. In contrast, for long passages of text, information from throughout the passage remains relevant—referring back to a specific sentence early in a document is intrinsic to the LM training objective.

However, the same may not necessarily be true for the sequential image language

modeling task. In the existing interleaved vision-language training corpora, images are likely to followed immediately by their relevant text. As a result, attending to much earlier images in the documents may not be crucial for achieving low LM loss. Additionally, none of the tested open VLMs are trained on image *generation* tasks conditioned on text while updating the core transformer weights. Overlooking this objective may prevent VLMs from robustly modeling the relationship between images and text, which could be crucial for performing well on these tasks.

Interesting examples of poor performance Our analysis unveiled a few particularly surprising results. For example, contrary to our expectation that models capable of handling interleaved input would perform better with it than with composite input, we found that GPT-4V actually favors composite input on many subtasks. Additionally, to our surprise LLaVA 1.6 performed much better than GPT-4V on the composite haystack task at all sizes. This outcome was driven by multiple factors: GPT-4V tends to refuses some tasks rather than guessing, while LLaVA always provides a guess. Furthermore, GPT4V often "hallucinates" multiple numbers instead of just one.

Possibility of memorization Some of the surprising performance results are driven by the training mix. For example, Mantis's significant lead on AI2D, and LLaVA's strong performance on the MNIST task may result from having more relevant data or those specific training sets included in their data. However, even considering that some models were trained on these tasks, the pronounced drops in generalization performance as the context length increases are even more striking. This illustrates that current VLMs fundamentally struggle to attend image sequences as well as they do with text.

Upper bound for video QA A more construct-valid test for real-world visual extractive reasoning would be QA over a long video where only a sub-element is required. E.g., putting the entire film *Star Wars* as input for QA pair (Q: "which character shoots a green alien in the Mos Eisley cantina?", A: Han Solo). Information from a single scene must be extracted to perform this task, thereby requiring extractive reasoning. However, it is plausible that sequences of related images, unlike sequences of unrelated images, are more in-distribution for long-context VLMs and extractive tasks involving them could be easier for them to solve. By providing completely unrelated images, our task may be harder than video-based VQA, and may represent an upper-bound for visual extractive reasoning.

5.5 Conclusion

Vision-language models struggle with long-context visual extractive reasoning. Many models fail to extract necessary information from even short contexts of 36, 25, or 16 images, resulting in near- or sub-baseline performance across tasks. LoCoVQA presents a simple benchmark-generating process applicable to any VQA or retrieval evaluation dataset, making it easy to assess extractive reasoning in VLMs. Our findings suggest that training tasks requiring attention across multiple context images to extract information–rather than simple single-image tasks–should be included in VLM training. By measuring this capacity it offers an appealing direction for future work.

Part III

Multilingual and multicultural multimodal systems

Chapter 6

Multilingual text-to-image models

I conclude this dissertation by using the tools, strategies, and lessons above to enable measurement and assessment for a key direction for social good: global accessibility of generative AI systems.

In this chapter, I present a benchmark for assessing conceptual coverage across languages (***CoCo-CroLa***) in T2I models. This project was inspired by a surprising incident I encountered when first playing with a diffusion model—an offensive collision between the concept of a "dog" in Japanese with pictures of women by the model (fig. 6.3). ***CoCo-CroLa*** represents an effort to scalably and automatically find failure modes like this as well as quantify T2I model cross-lingual performance.

More precisely, ***CoCo-CroLa*** measures the degree to which any generative text-to-image system provides parity between some test language and its training language by accurately and consistently generating images of tangible nouns. For each model we can assess "conceptual coverage" of a given target language relative to a source language by comparing the population of images generated for a series of tangible nouns in the source language to the population of images generated for each noun under translation in the target language. This technique allows us to estimate how well-

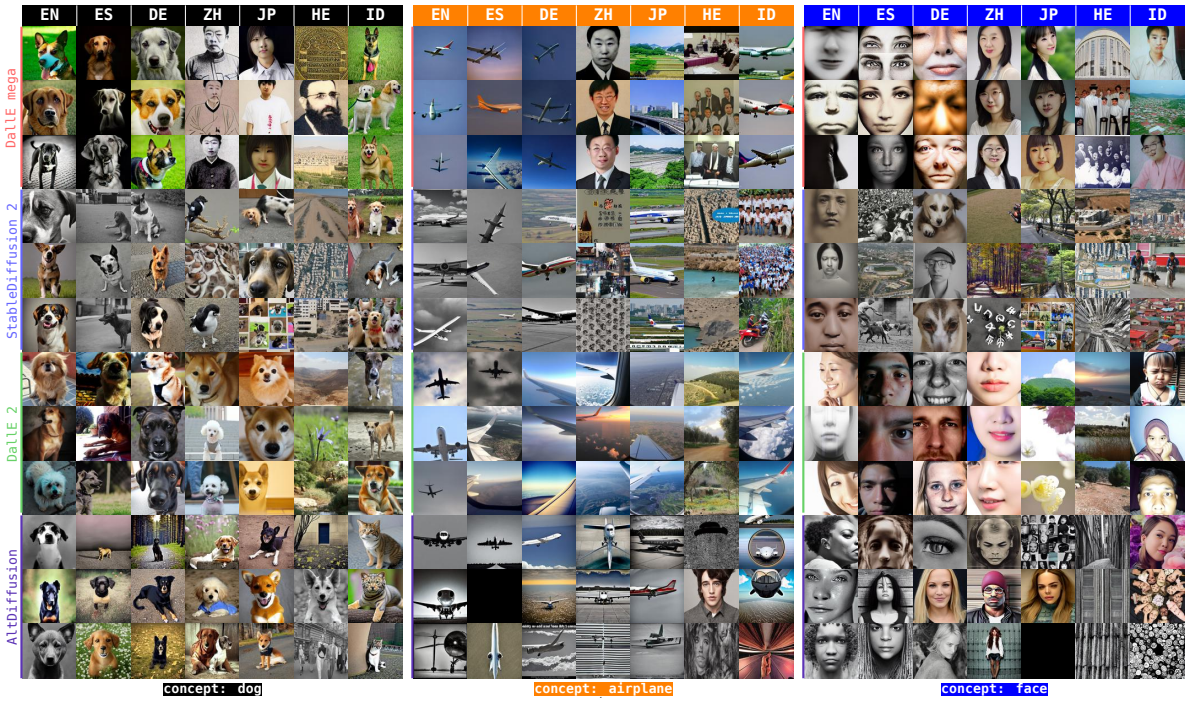


Figure 6.1: Figure 1: A selection of images generated by DALLE-mega, Stable Diffusion 2, DALLE-2, and AltDiffusion, illustrating their *conceptual coverage* of "**dog**," "**airplane**," and "**face**" across English, Spanish, German, Chinese (simplified), Japanese, Hebrew, and Indonesian. Coverage of the concepts varies considerably across model and language, and can be observed in the consistency and correctness of images generated under simple prompts.

suited a model is to a target language as well as identify model-specific weaknesses, spurious correlations, and biases without a-priori assumptions. We demonstrate how it can be used to benchmark T2I models in terms of multilinguality, and how despite its simplicity it is a good proxy for impressive generalization.

CoCo-CroLa satisfies many of the desiderata set out in **Part I**. It is a **dynamic** benchmark, generated entirely using automated systems, that can be scaled to any language and any set of tangible concepts. Furthermore, **saturation on this benchmark means something**. When a model achieves high performance on ***CoCo-CroLa*** for a given language, we can expect it to have a high degree of performance parity in that language to English. ***CoCo-CroLa*** is a measure of T2I model multilinguality.

Figure 6.2: We hypothesize that a model's ability to generate creative, compositional images depicting tangible concepts (e.g., astronaut, horse, soup, bear) is predicated on its ability to generate simple images of the concepts alone. Samples from [125].

6.1 Introduction

Neural text-to-image models convert text prompts into images [215, 216] using internal representations reflective of the training data population. Advancements in conditional language modeling [217], variational autoencoders [218], GANs [219], multimodal representations [162], and latent diffusion models [126] have given rise to sophisticated text-to-image (T2I) systems that exhibit impressive *semantic generalization capabilities*, with which they generate coherent, visually-appealing images with novel combinations of objects, scenarios, and styles [124]. Their semantic latent spaces [220] ground words to associated visuals [221]. Characterizing the limits of these systems' capabilities is a challenge. They are composed of elements trained on incomprehensibly large [222, 166], web-scale data [223, 224], hindering training-data-centric model analysis [225, 226] to address this problem.

Demonstrations of novel T2I model capabilities tend to rely on the subjective impressiveness of their ability to generalize to complex, novel prompts (fig. 6.2). Unfortunately, the space of creative prompts is in principle infinite. However, we observe that **impressive creative prompts are composed of known, tangible concepts**.

Can we directly evaluate a model's knowledge of these tangible concepts as a partial proxy for its capability to generalize to creative novel prompts? Perhaps. But

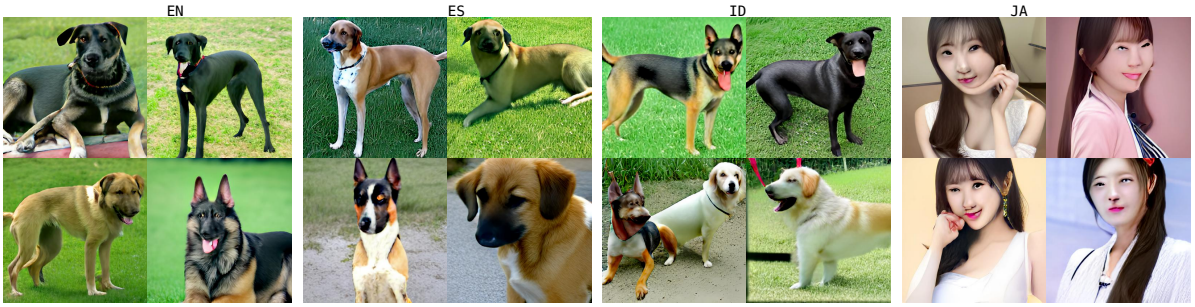


Figure 6.3: Although DALL-E mini [227] is ostensibly trained only on English data, when elicited with "big dog" in Spanish, Indonesian, and Japanese it generalizes the "dog" concept to ES and ID, while exhibiting an offensive concept-level collision in JA.

finding a diverse set of significant failure cases of basic concept knowledge for these models is challenging—in their training language.

We observe that when prompted with simple requests for specific tangible concepts in a constrained style, T2I models can sometimes generate consistent and semantically-correct images in languages for which they received limited training (Figure 1, fig. 6.3). We refer to this capacity as *concept possession* by said model in said language. At scale, we can assess the language-concept possession for a diverse array of concepts and languages in a model to attempt to describe its overall multilingual generalization capability. We refer to the degree of this capability as the model's multilingual *conceptual coverage*. In this work we:

1. Introduce objective measures of *multilingual conceptual coverage* in T2I models that compare images generated from equivalent prompts under translation (fig. 6.4).

2. Release ***CoCo-CroLa***, a benchmark set for conceptual coverage testing of 193 tangible concepts across English, Spanish, German, Chinese, Japanese, Hebrew, and Indonesian.

3. Validate the utility of *conceptual coverage analysis* with a pilot study suggesting that generalization to complex, creative prompts follows concept possession.

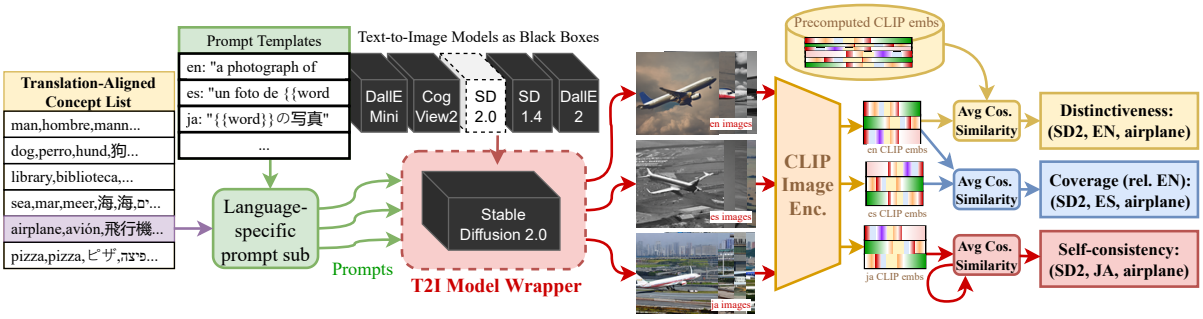


Figure 6.4: ***CoCo-CroLa*** assesses the cross-lingual coverage of a concept in a model by plugging all the term translations into prompt templates, generating a set of images from a model under test, extracting their corresponding CLIP embeddings, and computing concept-level **distinctiveness**, **coverage**, and **self-consistency** for the concept with respect to each language. (Demo and code available at `github.com/michaelsaxon/CoCoCroLa`)

Our benchmark enables fine-grained concept-level model analysis and identification of novel failure modes, and will guide future work in increasing the performance, explainability, and linguistic parity of text-to-image models.

6.2 Motivation & Related Work

This work is an attempt to produce a scalable, precise technique for characterizing conceptual coverage with **minimal assumptions** about the concepts or models themselves. In this section we lay out our motivations alongside relevant related work.

Benchmarks enabling model comparability have been a driving force in the development of pretrained language models (LM) [107]. For classification and regression tasks, evaluation under fine-tuning [228, 229] is a straightforward and practical proxy for pretrained LM quality [230] (e.g., for encoder-only transformer networks [105]). For these classification models, higher performance on benchmark datasets [231, 232, 26] became the epitome of LM advancement. However, other important qualities in models including degree of social biases [233] and robustness [234] arising

from biases in training data [69] can only be captured by more sophisticated benchmarks that go beyond simple accuracy [235]. CheckList represented an influential move in this direction by benchmarking model performance through *behavioral analysis under perturbed elicitation* [236].

In contrast, generative large language models (LLMs) such as GPT-3 [237] have a broader range of outputs, use-cases, and capabilities, making evaluation more difficult. For many text-generative tasks such as summarization and creative text generation, the crucial desired quality is subjective, and challenging to evaluate [238]. However, as these LLMs operate in a text-only domain, existing supervised tasks could be ported to few-shot or zero-shot evaluations of LLM capabilities [239]. While performance on these benchmarks isn't directly indicative of the impressive generative performance and generalization capabilities, they are a means to measure improvement [240].

Text-to-image models are even more difficult to evaluate than LLMs. Unlike in LLMs, there are no objective evaluation tasks that can be directly ported as proxy evaluations. For example, while GPT-3 was introduced with impressive zero-shot performance across many classification tasks, the T2I model DALL-E 2 was primarily introduced with human opinion scores and cool demo images [125]. Prior efforts in developing T2I evaluations such as Drawbench [127] and DALL-Eval [157] fall into the trap of trying to build "everything benchmarks" for which whether the benchmark accurately reflects the practical task being asked of the computer in its real-world context is difficult to assess [23]. We instead seek to build an *atomic benchmark* which narrowly and reliably captures a specific characteristic—conceptual knowledge as reflected by a model's ability to reliably generate images of an object.

Multilingual conceptual coverage is a high-variation T2I model performance setting. (Figure 1) Perhaps more importantly, it has immediate value, as work on improving T2I model multilinguality has has been proposed, but hampered by a lack

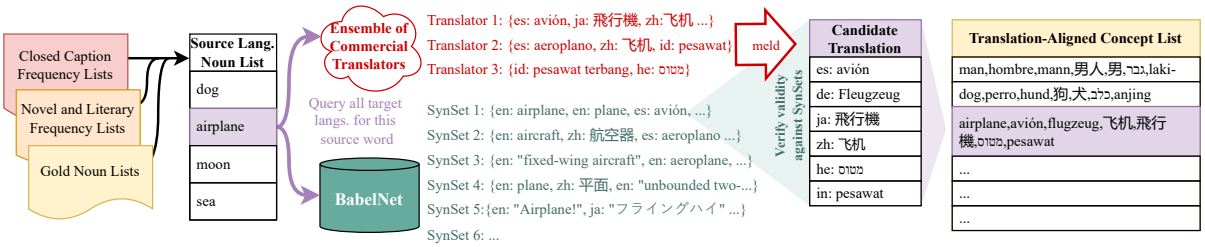


Figure 6.5: A diagram of our approach for producing the aligned noun concept list across the target language set using an ensemble of cloud translation services and BabelNet. Full description of this method in §6.4.1.

of evaluation metrics.

Chen et al. [241] introduce AltCLIP and AltDiffusion, models produced by performing multilingual contrastive learning on a CLIP checkpoint for an array of non-English languages including Japanese, Chinese, and Korean. Without an objective evaluation benchmark, they can only demonstrate their improvement through human evaluation of impressive but arbitrary examples. ***CoCo-CroLa*** improves this state of affairs by enabling CheckList-like direct comparison of techniques for reducing *multilingual conceptual coverage* disparities as an objective, capabilities-based benchmark.

Excitingly, we find that **conceptual coverage is upstream of the impressive T2I model creativity** that model developers and end-users are fundamentally interested in. This means that not only is ***CoCo-CroLa*** an objective evaluation of T2I system capabilities, it is also a **proxy measure for the deeper semantic generalization capabilities we are interested in enhancing** in second languages, as we demonstrate in §6.5.6.

6.3 Definitions & Formulations

We define a multilingual *concept* over languages L as a set of words in each language carrying the same meaning and analogous colloquial use. We refer to the

equivalent translation of concept c_k in language ℓ as $c_{k,\ell}$.

Given a set of concepts C, test language ℓ, a *minimal eliciting prompt*[1] MP_ℓ, text-to-image model f, and a desired number of images-per-concept n, we sample $n|C||L|$ images $I_{c_k,\ell,i}$, where

$$I_{c_k,\ell,i} \sim f(\text{MP}_\ell(c_{k,\ell})) \tag{6.1}$$

For every concept word in the language ℓ $c_k \in C$.

Given an image feature extractor F, some similarity function $\textsc{Sim}(\cdot,\cdot)$, we assess whether f *possesses* concept $c_{k,\ell}$ in the test language if using the following metrics from the concept-image set $\{I_{c_{k,\ell},i}\}_{i=0}^{n}$ (***CoCo-CroLa*** scores in fig. 6.4):

Distinctiveness. The images are *distinct* if they tend to not resemble the population of images generated for other concepts in the target language.

Formally, we compute our (inverse) distinctiveness score $\text{Dt}(f,\ell,c_k)$ relative to m images sampled from other concepts in C:

$$\text{Dt} = \frac{1}{mn}\sum_{j=0}^{m}\sum_{i=0}^{n}\textsc{Sim}(F(I_{c_k,\ell,i}), F(I_{c_r,\ell,s})), \tag{6.2}$$

$$c_{r,\ell} \sim C \setminus c_{k,\ell}, \quad s \sim U\{0,n\} \tag{6.3}$$

Self-consistency. The images are *self-consistent* if they tend to resemble each other as a set.

[1]We define a minimal eliciting prompt as a short sentence with a slot for concept work insertion, intended to enforce style consistency without interfering with the concept.

Formally, we compute the self-consistency score $\text{Sc}(f, \ell, c_k)$ as:

$$\text{Sc} = \frac{1}{n^2 - n} \sum_{j=0}^{n} \left(\sum_{i=0}^{n} \textsc{Sim}(F(I_{c_{k,\ell},i}), F(I_{c_{k,\ell},j})) - 1 \right) \tag{6.4}$$

We subtract 1 from each step in the numerator and n from the denominator so that identical matches generated image to itself.

Correctness. The images are *correct* if they faithfully depict the object being described.

Rather than assess this using a classification model (hindering generality depending on the pretrained classifier), we use faithfulness relative to a source language ℓs, cross consistency $\text{Xc}(f, \ell, c_k, \ell s)$ as a proxy:

$$\text{Xc} = \frac{1}{n^2} \sum_{j=0}^{n} \sum_{i=0}^{n} \textsc{Sim}(F(I_{c_{k,\ell},i}), F(I_{c_{k,\ell s},j})) \tag{6.5}$$

It is important to note that in the source language (eg. English), $\text{Xc} \approx \text{Sc}$, as for that language both metrics are essentially comparing the same (concept, language) pair to itself.

Thus we augment with a second language-grounded correctness score, utilizing the average text-image similarity score of the English concept text against the set of generated images, for a CLIP image encoder F and text encoder F_t, Wc:

$$\text{Wc} = \frac{1}{n} \sum_{i=0}^{n} F_t(c_{k,\ell_s}) \cdot F(I_{c_{k,\ell},i}) \tag{6.6}$$

Model	Authors (Year)	Repository	Training Language
DALL-E Mini	[227]	`github:borisdayma/dalle-mini`	EN
DALL-E Mega			
CogView 2	[242]	`github:THUDM/CogView`	ZH
StableDiffusion 1.1	[126]	`HF:CompVis/stable-diffusion-v1-1`	EN
StableDiffusion 1.2		`HF:CompVis/stable-diffusion-v1-2`	
StableDiffusion 1.4		`HF:CompVis/stable-diffusion-v1-4`	No language filter
StableDiffusion 2		`HF:stabilityai/stable-diffusion-2`	
DALL-E 2	[125]	`openai.com/dall-e-2/` (no checkpoints)	No language filter
AltDiffusion m9	[241]	`HF:BAAI/AltDiffusion-m9`	EN, ES, FR, IT, RU, ZH, JA, KO

Table 6.1: The set of text-to-image models we evaluated with ***CoCo-CroLa*** v1.0. Some monolingual models may integrate pretrained elements such as CLIP checkpoints that have been trained on multilingual data.

6.4 Approach

We compute distinctiveness, self-consistency, and correctness scores across English, Spanish, German, Chinese (Simplified), Japanese, Hebrew, and Indonesian on the models listed in table 7.2.

For each (language, concept) pair, we generate 10 images for analysis. We use a CLIP [162] checkpoint from HuggingFace[2] as our semantic visual feature extractor F, and cosine similarity as our similarity function ($\textsc{Sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b} / ||\mathbf{a}|| ||\mathbf{b}||$). We collect a translation-aligned concept list C using techniques described in §6.4.1 and depicted in fig. 6.5. We release our list generation code, testing code, feature extraction code, and final concept list as ***CoCo-CroLa*** v1.0[3].

6.4.1 Translation-aligned concept set collection

We automatically produce an aligned multilingual concept list, where meaning, colloquial usage, and connotations are preserved as well as possible. We identify *tangible nouns* describing physical objects, animals, people, and natural phenomena

[2] `HF:openai/clip-vit-base-patch32`.
[3] Demo and code: `github:michaelsaxon/CoCoCroLa`

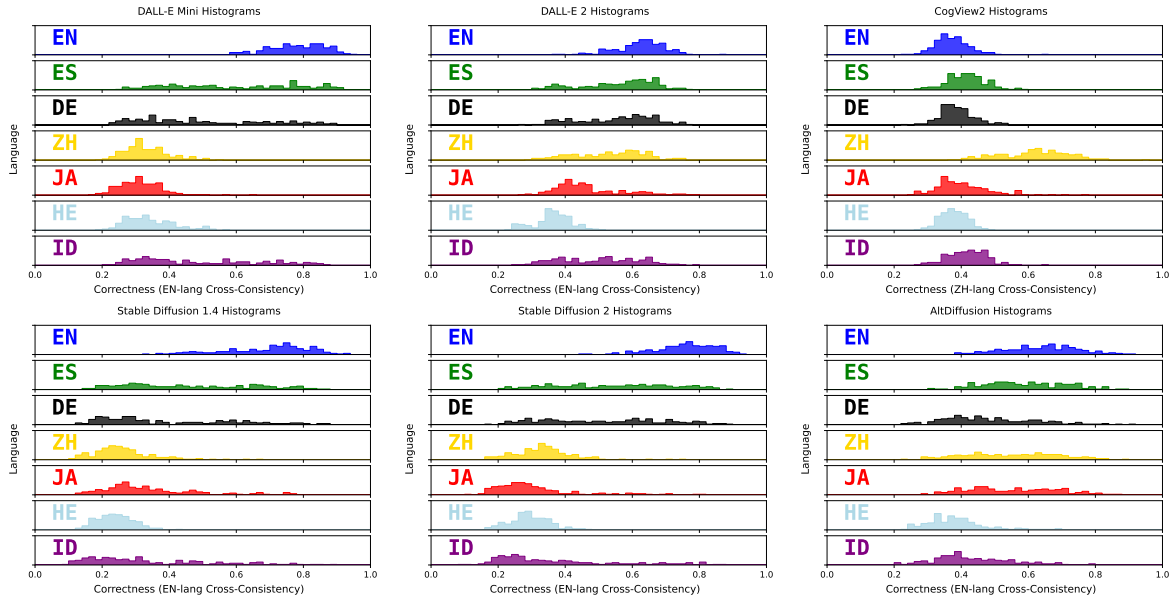


Figure 6.6: Histograms of the distribution of **correctness** cross-consistency (Xc) for each test language for six assessed models. Rightward probability mass reflects better conceptual coverage.

as a class of concepts that are both straightforward to evaluate and tend toward relative ubiquity in presence as words across languages and cultures, to the extent this ubiquity can be ensured through automated means.

Automated production is desirable for this task, as it enables new languages to be easily added in the future. To minimize translation errors we utilize both a large knowledge graph of terminology and an ensemble of commercial machine translation systems to produce an aligned concept list (fig. 6.5).

Details on producing the concept list

Source language term lists. We first produce a list of English nouns by collating words in term frequency lists extracted from TV closed captions and contemporary fiction novels from Wiktionary[4], and filter for the 2000 most frequent words in this

[4] `en.wiktionary.org/wiki/Wiktionary:Frequency_ lists/Contemporary_fiction, .../TV/2006/1-1000`

combined list, and augment it with class label names from CIFAR100 [243].

Finding good translations. We feed the list of English words into a custom translation pipeline, which simultaneously queries BabelNet [244], and an ensemble of four commercial translation systems: Google Translate, Bing Translate, Baidu Translate, and iTranslate[5].

In response to an English query, the BabelNet API returns a collection of "SynSets," subgraphs of a combined multilingual word and entity graphs centered on a node the query word maps to (see fig. 6.5 for examples). Each subgraph links to multiple other nodes, containing terms in both the source language and the target language. These edges can represent, for example, the titles of Wikipedia articles in different language editions of Wikipedia that are marked as being equivalent, thus ensuring that by checking against SynSet edges, a degree of human validation is included automatically. The synset also contains information about whether a given word is a noun. If it is not a noun, the candidate concept is discarded.

To choose the best translation from those edges, the returned translations into the target languages of the English term from the commercial translation services are *melded* by first sorting all returns by number of languages in the return query (in the case that one translation service covers more languages than others), and filling in missing translations by prioritizing alignment in the shared language translations. If any target language is missing a word for a concept at the conclusion of this process, that concept is discarded from the final list.

Post-filtering. Once a list of melded translations from the commercial service is returned, each row in the candidate aligned concept list is checked against the

[5]Using the `translators` PyPi package.

corresponding BabelNet SynSets to ensure each translation is present as a connected node, for pseudo-human evaluation. At the end of this process, a list of approximately 250 concepts is returned. Finally, we manually remove terms that are verb-noun collisions (e.g. hike) to ensure this ambiguity didn't drive any poor translations. The final list for ***CoCo-CroLa*** v1.0 contains 193 concepts.

6.4.2 Making minimal eliciting prompts

As discussed in §6.3, an ideal prompt template would enforce stylistic consistency in the generated outputs without introducing biases that interfere with the demonstration of concept possession. Following Bianchi et al. [152] we build simple prompts of the form, "a photograph of _____", which we manually translate into target languages. This simple template-filling approach will introduce grammatical errors for some languages. We briefly investigate if this matters in §6.4.2.

Validating the prompt templates

As mentioned in §6.4.2, the simple template-based approach to generating prompts for concepts leads to the introduction of grammatical errors, e.g. "a photograph of dog."

However, it is questionable whether small grammatical or logical errors like missing articles matters for high-resourced, well-covered languages like English. After all, the models are clearly able to generate high quality "photograph of dog" pictures without the word "a" in the sentence (Figure 1). But, does the prompt phrasing matter for lower-performance languages in a model? To investigate the impact of prompt phrasing on conceptual coverage, we tested a variety of English, Spanish and Chinese prompt phrasings on the concepts "dog," "sea," "airplane," and "ship" (selected for

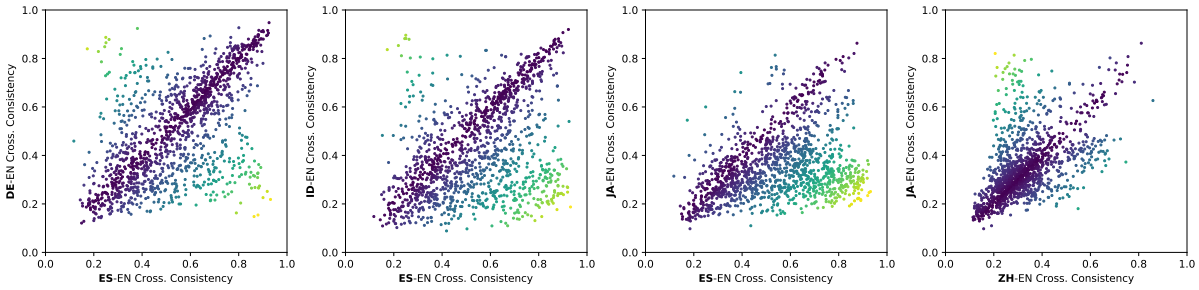


Figure 6.7: The correctness score for every (concept, model) pair for (right to left) ES vs DE, ES vs ID, ES vs JA, and ES vs JA. Languages sharing scripts (ES/DE/ID and JA/ZH) are more correlated than those that don't (ES/JA).

their wide distribution across the cross-correlation correctness metric range).

For the English prompts, we experimented with including the articles "a," "the," "my," and "an," as well as using the words "photograph," "image," "photo," and "picture." For Spanish, we used variations on the phrase "un foto de" (a photo of), including the same set of articles in English "un/una," "el/la," "mi," "tu," (your) and "nuestra/o" (our). For Chinese, we tried examples that both included and excluded the possessive particle "的" (de), as well as the words "照片" (zhaopian) and "图片" (tupian) for picture/photograph, and including or excluding the prepended phrase "一张" (yi zhang) to create the meaning "one photograph." We reran the full 193 concept image generations in those three languages for Stable Diffusion 2 and AltDiffusion.

We found limited impact across all of these dimensions. Full details available in our anonymous demo at `saxon.me/coco-crola`.

6.4.3 Applying the metrics for analysis

We assess Dt, Sc, Xc, and Wc for each (concept, language) pair for each model. Using these we compare models and assess the validity of conceptual coverage as a proxy for generalization.

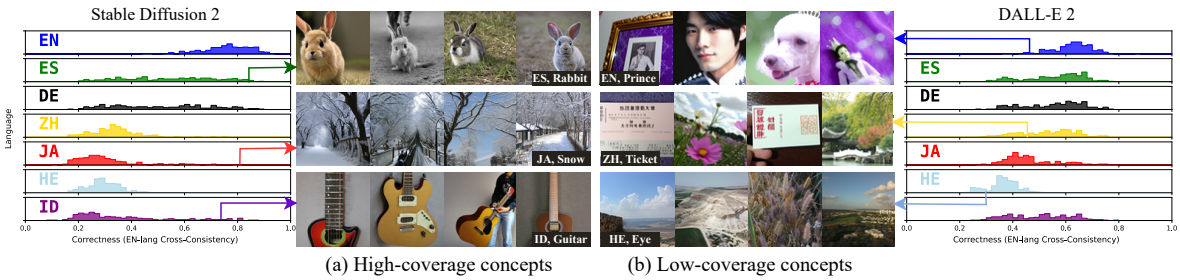


Figure 6.8: We automatically identify (a) high-coverage concepts in Stable Diffusion 2 (ES, rabbit), (JA, snow), (ID, guitar) and (b) low-coverage concepts in DALL-E 2 (EN, prince), (ZH, ticket), (HE, eye) using **correctness** Xc.

6.5 Findings

fig. 6.6 shows histograms for the distributions of the cross-consistency correctness proxy score Xc for each concept, relative to the training language (either English or Chinese) for DALL-E Mini, DALL-E 2, CogView 2, Stable Diffusion 1.4, Stable Diffusion 2, and AltDiffusion across the seven test languages. This plot clearly depicts that for the primarily English-trained models (DALL-E Mini, Stable Diffusion 1.4, Stable Diffusion 2), English-language performance (a high-mean distribution of high-EN-EN consistency concepts) is considerably better than the other languages. Similarly, for CogView2, trained on Chinese, the Chinese distribution of ZH-ZH scores is considerably better than the others, which do equally bad.

DALL-E 2 recieved open-ended multilingual training, and exhibits more consistent acceptable performance across the European and East Asian languages being tested. AltDiffusion, which has had its CLIP text encoder contrastively trained against multilingual representations on 9 languages (including ES, DE, ZH, and JA) exhibits higher performance on its training languages than its non-training languages (HE and ID).

Correctness distributions for Spanish, German, and Indonesian look roughly similar (in terms of mean and variance) for all models but AltDiffusion. This is

particularly interesting because they are the three non-English languages that also use the Latin alphabet. fig. 6.7 compares the **correctness** Xc score for every concept, in every model, across pairs of languages that fully or partially share scripts (ES, DE, ID), (ZH, JA) and two languages that don't (JA, ES). Across pairs of languages that share scripts, there is a high correlation between possession of a given concept in one language and the other. A consistent trend across all models was poor performance on Hebrew, which is both considerably lower-resource compared to the other six test languages, and uses its own unique writing system.

6.5.1 Correctness feature captures possession

fig. 6.8 shows how choosing samples of an image generated by a model, elicited by a high- or low-**correctness** score naturally reveals in which languages which concepts are possessed (e.g., Stable Diffusion 2 possesses ES:rabbit, JA:snow, and ID:guitar.) When a model possesses a concept, the outputted images are often visually similar with the tangible concept set in similar scenarios.

6.5.2 Types of concept non-possession

A model *not* possessing a concept can manifest in a few different scenarios depicted in fig. 6.8 (b). DALL-E 2 doesn't possess "prince" in English because it outputs a variety of different images, including human portrait photos, and pictures of photos, toys, and dogs. The existence of these *non-specific* error cases reflects the imperfect nature of our automated concept collection pipeline. Removing these poorly specified concepts is one way we plan to improve ***CoCo-CroLa*** in future releases.

A second type of possession failure we observe, we dub *specific collisions*. For example, Figure 1 and fig. 6.3 show JA collisions for the DALL-E mini/mega family.

	EN		ES		DE		ZH		JA		HE		ID		Avg	
Model	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc
DALL-E Mega	**81**	**28**	**65**	26	**64**	**26**	29	21	32	21	28	19	**51**	**25**	50	**24**
DALL-E Mini	78	27	59	25	50	23	33	21	31	21	34	**20**	49	24	48	23
SD 1.1	69	26	52	23	46	22	32	19	37	21	28	17	39	21	43	21
SD 1.2	71	26	48	23	44	22	28	19	35	21	24	17	37	21	41	21
SD 1.4	69	26	46	23	40	22	26	20	34	21	24	17	34	21	39	21
SD 2	76	27	54	24	51	24	34	19	31	21	29	17	37	21	45	22
CogView 2	37	20	42	20	39	20	**62**	**25**	40	21	**38**	**20**	42	20	43	21
DALL-E 2	61	27	55	**27**	54	**26**	44	**25**	42	22	36	19	42	**25**	48	24
AltDiffusion m9	64	26	59	25	49	22	55	**25**	**55**	**25**	**38**	**20**	43	22	**52**	23
Avg	67	26	53	24	49	23	38	22	38	21	31	18	42	22		

Table 6.2: **Correctness** scores (Xc and Wc) averaged for all concepts within a column language for all models. Note that Xc for CogView2 is relative to ZH rather than EN. AltDiffusion performs best in terms of total average Xc, and number of Xc or Wc column "wins." DALL-E Mega performs best on Latin languages and average Wc.

Both models consistently generate images of humans for "dog" but pictures of green landscape scenes for "airplane." While these generated concepts are incorrect, they represent an incorrect mapping to a different concept rather than a mere lack of conceptual possession. We also observe cases where specific collisions only occur part of the time, such as in the case of DALL-E 2 and ZH:ticket (fig. 6.8 (b)).

Finally, we observed cases of *generic collisions*. For example, DALL-E 2 consistently generates images of desert or Mediterranean scenery when prompted with "eye" in Hebrew (fig. 6.8 (b)). This pattern shows up across a diverse set of models and prompts. Figure 1 shows how across "dog," "airplane," and "face," DALL-E mega, Stable Diffusion 2, and DALL-E 2 seem to generate vaguely-Israel-looking outdoor scenes regardless of eliciting concept. This is probably reflective of a small sample-size bias in the training data.

6.5.3 Model comparison

table 6.2 shows the the use of correctness scores in the ***CoCo-CroLa*** benchmark to compare the 9 models. As expected, given its multilingual training regimen, AltDiffusion m9 outperforms the other T2I models on average, and in terms of total wins. It is particularly strong relative to the other models in Japanese and Chinese (with the exception of the Chinese-only CogView 2, which is best on Chinese but worst on average overall for both Xc and Wc).

However, despite the strong average performance of AltDiffusion, there's a lot of room for improvement. For example, its improvements in terms of JA and HE performance come at a cost of significantly reduced EN and DE performance relative to Stable Diffusion 2, its initialization checkpoint. The ***CoCo-CroLa*** benchmark can guide future work in adapting T2I models to further multilinguality without losing conceptual coverage on source languages.

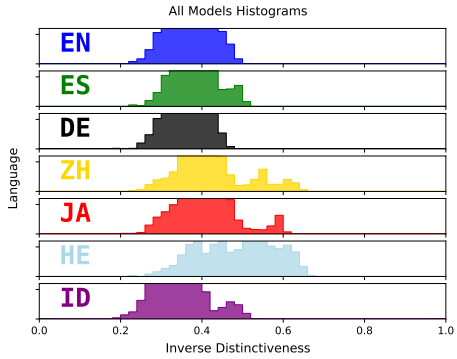
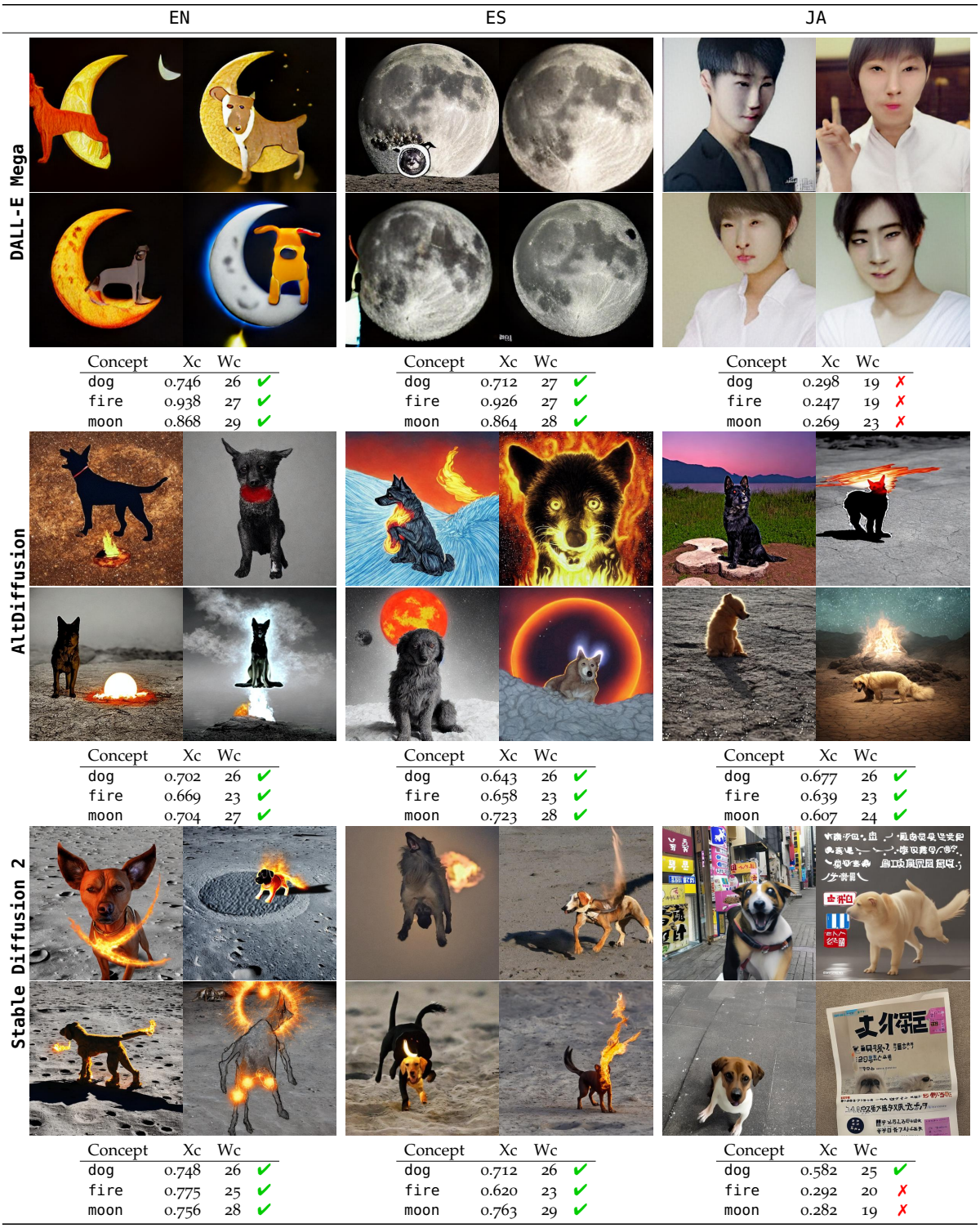


Figure 6.9: Histograms of the **inverse distinctiveness** scores for all models and all concepts.

6.5.4 Distinctiveness captures generic collisions

fig. 6.9 shows the distribution of the **inverse distinctiveness** score Dt. On this plot, more rightward probability mass indicates a distribution of concepts for which

Figure 6.10: "a bird using a keyboard in the snow"



Concept	Xc	Wc	
dog	0.746	26	✔
fire	0.938	27	✔
moon	0.868	29	✔

Concept	Xc	Wc	
dog	0.712	27	✔
fire	0.926	27	✔
moon	0.864	28	✔

Concept	Xc	Wc	
dog	0.298	19	✗
fire	0.247	19	✗
moon	0.269	23	✗

Concept	Xc	Wc	
dog	0.702	26	✔
fire	0.669	23	✔
moon	0.704	27	✔

Concept	Xc	Wc	
dog	0.643	26	✔
fire	0.658	23	✔
moon	0.723	28	✔

Concept	Xc	Wc	
dog	0.677	26	✔
fire	0.639	23	✔
moon	0.607	24	✔

Concept	Xc	Wc	
dog	0.748	26	✔
fire	0.775	25	✔
moon	0.756	28	✔

Concept	Xc	Wc	
dog	0.712	26	✔
fire	0.620	23	✔
moon	0.763	29	✔

Concept	Xc	Wc	
dog	0.582	25	✔
fire	0.292	20	✗
moon	0.282	19	✗

Figure 6.11: "a dog made of fire standing on the moon"

distinctiveness is **low** relative to a generic sample of images produced by a given model in that language. The four Latin script languages (EN, ES, DE, ID) exhibit the lowest inverse distinctiveness, and are thus the least prone to producing generic failure images. Hebrew is an outlier for having high concept-level Dt. Across many models. This is probably because it is the most low-resource language in our list by far, and doesn't benefit from script sharing.

6.5.5 Ranking concepts by Xc

For a given model and language, ***CoCo-CroLa*** can be used as a concept-level analysis tool. For example, by performing the same ranking over a specific (model, language) pair, we can find the most well-covered and poorly-covered concepts for that pair. For all models and languages, an interactive ranking demo based on ascending and descending Xc and Wc is available at `saxon.me/coco-crola/`. For example, we found "snow" to be a concept possessed in EN and ES for DALL-E Mega, AltDiffusion, and Stable Diffusion, but only possessed in JA by Stable Diffusion 2. A similar situation holds for "dog" and "fire," with AltDiffusion.

6.5.6 Concept possession as a proxy

In this section we will discuss how **an inability for a model to use some concept in complex, creative prompts is implied by our detected non-possession of said concept,** thereby validating the ***CoCo-CroLa*** atomic evaluation paradigm.

To investigate this we manually translated two creative prompts including concepts found to be differentially present in DALL-E Mega, AltDiffusion, and Stable Diffusion §6.5.5 from English into Spanish and Japanese. The prompts were: "a bird using a keyboard in the snow," (ES: "un pájaro usando un teclado en la nieve," JA: "雪にキー

ボードを使っている鳥") and "a dog made of fire standing on the moon," (ES: "un perro hecho de fuego pisando en la luna," JA: "火でできた犬が月に立っている").

fig. 6.11 clearly shows that, using thresholds for non-possession of Xc < 0.5 and Wc < 25, **if a concept is not possessed by a model according to *CoCo-CroLa*, it will be unable to successfully generate creative images containing it**.

However, other capabilities including compositionality and perhaps a sort of *verb-level conceptual possession* are probably also required in order to make the converse (possession implies capability to generate creatively) to be true—yet we have no method to capture such possession. This is a promising direction for future work.

Thus, concept-level coverage in a model can be used as a proxy for generalization capabilities to more complex prompts containing the concept, at least in the case of tangible noun concepts. This is good news, as it **enables assessment of the infinite space of creative prompts from a finite, constrained set of atomic concepts.**

6.6 Conclusion

Multilingual analysis of text-to-image models is desirable both to improve the multicultural accessibility of T2I systems and to deepen our understanding of their semantic capabilities and weaknesses. Analyzing a model's *conceptual coverage* is a simple and straightforward way to do this.

We demonstrated that these concepts are core building blocks for producing impressive images, and that analyzing the concepts is a useful proxy for assessing the creative capabilities of T2I models.

Our technique, ***CoCo-CroLa*** is a first step toward further work in this domain. Following our recipe, larger benchmarks containing more languages and concepts can be built easily.

Limitations

The ***CoCo-CroLa*** benchmark generating procedure is intended to yield multilingual evaluations that can be scaled to even larger sets of concepts and languages without experienced annotators. In the interests of both concept and language quantity scale, we opted for an automated procedure which leverages machine translation systems, which can introduce translation errors. Furthermore, variation in the nuance or normative meaning of concepts, particularly culturally contested ones such as "face," [245] "person," or "man" will inevitably drive some variance in expected outputs by users across language communities. This cultural variation will place an unavoidable upper bound on the performance of inherently cross-cultural benchmarks such as ***CoCo-CroLa***.

Additionally, typological variation between languages can introduce complications in applying our framework. For example, while simple template filling for prompting is straightforward in Chinese, which requires no word-dependent articles, in English phonological properties of the word govern the preceding article, and in Spanish and German grammatical gender do the same. Hebrew has gendered nouns, adjectives, and verbs but not articles, on the other hand. Overall, it appears that these have limited influence as grammaticality isn't a crucial feature in the prediction of image tokens performed in T2I models, §6.4.2.

Finally, using the ensembled commercial translation systems for single-word translations in ***CoCo-CroLa*** turns out to be suboptimal; it can introduce ambiguity and word sense errors that challenge the construct validity of some evaluations. I will discuss the severity of this issue and how to mitigate it in **Chapter 7**.

Chapter 7

Problems in multilingual evaluation

The dynamic synthesis pipeline ***CoCo-CroLa*** employs has many advantages. It can be scaled to any (high-resource) language where frequency list, knowledge graph, or machine translation services are available, and it meets the **dynamic** desideratum from **Part I**. Unfortunately it has an Achilles' heel: it relies on single-word calls to machine translation services, which may introduce translation errors. While the "melding" step using BabelNet was intended to disambiguate each language's translation candidates and mitigate word sense errors, the consensus mechanism leveraged may (and as we will soon see, did) actually *propagate these sense errors across languages*.

For example, the word *rock* has two common noun senses: *a stone* or *a music genre*. In the case of *rock* it turns out that the ***CoCo-CroLa*** pipeline selects the music sense for most languages. This then renders the correctness scores for *rock* uninformative.

In this chapter we explore the impact of these kinds of translation errors on ***CoCo-CroLa*** using some manual assessment. We will see that *in the text domain alone* we can actually predict the severity of a mistranslation error on the construct validity of ***CoCo-CroLa*** for that concept *in the image domain*—without actually generating any images—by leveraging multilingual joint embedding spaces.

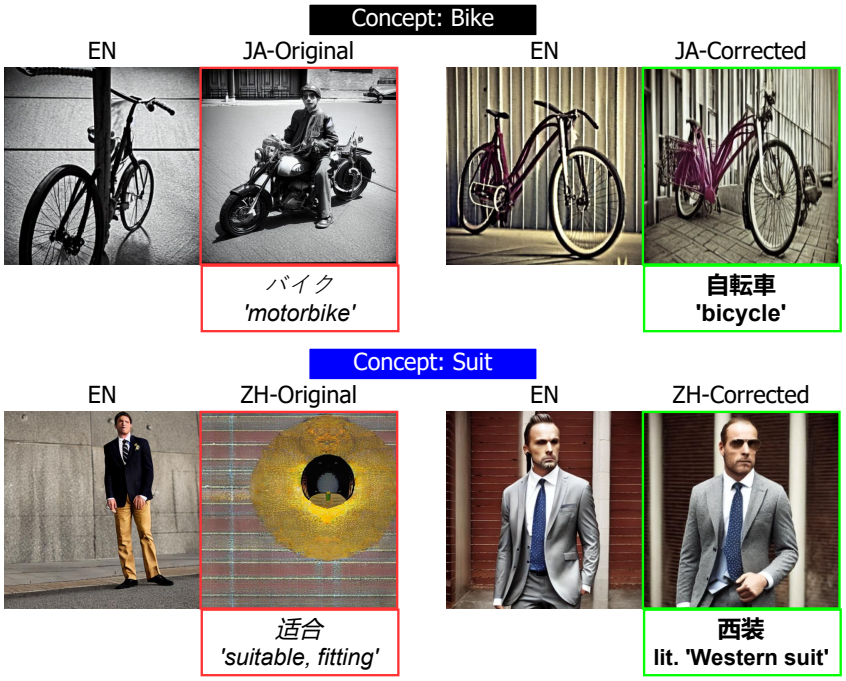


Figure 7.1: The CoCo-CroLa benchmark mistranslated concepts such as *bike* in JA and *suit* in ZH. With correct translations (right) AltDiffusion does in fact "possess" them; originally (left) they were false negatives.

7.1 Introduction

With growth in the popularity of generative text-to-image (T2I) models has come interest in assessing their capabilities across many dimensions, including multilingual accessibility. The ***CoCo-CroLa*** [150] benchmark attempts to capture how well "concept-level knowledge" within a T2I model is accessible across different input languages. It compares the output image populations of a system under test when prompted to generate images of 193 tangible concepts in 7 *test languages* to the images generated from a semantically equivalent prompt in a *source language*. It and similar benchmarks rely on correct translations for validity, lest "possessed" concepts be mistakenly assigned false negatives.

We find a strict *error candidate rate* of 4.7% for Spanish (ES), 8.8% for Chinese

Concept	Language	Original	Corrected	Reason for Correction
Rock	Japanese	ロック	岩	ロック, *rokku*, refers principally to "rock music" instead of stones in nature.
Flame	Spanish	*llama*	*flama*	*Llama*, though a correct translation for "flame," coincides with the animal in English.
Ground	Japanese	接地	地面	接地 refers to an electrical ground rather than the surface of the earth.
Table	Chinese	表	桌子	表 means a tabular form or a spreadsheet, not a four-legged furniture.
Milk	Japanese	乳	牛乳	乳 may mean breast or any kind of milk. 牛乳 means the milk produced by cows.
Tent	Spanish	*tienda*	*...de acampar*	*Tienda* alone more often means "store," *tienda de acampar* specifies (camping) tent.
Teacher	Japanese	先生	教師	先生 is a common title to address an educated person, e.g., teacher, doctor, lawyer.
Father	Chinese	爸爸	父亲	爸爸 is the colloquial addressing equivalent to 'daddy'. 父亲 is more formal.

Table 7.1: Example error candidates from the CoCo-CroLa benchmark in Japanese, Chinese, and Spanish.

(ZH), and 12.9% for Japanese (JA) in the ***CoCo-CroLa*** v1 (CCCL) concept translations through manual analysis by fluent speakers. These error candidates are not filtered by severity. While some candidates are severe translation errors that drive false negatives (fig. 7.1), others are marginal annotator disagreements that might not matter (table 7.1). In this work, **we investigate when and why translation changes actually impact CCCL results** to improve future T2I multilinguality benchmarks. We:

1. Write *candidate corrections* for CCCL in Spanish, Japanese, and Chinese, evaluated on four T2I models.

2. Introduce a text-domain comparison metric ΔSEM to predict correction significance.

3. Analyze our candidates by ΔSEM and image correctness improvement and apply impactful ones to CCCL as v1.1.

4. Report insights and considerations for future semantic T2I evaluations we uncovered.

7.2 Motivation & Approach

The **C**o**C**o-**C**ro**L**a benchmark (CCCL) evaluates a T2I model's ability to generate images of an inventory of tangible concepts when prompted in different languages

[150]. Given a tangible concept c, written in language ℓ as phrase c_ℓ, the i-th image produced by a multilingual T2I model f on the concept c_ℓ can be expressed as:

$$I_{c_\ell,i} \sim f(c_\ell) \tag{7.1}$$

The images generated in language ℓ are considered *correct* if they are faithful to their equivalent counterparts in the source language ℓ_s. This is measured by the CCCL benchmark by a **correctness metric** for a single concept c as the *cross-consistency* score $X_c(f, c_\ell, c_{\ell_s})$:

$$X_c = \frac{1}{n^2} \sum_{i=0}^{n} \sum_{j=0}^{n} \text{SIM}_F(I_{c_\ell,i}, I_{c_{\ell_s},j}) \tag{7.2}$$

where we sample n images per-concept per-language (we use 9), and $\text{SIM}_F(\cdot,\cdot)$ measures the cosine similarity in feature space by image feature extractor F. In practice, the default source language ℓ_s is English and F is the CLIP visual feature extractor [162].

7.2.1 Translation Errors in *CoCo-CroLa*

CCCL requires correct translations of each concept c from the source language ℓ_s into a set of semantically-equivalent translations in each test language ℓ. **Chapter 6** [150] built CCCL v1's concept translation list using an automated approach so as to allow new languages to be easily added without experts in each new language. Unfortunately, this approach introduces translation errors (table 7.1).

Here, we check the Spanish, Chinese, and Japanese translations using a group of proficient speakers who identify a set of *translation error candidates* that may not sufficiently capture a concept's intended semantics in English, for various reasons.

Some of the candidate errors, such as the error for *rock* in Japanese (table 7.1),

represent severe failures to translate a concept into its common, tangible sense—it is incoherent to test a model's ability to generate pictures of rocks by prompting it with "rock music." However, other candidate errors, such as *father* in Chinese are still potentially acceptable translations, but deviate from the annotators' preferred level of formality or specificity.

To decide which corrections ought to be integrated in future T2I multilinguality benchmarks, quantifying both the significance of each translation correction is and its impact on the CCCL score for its concept is desirable.

7.2.2 Quantifying Error Correction & Impact

Characterizing the *impact* of a translation correction on model behavior is simple; we check ΔX_c, the change in the CCCL score going from the original concept translation c_ℓ to the corrected c'_ℓ,

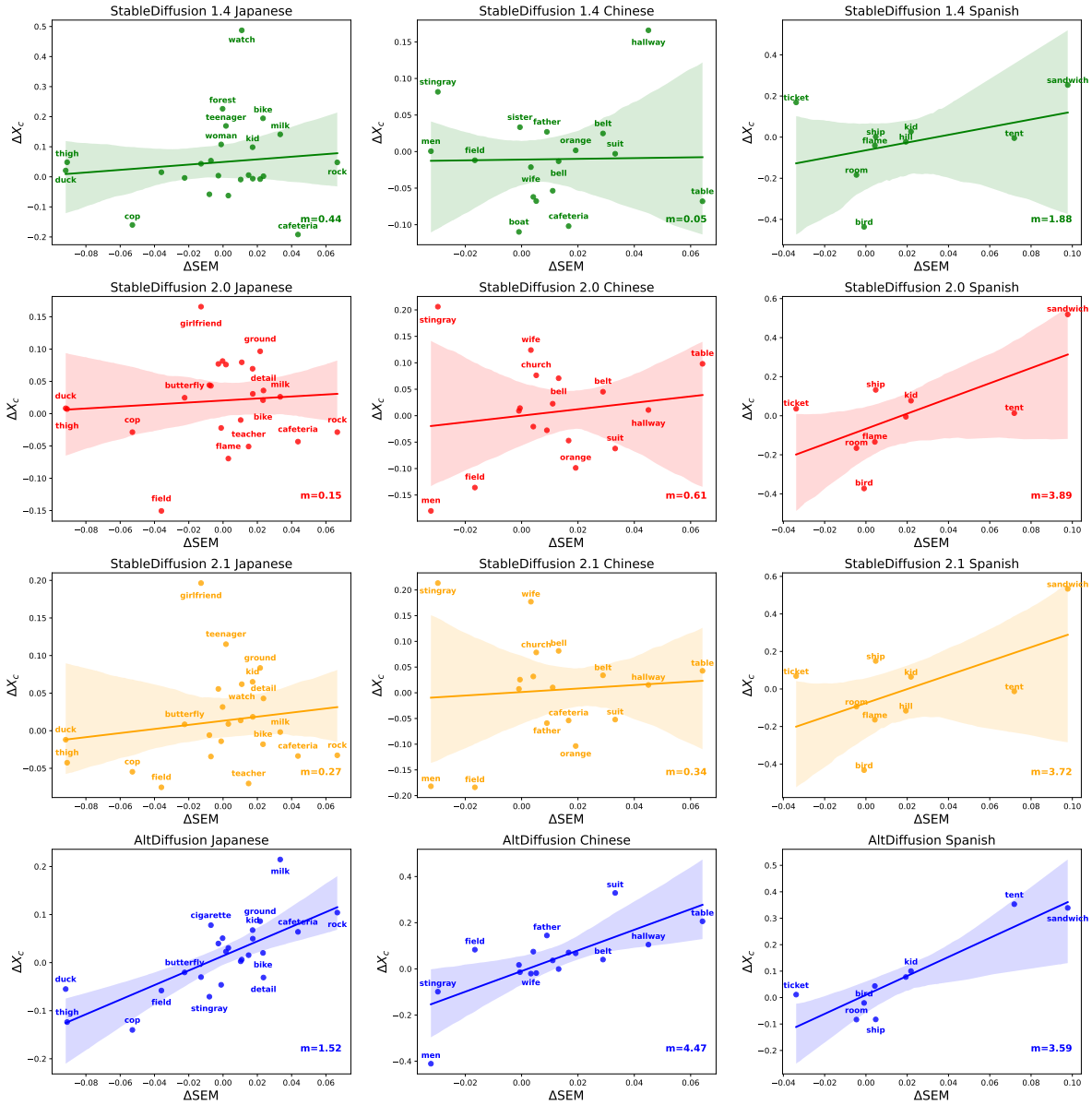
$$\Delta X_c(c, \ell) = X_c(f, c'_\ell, c_{\ell_s}) - X_c(f, c_\ell, c_{\ell_s}) \tag{7.3}$$

by comparing the generated population of images elicited from the corrected term $I_{c'_\ell}$ to the candidate translation error-conditioned images I_{c_ℓ}.

We quantify the significance of the translation correction as the *improvement in semantic similarity* $\Delta\text{SEM}(c_{\ell_s}, c_\ell, c'_\ell)$ using a text feature extractor F_t and cosine similarity metric $\text{SIM}(\cdot, \cdot)$

$$\Delta\text{SEM} = \text{SIM}_{Ft}(c_{\ell_s}, c'_\ell) - \text{SIM}_{Ft}(c_{\ell_s}, c_\ell) \tag{7.4}$$

We use embeddings from the multilingual SentenceBERT [246] text embedder `OpenAI CLIP-ViT-B32` model as F_t.



(a) Japanese (b) Chinese (c) Spanish

Figure 7.2: Scatterplots showing the impact of the corrections to each concept in JA, ZH, and ES on the conceptwise improvement to the CCCL correctness score, ΔX_c, as a function of ΔSEM. Slopes m at bottom-right in **bold**.

7.3 Results & Analyses

We generate output images using StableDiffusion 1.4, 2.0, 2.1 [126] and AltDiffusion [241], for all concepts corrected by our annotators in English, Spanish, Chinese, and Japanese, using both the original concept translations c_ℓ from ***CoCo-CroLa*** v1 [150]

and the corrected translations c'_ℓ. Model details are provided in Appendix 7.5.2.

fig. 7.2 shows the relationship between ΔSEM and ΔX_c for all corrected concepts for StableDiffusion 1.4, 2.0, 2.1, and AltDiffusion[1]. Note the pronounced, significant positive slope of the correlations between the two variables for AltDiffusion in all languages (4th row) and in Spanish for all models (third column). Here a positive slope means that *higher-improvement* translation corrections (assessed by increased proximity to the English word in a shared embedding space) reliably correct the generated images more than the modest candidates.

These same high-slope model/language pairs (eg., JA & AltDiffusion) were found by [150] to be "well-possessed" (high average X_c across correct concepts) in ***CoCo-CroLa*** v1. In other words, *valid corrections only matter for languages a model already "knows."* Correct Klingon is just as useless as incorrect Klingon to a non-Klingon model.

Table 7.3 (§7.5.3) shows the same slopes m with PCCs, p-values, and intercepts for the each model and language's ΔSEM to ΔX_c relationship. The high-slope language/model pairs also tend to have higher PCC with more statistical significance.

StableDiffusion 1.4 was trained on the primarily-Latin script `LAION-en-2b` [224], and thus lacks capabilities in non-Latin script languages JA and ZH. Consequently, there is no significant relationship between more semantically divergent corrections with high ΔSEM and larger improvements to concept correctness ΔX_c for SD 1.4 on those languages. Meanwhile, AltDiffusion—which conditions output images on the multilingual XLM-Roberta encoder [106]–benefits from all significant corrections in all languages with statistical significance.

[1]Error margins are 95% regression-fit confidence intervals.

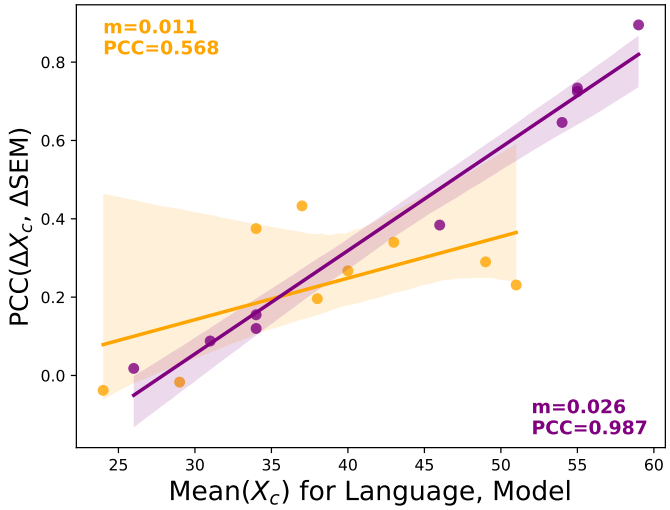


Figure 7.3: Languages with a high correlation between textual correction significance and image improvement (PCC) are more "well-understood" by the model (X_c), for both **real**- and **pseudo**-corrections.

7.3.1 Pseudocorrection Experiment

Unfortunately our ability to use the aforementioned corrections to confirm our hypothesis that *T2I model language capability can be estimated from the impact of translation corrections on image-domain performance* is hindered by the small quantity of correction candidates we found. We bypass this problem with a *pseudocorrection experiment*—simulating a larger set of corrections by generating artificial errors in the other CCCL languages. We generate 10 synthetic erroneous *pseudo-original translations* for each concept in German, Indonesian, and Hebrew by randomly sampling the translations for other concepts within-language. Each concept's "correction" is its original translation.

For example, we assign the concept *eye* the Indonesian word *guru* (EN:teacher) as its pseudo-original. We then "correct" this word to *mata*, the original correct translation, and assess ΔX_c and ΔSEM with c_{ℓ_s}:*eye*, c_ℓ:*guru* and c'_ℓ:*mata*.

This gives us 1,930 ΔX_c, ΔSEM pairs for each language and model, with which

we evaluate the same correlation relationship as before (plot in Appendix fig. 7.7). We report Pearson's correlation coefficient (PCC) for each of these pairs along with the average CCCL X_c reported in [150] in fig. 7.3. The same relationship for real corrections holds for pseudocorrections, demonstrating that text-only multilingual semantic similarity features can predict the impact of a translation correction on the output image correctness.

7.4 Discussion & Conclusions

Our findings motivate important considerations for building future T2I semantic evaluations [127, 157, 149].

Subjectivity A reliable T2I multilinguality assessment must report true knowledge failures—examples where a model fails to generate correct images of a concept, when it is correctly prompted to do so. Correct translations are required.

Unfortunately, choosing one "correct translation" is in inherently subjective task. This study tackled this subjectivity by casting a wide net of error candidates, and testing their impact. Consequential errors caused *false negatives* where a concept is erroneously marked as "not possessed" (fig. 7.1).

CCCL's tangible concept constraint and corpus-based approach to finding concepts helps combat subjectivity [150]. In the tangible sense it's fair to say "orange" is correctly translated in Spanish to *naranja* (the fruit) rather than *anaranjado* (the adjective).

In prompting the T2I model we assume this tangible noun context is induced by using "a picture of an X"-style prompts. While our results show this works, it is a model-specific phenomenon and future work should examine more prompt templates.

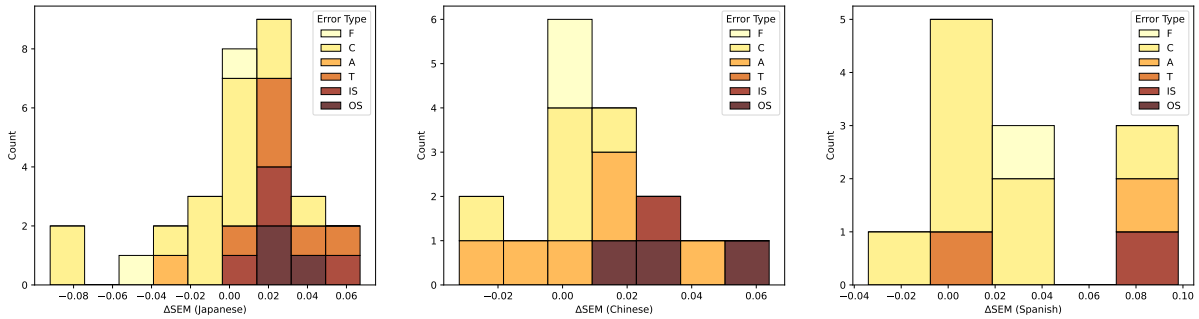


Figure 7.4: Histograms for the error counts in JA, ZH, and ES vs Δ_{SEM}, colored by error type. From lightest, they are F:*formality*, C:*commonality*, A:*ambiguity*, T:*transliteration*, IS:*incoming sense error*, OS:*outgoing sense error*. The error types are defined in §7.5.1. Severe error types will exhibit more rightward distributional mass.

Future work grounded in prototype theory [247] may enable identification of culturally universal concepts for assessment.

Need to assess Multiple Translations One challenge in multilinguality assessments is *incoming duplicates*, where multiple ways of writing a translation really are equally correct. Our homograph errors have examples, such as cigarette in Japanese. たばこ, タバコ, and 煙草 are all translations of cigarette with identical reading, *tabako*. Why should a metric of model-language capabilities only assess one correct translation rather than all?

More significant multiple translation problems arise in languages with gendered human-referent terms. For example, in Spanish *maestro* refers to a male teacher, while *maestra* a female one. Should a test of a model's Spanish knowledge of "teacher" as a concept test that both translations work equally well? CCCL v1 is incapable of assessing these attributes. Future benchmarks should contain this flexibility, so multiple incoming translations [248] can be assessed for the same concept, while also tracing semantically-encoded secondary attributes such as gender between the source and test language.

Error Severity and Error Type fig. 7.4 shows the distributions of error types for each language with respect to Δ_{SEM}, our proxy for correction significance or error severity. Across all three languages, the *sense errors* (OS and IS) are the most severe, while the formality and commonality errors are the least severe (defined in §7.5.1).

Our original estimated error rate (sum of all candidates per language) is a worst-case bound, the significant-to-evaluation-validity error rate is lower. Our impact and significance results show that some of our suggestions (mainly formality and commonality errors) may be more nitpick than correction.

Some concepts in CCCL are inherently erroneous due to intangibility. For example, *history*, *film*, and *jump* are all present in v1 of CCCL, picked up for being high-frequency noun concepts across multiple languages in the corpora. There is no sensible prototypical way to generate images "of" those concepts. We removed these for CCCL v1.1; Future benchmarks should avoid including them.

Image-Image Metric Blind Spots We observed interesting borderline (potential false positive) cases where ***CoCo-CroLa*** scored mistranslated concepts as possessed. For example, *bike* in Japanese. fig. 7.1 shows that under the erroneous translation, AltDiffusion generates pictures of *motorcycles* rather than *bicycles* as it does in English. However, X_c doesn't actually change much under this correction as shown in fig. 7.2 & table 7.4. The CLIP similarity score in CCCL is blind to the difference between a bicycle and motorcycle. Mistranslations where visual structural similarity is present are sometimes invisible to the image metrics.

Tangible object translation as an MT domain Single word concepts are not central to the distribution of machine translation training data. By providing the individual English tangible nouns as input we may expect an unreasonable amount of implicit

commonsense reasoning from commercial MT systems—the correct sense out of many had to be selected for success. Furthermore, the use of the BabelNet knowledge graph as a consensus mechanism reinforced some sense errors. For example, the *rock* sense error for JA (music genre rather than physical object, table 7.4) was also present in Hebrew, probably due to shared edges in the knowledge graph. Given previous interest in assessing the performance of MT translation in diverse domains [249], we think both the word-level translation of concepts under domain constraints without context (as we tried to do in CCCL previously) and treating input prompts for T2I systems (ie, captions) [250, 251] as a target domain for MT evaluation would be interesting and useful future directions.

Future benchmarks should leverage context with sentences as input to MT (eg, "watch for falling rocks") rather than the decontextualized concept words alone to improve robustness. LLMs could generate diverse English sentence examples, and could potentially also extract the final concept translations out of the multiple sentence translations.

Limitations

Trivially, human annotators for every language would remove false-negative mistranslations from future benchmarks, but there's a trade-off between easy scalability and certainty of correctness.

Our work incorporates human efforts of both native and proficient but non-native language speakers to propose and resolve translation error candidates caused by the machine translation pipeline in the original ***CoCo-CroLa*** benchmark. This could potentially bring human biases into the nuance of factors such as words' choices, introducing less culturally neural expressions as a result.

The assumption of *translatability* that underlies CCCL in general is a challenge. As a practical use-based test of functional fairness, using heuristics and only common everyday objects that can be reasonably assumed universal is acceptable, but more linguistic and even philosophical work is needed to really motivate fairness across languages and cultures when underlying assumptions differ.

7.5 Supplementary materials

7.5.1 Error candidate typology

Commonality (C). When a selected translated term doesn't appear to reflect the most common, colloquial, contemporary, or "natural" way that native speakers of the language would use in reference to the concept in a photograph or conversation. For example, in Chinese "瓶子" is a more conversational and contemporary way of writing *bottle* than "瓶," which reads literary and archaic.

Outgoing Sense Error. (OS) The translated term picks an alternative (and often less tangible) sense from the source concept. For example, the original Chinese translation for *Table* diverges to the sense of 'spreadsheet, tabular', instead of the presumptive home furniture item.

Incoming Sense Error. (IS) The translated term, while aligned to the correct source concept sense, picks a phrasing for which other senses in the target language exist that the annotators expect will confound model behavior, where another (often more common) disambiguated translation also exists. For example, the original Spanish translation for *tent* is given as *tienda* alone, which can also mean 'store, shop', in addition to 'a tent,' whereas the corrected translation *tienda de acampar* refers to a camping tent alone.

Ambiguity (A). The translated term introduces a word with multiple meanings from the unambiguous source concept. For example, the Japanese translation for *Milk* originally uses a single character that can mean any kind of animal or human milk, or even the organ of the breast.

Formality. (F) The translated term uses an expression in an improper formality. For example, the original Chinese translation for *Father* is only heard in casual conversations.

Transliteration (T). When one of the above errors occurs with . For example, the transliteration of *Rock* in Japanese is commonly related to 'Rock Music', rather than stones found in nature.

7.5.2 Computational Experiments Details

Models Employed See table 7.2.

Model	# Param	Repository (`huggingface.co/...`)
StableDiffusion 1.4	860M	`CompVis/stable-diffusion-v1-4`
StableDiffusion 2	NA	`stabilityai/stable-diffusion-2`
StableDiffusion 2.1	NA	`stabilityai/stable-diffusion-2`
AltDiffusion m9	1.7B	`BAAI/AltDiffusion-m9`

Table 7.2: The set of text-to-image models we evaluated with (Table adapted from [150].

Experimental Setup We generated 9 images for each (language, model, concept) triple and evaluated X_C using identical methods and codeas described in CCCL [150].

7.5.3 Full Analysis Numbers

See table 7.3.

Model	Language	PCC	p	m	b
SD1-4	Japanese	0.120	0.577	0.437	0.049
SD1-4	Chinese	0.018	0.944	0.051	-0.011
SD1-4	Spanish	0.384	0.307	1.877	-0.064
SD2	Japanese	0.088	0.684	0.155	0.020
SD2	Chinese	0.155	0.554	0.608	0.000
SD2	Spanish	0.646	0.060	3.891	-0.067
AD	Japanese	0.734	0.000	1.519	0.014
AD	Chinese	0.725	0.001	4.472	-0.010
AD	Spanish	0.895	0.001	3.588	0.010
SD2-1	Japanese	0.162	0.448	0.272	0.013
SD2-1	Chinese	0.078	0.765	0.340	0.001
SD2-1	Spanish	0.574	0.106	3.722	-0.075

Table 7.3: Stats for Pearson correlation and linear best fit between ΔSEM and ΔX_c for each model and language. p represents the p-value for the PCC, m and b the slope and intercept for the best-fit line.

7.5.4 Further Related Work

ConceptBed [252] evaluates monolingual concept-level knowledge in T2I, and its concept inventory could extend and improve CCCL's. T2I-CompBench assesses compositionality in T2I [149], leveraging VQA and image segmentation. Assessment model weaknesses, such as [253]'s VQA spurious correlations [254] remain a challenge.

Other benchmarks in vision-and-language also require correction and improvement. Others found and filtered unsolvable cases in *Who's Waldo* [255]. [256] exploit repeated texts in QA pairs on *VCR* [257]. While manual techniques can find and clean these errors, automated approaches would be preferable, such as the PECO method [69] for finding model-used shortcuts in NLI. Semi-human-in-the-loop approaches [37] may improve the sourcing and cleaning of future CCCL versions.

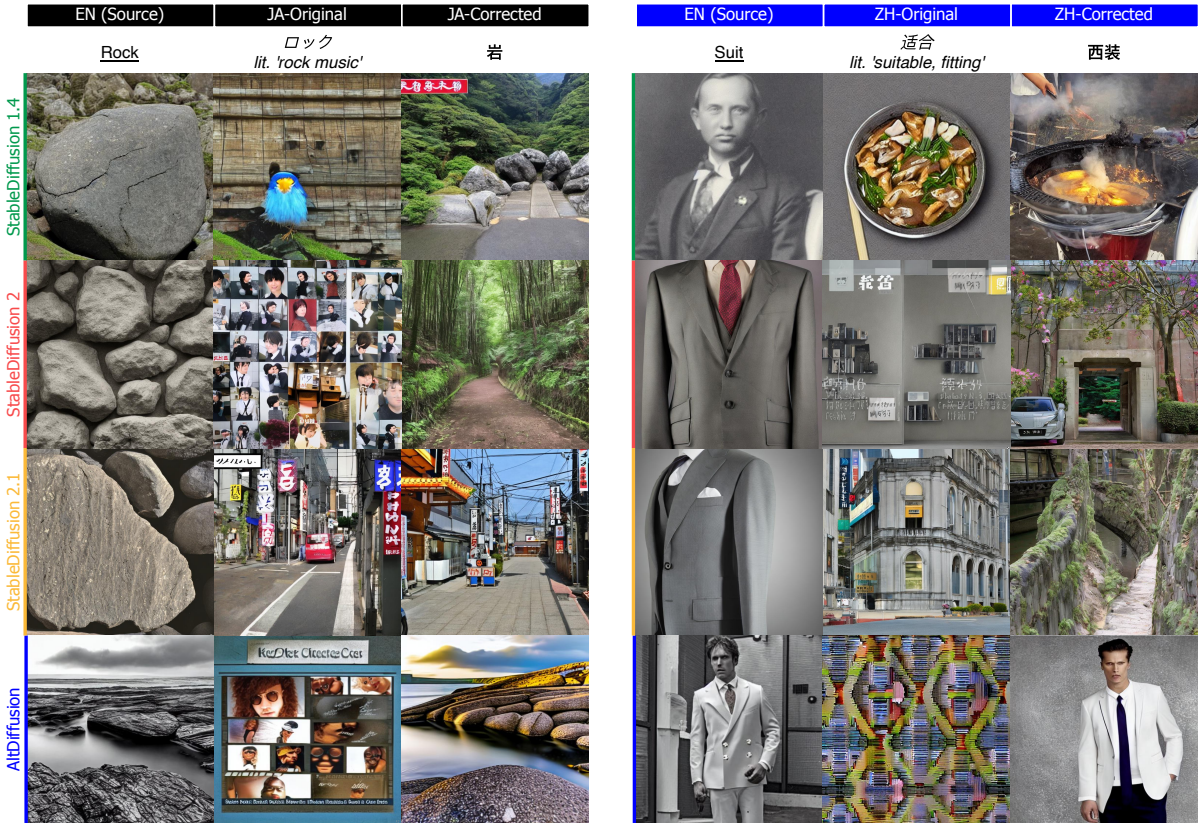


Figure 7.5: Qualitative examples of selected mistranslated concepts found in ***Co-Co-CroLa*** generated by AltDiffusion and multiple versions of Stable Diffusion - **Left:** "Rock" in Japanese, **Right:** "Suit" in Chinese. Noticeably, we observe that T2I models such as Stable Diffusion 2 do not benefit from correcting the translations, as their outputs in the aforementioned languages remain irrelevant similarly to using random prompts.

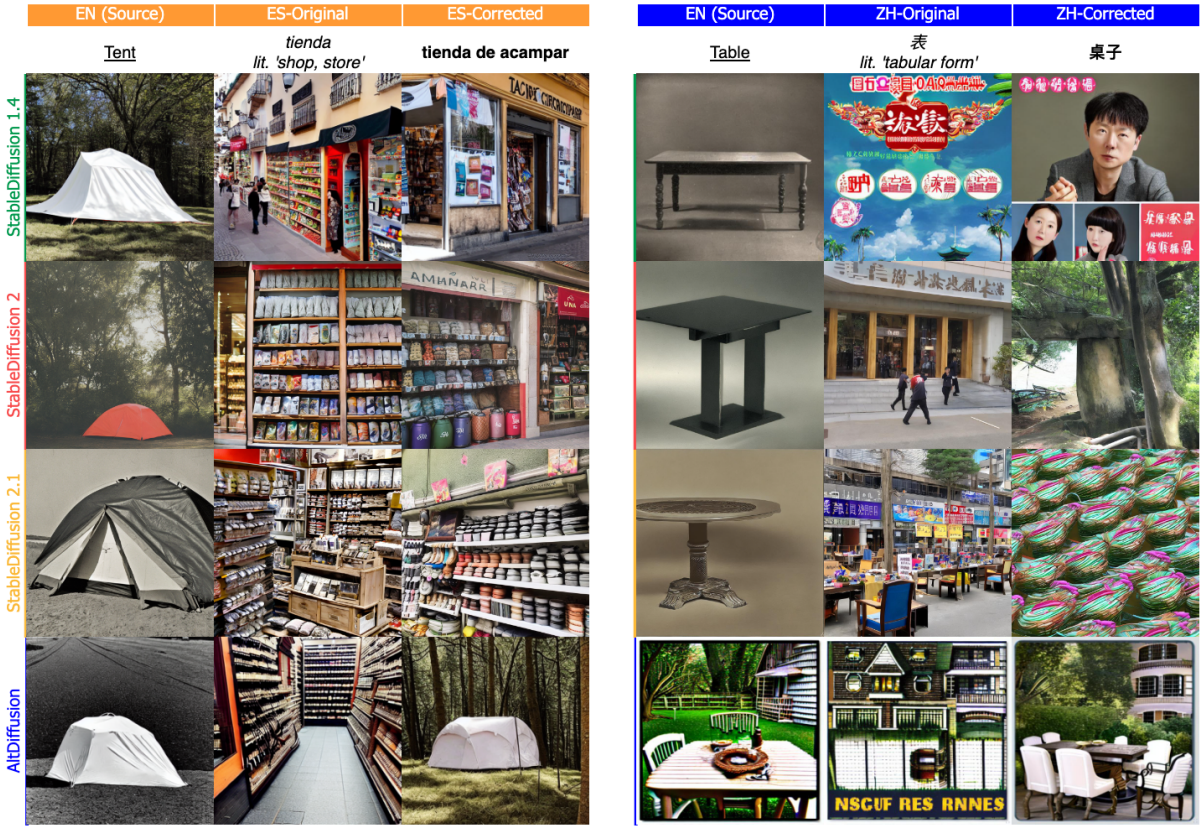


Figure 7.6: Further qualitative examples of selected mistranslated concepts found in ***CoCo-CroLa*** generated by AltDiffusion and multiple versions of Stable Diffusion - **Left:** "Tent" in Spanish, **Right:** "Table" in Chinese. Noticeably, we observe that T2I models such as Stable Diffusion 2 do not benefit from correcting the translations, as their outputs in the aforementioned languages remain irrelevant similarly to using random prompts.

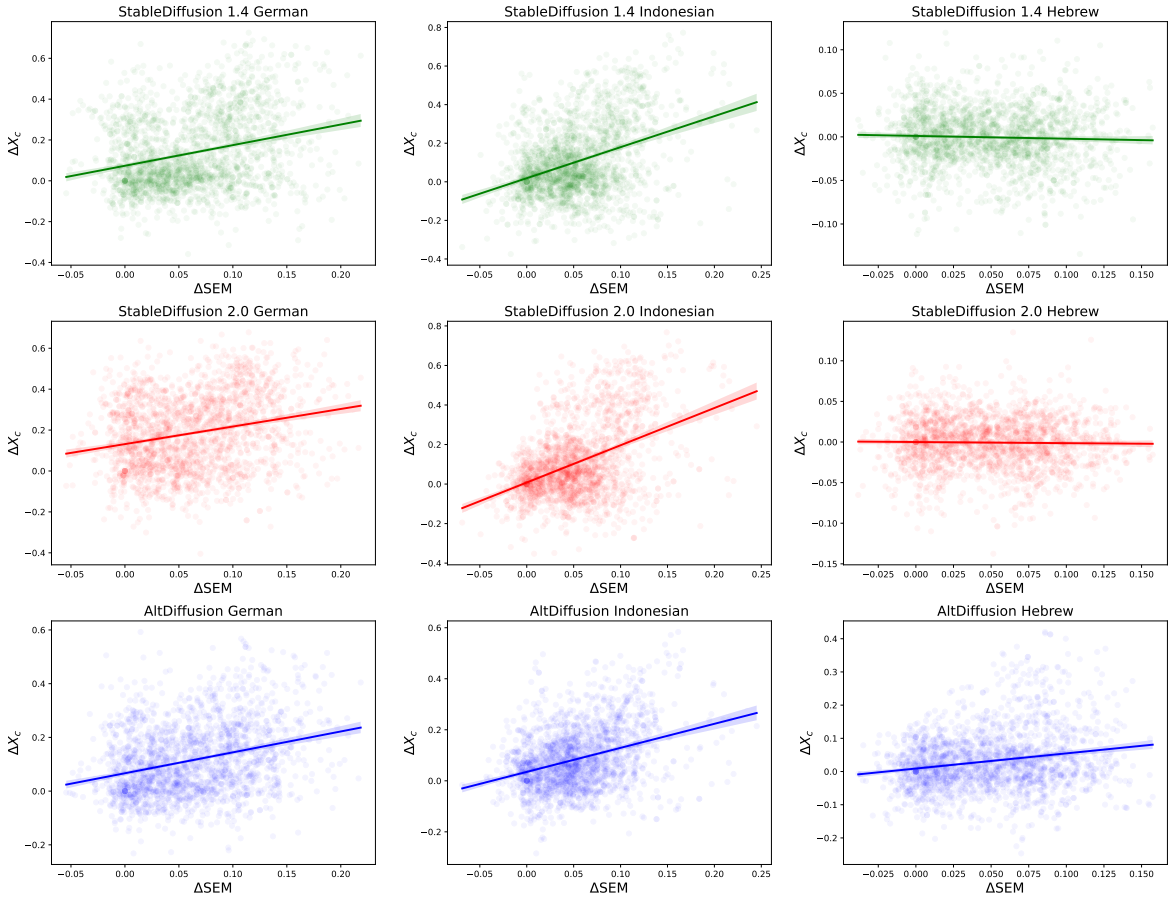


Figure 7.7: Scatterplots for the pseudocorrection experiments. Transparent circles are used to make distribution mass more visible.

Concept	Original	Corrected	Type	ΔSEM	ΔX_c (CCCL Improvement) for model			
					SD 1.4	SD 2	SD 2.1	AD
All Japanese-language error candidates:								
duck	鴨	アヒル	C	-0.092	0.021	0.008	-0.012	-0.055
thigh	腿	ふともも	C	-0.091	0.048	0.007	-0.043	-0.124
cop	警官	お巡りさん	F	-0.053	-0.160	-0.029	-0.055	-0.140
field	分野	田んぼ	A	-0.036	0.015	-0.151	-0.075	-0.058
butterfly	蝶	蝶々	C	-0.022	-0.004	0.025	0.009	-0.020
girlfriend	ガールフレンド	彼女	C	-0.013	0.044	0.166	0.196	-0.030
stingray	アカエイ	エイ	C	-0.008	-0.058	0.044	-0.006	-0.071
cigarette	煙草	たばこ	C	-0.007	0.054	0.043	-0.034	0.078
tail	尾	尻尾	C	-0.003	0.004	0.077	0.056	0.040
woman	女性	女	C	-0.001	0.108	-0.022	-0.014	-0.046
forest	森林	森	C	-0.000	0.226	0.081	0.032	0.051
teenager	ティーンエイジャー	少年	C, T	0.002	0.169	0.076	0.115	0.023
flame	火炎	炎	C	0.003	-0.062	-0.070	0.009	0.031
father	父	父親	F	0.010	-0.009	-0.010	0.014	0.003
watch	時計	腕時計	IS	0.011	0.487	0.080	0.062	0.006
teacher	先生	教師	IS	0.015	0.006	-0.051	-0.070	0.016
kid	キッド	子ども	C, T	0.017	0.098	0.070	0.065	0.068
doctor	先生	医者	IS	0.017	-0.006	0.031	0.018	0.050
ground	接地	地面	OS	0.022	-0.008	0.097	0.084	0.086
bike	バイク	自転車	OS, T	0.023	0.195	0.021	-0.018	0.020
detail	ディテール	詳細	C, T	0.024	0.002	0.036	0.043	-0.031
milk	乳	牛乳	OS	0.033	0.141	0.026	-0.002	0.215
cafeteria	カフェテリア	食堂	C, T	0.044	-0.192	-0.043	-0.034	0.064
rock	ロック	岩	IS, T	0.067	0.048	-0.029	-0.033	0.104
All Chinese-language error candidates:								
men	男人	很多人	A	-0.032	0.001	-0.180	-0.182	-0.411
stingray	黄貂鱼	鳐鱼	C	-0.030	0.082	0.206	0.213	-0.099
field	领域	田野	A	-0.017	-0.012	-0.136	-0.184	0.083
boat	船	小船	F	-0.001	-0.110	0.009	0.008	0.017
sister	姐姐	姐妹	F	-0.001	0.033	0.014	0.026	-0.014
wife	老婆	妻子	C	0.003	-0.021	0.124	0.177	-0.021
bottle	瓶	瓶子	C	0.004	-0.062	-0.021	0.032	0.075
church	教会	教堂	A	0.005	-0.068	0.076	0.078	-0.018
father	爸爸	父亲	C	0.009	0.027	-0.028	-0.059	0.145
mouth	口	嘴	C	0.011	-0.054	0.023	0.010	0.037
bell	钟	铃	A	0.013	-0.013	0.071	0.081	-0.001
cafeteria	自助餐厅	食堂	A	0.017	-0.102	-0.047	-0.054	0.071
orange	橙色	橙子	OS	0.019	0.002	-0.099	-0.104	0.067
belt	带	皮带	IS	0.029	0.025	0.045	0.034	0.040
suit	适合	西装	OS	0.033	-0.003	-0.062	-0.052	0.329
hallway	门厅	走廊	A	0.045	0.166	0.011	0.015	0.105
table	表	桌子	OS	0.064	-0.068	0.098	0.043	0.206
All Spanish-language error candidates:								
ticket	boleto	billete	C	-0.034	0.169	0.036	0.069	0.011
room	habitación	cuarto	C	-0.005	-0.184	-0.166	-0.094	-0.083
bird	pájaro	ave	C	-0.001	-0.437	-0.373	-0.433	-0.020
flame	llama	flama	T, C	0.004	-0.040	-0.134	-0.164	0.044
ship	navío	barco	C	0.005	0.002	0.132	0.149	-0.083
hill	cerro	colina	C	0.019	-0.023	-0.005	-0.116	0.078
kid	cabrito	joven	C, F	0.022	0.027	0.077	0.065	0.100
tent	tienda	tienda de acampar	A, IS	0.072	-0.005	0.013	-0.013	0.353
sandwich	emparedado	sándwich	C	0.098	0.254	0.519	0.534	0.339

Table 7.4: All identified concept translation error candidates in the original Co-Co-CroLa and their corresponding corrections in Japanese, Chinese, and Spanish. Each section is sorted in ascending order of ΔSEM. Error types are defined in §7.5.1

Part IV

Conclusion

Chapter 8

Conclusion

This dissertation describes an agenda where a similar mindset is applied to design, critique, and apply high standards for quality evaluation of high-level capabilities in generative models. Each of project is anchored in critical examination of the measurement targets and techniques driving AI advancement.

I believe this approach can serve as a corrective to the problematic trends in evaluation described in **Chapter 2**. The core problem in our approach to evaluation is a muddling of two conflicting goals:

- Scientifically demonstrating human-like abstract intelligence in AI systems.
- Providing useful metrics to organize R&D efforts and demonstrate the trustworthiness of commercial AI systems to customers.

However, these benchmarks may rely on flawed datasets (**Chapter 3**), use problematic metrics with only tenuous connection to real-world needs (**Chapter 4**), or fundamentally lack construct validity (**Chapter 5**). Benchmarks are trying to achieve both goals while failing to do either.

I introduced and critiqued a dynamic, constrained evaluation of multilingualism in text-to-image models in **Chapter 6** and **Chapter 7** with an eye toward realizing best practices for scientifically-valid evaluation, and learned lessons that I will carry forward in future work.

Per Kuhn's "The Structure of Scientific Revolutions," pre-paradigmatic fields are disorganized and diverse, lack consensus theories, and have no commonly accepted observational frame [258]. Our understanding of AI capabilities is clearly pre-paradigmatic. My long-term research goal is to help **develop a better AI evaluation discipline**, **introduce novel multimodal evaluations** and improve vision-language models, and text-to-image, text-to-video, and text-to-audio models, and **advance the multilingual and multicultural utility** of these multimodal systems.

8.1 Building an Evaluation Discipline

Many AI conferences already hold benchmarks and evaluation tracks. Work on metric development and benchmarking best practices has regularly appears at ICLR [209], *ACL [259], CVPR [260], and NeurIPS [261]. In my *Model Metrology* paper (excerpted in **Chapter 2**) we treat these as seeds for an AI evaluation community [77].

The Conference on Machine Translation (WMT) [262] has run shared tasks that *simultaneously benchmark translation systems and quality metrics* since 2006 [263]. In so doing, the MT community continually refreshes their measurement practices as the field advances. Similarly, the Generation, Evaluation & Metrics (GEM) Workshops [264], emphasizes research on metrics for evaluating generated text against gold references to better compare competing submissions. Their GEM benchmark was an early attempt at building a protocol for living text generation assessment where new tasks and metrics could be slotted into a comprehensive benchmark over time [265].

Right now we effectively rely on happy accidents to spread evaluation knowledge across these disciplinary barriers. By organizing this work not as "`[domain]`'s evaluation research" but as *core model metrology research applied to* `[domain]`, we can render this knowledge transfer routine. Model metrology venues should bring the culture of WMT and GEM to evaluation writ large, so lessons from all these domains can be shared—to start we should collect and promote these disparate threads of capability measurement research as a cohesive whole.

A **model metrology community** that resolves the contradictory nature of current evaluations and benchmarks must be built through conscious engagement between with domain experts and model consumers. Academic metrologists and metrology venues should actively solicit specifications for new tasks. These expert-designed settings could be framed as shared tasks or even as concrete evaluation bounties.

This community may effect sociological change toward valuing grounded and concrete work by providing opportunities for genuine technical novelty (eg., benchmark generating processes) alongside crucial data work on real-world constraints. Practical model metrology will probably emerge as a profitable industry where consultants operate similarly to penetration testing teams in information security, customizing stress tests for each scenario based on their broad knowledge of evaluation.

8.2 Future Directions

Improving multimodal AI. Surprising evaluation results stir up attention. LoCoVQA (**Chapter 5**) was covered in the press for demonstrating the insufficient nature of commerical LM evaluation[1]. Careful evaluation is an important public good! However, these evaluations should also be used to improve the systems themselves.

[1]TechCrunch: "Gemini's data-analyzing abilities aren't as good as Google claims" (Kyle Wiggers)

I want to build knowledge measures for higher-level capabilities like verbs and object attributes in LMs and T2I systems following from my prior work. I will use these assessments and the surprising discoveries made along the way to guide the training of better multilingual and prompt-coherent VLMs and T2I models.

Advancing multilingual and multicultural systems. While developing ***CoCo-CroLa*** (**Chapter 7**) I identified some intriguing future directions. For example, within "possessed" concepts T2I models expressed both distinct cultural knowledge and language-contingent social biases in their outputs. Stable Diffusion picks popular dog breeds in different cultures when prompted to generate "a dog" in their respective language, producing pictures of labrador retrievers and poodles in English and Spanish, but pictures of shiba inus when prompted in Japanese. This constitutes a cute example of cultural knowledge that would be cool to preserve under system modifications. However, the same model inappropriately sexualizes pictures of women in Spanish and Japanese, but not in English. An interesting and important question presents itself: can a model's cultural knowledge be preserved while multilingual performance is improved, and cross-lingual biases are eliminated?

Answering this question requires not just investment in *methods for achieving the goal* but also *metrics to track progress*. Thanks to ***CoCo-CroLa*** multilingual performance is now quantifiable; cultural knowledge in generated images is not.

In future work I aim to introduce sophisticated tracking and evaluation techniques for cultural capabilities in multimodal models, and use them to develop mitigations to complex multivariate problems in global generative AI, such as mitigating the competing demands of cultural knowledge, multilingual performance, and bias mitigation described above.

I plan to seek out interdisciplinary and multicultural collaborators who can provide

further insights into their generative AI system needs. I think my work will be useful for the digital humanities; systems that faithfully process non-English information, and represent diverse cultural concepts can be useful in analyzing global media trends or processing and cataloguing cultural information.

Complex verifiers to improve reasoning models. A shift from language models to so-called *reasoning models* kicked off in 2024 with the discovery that *inference-scaling*—training models to expend more runtime compute generating chains of thought [266], debate [267], and self-corrections [268]—can greatly increase the performance of models. However, this increase in the robustness and and correctness of answers generated from inference scaling [269] can only take place on domains with readily-available verifiers to provide a training signal.

On average, there is a direct relationship between amount of inference spend and performance on benchmarks of a variety of "reasoning tasks" [270]. However, these reasoning tasks have largely been limited to logic puzzles, math, and code. In future work I aim to broaden the scope of tasks which are legible to inference scaling approaches by producing sophisticated verifiers for complex desiderata beyond simple correctness.

Bibliography

[1] A. Saka, R. Taiwo, N. Saka, B. A. Salami, S. Ajayi, K. Akande, and H. Kazemi, *Gpt models in construction industry: Opportunities, limitations, and a use case validation*, *Developments in the Built Environment* **17** (2024) 100300.

[2] S. Goyal, E. Rastogi, S. P. Rajagopal, D. Yuan, F. Zhao, J. Chintagunta, G. Naik, and J. Ward, *Healai: A healthcare LLM for effective medical documentation*, in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, (New York, NY, USA), p. 1167–1168, Association for Computing Machinery, 2024.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et. al.*, *On the opportunities and risks of foundation models*, *ArXiv preprint* **abs/2108.07258** (2021).

[4] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, *et. al.*, *Constitutional AI: Harmlessness from AI feedback*, *ArXiv preprint* **abs/2212.08073** (2022).

[5] P. Cihon, *Standards for AI governance: International standards to enable global coordination in AI research & development*, *Future of Humanity Institute. University of Oxford* (2019) 340–342.

[6] NIST, *Artificial Intelligence Safety Institute consortium: Cooperative research and development agreement*, tech. rep., National Institue for Standards and Technology, 2023.

[7] S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, and E. Gómez, *Measuring the occupational impact of AI: Tasks, cognitive abilities and AI benchmarks*, *Journal of Artificial Intelligence Research* **71** (2021) 191–236.

[8] L. Li, J. Xu, Q. Dong, C. Zheng, X. Sun, L. Kong, and Q. Liu, *Can language models understand physical concepts?*, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 11843–11861, Association for Computational Linguistics, 2023.

[9] M. R. Morris, J. Sohl-dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg, *Levels of AGI: Operationalizing Progress on the Path to AGI*, *ArXiv preprint* **abs/2311.02462** (2023).

[10] J. Jumelet and D. Hupkes, *Do language models understand anything? on the ability of LSTMs to understand negative polarity items*, in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (T. Linzen, G. Chrupała, and A. Alishahi, eds.), (Brussels, Belgium), pp. 222–231, Association for Computational Linguistics, 2018.

[11] E. M. Bender and A. Koller, *Climbing towards NLU: On meaning, form, and understanding in the age of data*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 5185–5198, Association for Computational Linguistics, 2020.

[12] S. Park, S. Ryu, and E. Choi, *Do language models understand measurements?*, in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 1782–1792, Association for Computational Linguistics, 2022.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et. al.*, *Gpt-4 technical report*, *ArXiv preprint* **abs/2303.08774** (2023).

[14] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et. al.*, *Sparks of artificial general intelligence: Early experiments with gpt-4*, *ArXiv preprint* **abs/2303.12712** (2023).

[15] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, *et. al.*, *Training compute-optimal large language models*, *ArXiv preprint* **abs/2203.15556** (2022).

[16] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, *et. al.*, *Governing AI safety through independent audits*, *Nature Machine Intelligence* **3** (2021), no. 7 566–571.

[17] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord, *Are we done with imagenet?*, *ArXiv preprint* **abs/2006.07159** (2020).

[18] S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, and M. Samwald, *Mapping global dynamics of benchmark creation and saturation in artificial intelligence*, *Nature Communications* **13** (2022), no. 1 6793.

[19] C. Singer, *Notes on the early history of microscopy*, *Proceedings of the Royal Society of Medicine* **7** (1914), no. Sect_Hist_Med 247–279.

[20] A. E. F. La Berge, *The history of science and the history of microscopy*, *Perspectives on Science* **7** (1999), no. 1 111–142.

[21] D. Leverington, *A History of Astronomy: from 1890 to the Present*. Springer Science & Business Media, 2012.

[22] M. W. Davidson and M. Abramowitz, *Optical microscopy*, *Encyclopedia of imaging science and technology* **2** (2002), no. 1106-1141 120.

[23] D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada, *AI and the Everything in the Whole Wide World Benchmark*, in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (J. Vanschoren and S. Yeung, eds.), vol. 1, Curran, 2021.

[24] D. Donoho, *50 years of data science*, *Journal of Computational and Graphical Statistics* **26** (2017), no. 4 745–766.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et. al.*, *Imagenet large scale visual recognition challenge*, *International journal of computer vision* **115** (2015) 211–252.

[26] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *SuperGLUE: A multi-task benchmark and analysis platform for natural language understanding*, *Advances in Neural Information Processing Systems* **32** (2019) 3261–3275.

[27] Y. Ge, W. Hua, K. Mei, J. Ji, J. Tan, S. Xu, Z. Li, and Y. Zhang, *Openagi: When LLM meets domain experts*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[28] Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, J. Zhang, J. Li, and L. Hou, *Benchmarking foundation models with language-model-as-an-examiner*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[29] S. Piantadosi and F. Hill, *Meaning without reference in large language models*, in *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022.

[30] M. Mitchell and D. C. Krakauer, *The debate over understanding in AI's large language models*, *Proceedings of the National Academy of Sciences* **120** (2023), no. 13 e2215907120.

[31] J. Michael, A. Holtzman, A. Parrish, A. Mueller, A. Wang, A. Chen, D. Madaan, N. Nangia, R. Y. Pang, J. Phang, and S. R. Bowman, *What do NLP researchers believe? results of the NLP community metasurvey*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 16334–16368, Association for Computational Linguistics, 2023.

[32] H. Neri and F. Cozman, *The role of experts in the public perception of risk of artificial intelligence*, *AI & society* **35** (2020) 663–673.

[33] V. M. Ambartsoumean and R. V. Yampolskiy, *AI risk skepticism, a comprehensive survey*, *ArXiv preprint* **abs/2303.03885** (2023).

[34] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, *The FERET evaluation methodology for face-recognition algorithms*, *IEEE Transactions on pattern analysis and machine intelligence* **22** (2000), no. 10 1090–1104.

[35] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, *et. al.*, *Training verifiers to solve math word problems*, *ArXiv preprint* **abs/2110.14168** (2021).

[36] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, *Measuring massive multitask language understanding*, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[37] M. Ho, A. Sharma, J. Chang, M. Saxon, S. Levy, Y. Lu, and W. Y. Wang, *WikiWhy: Answering and explaining cause-and-effect questions*, in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

[38] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, *A large annotated corpus for learning natural language inference*, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (L. Màrquez, C. Callison-Burch, and J. Su, eds.), (Lisbon, Portugal), pp. 632–642, Association for Computational Linguistics, 2015.

[39] K. Paster, "Testing language models on a held-out high school national finals exam." `https://huggingface.co/datasets/keirp/hungarian_national_hs_finals_exam`, 2023.

[40] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, *LiveCodeBench: Holistic and contamination free evaluation of large language models for code*, *ArXiv preprint* **abs/2403.07974** (2024).

[41] T. Liao, R. Taori, I. D. Raji, and L. Schmidt, *Are we learning yet? A meta review of evaluation failures across machine learning*, in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[42] S. W. O'Leary-Kelly and R. J. Vokurka, *The empirical assessment of construct validity*, *Journal of operations management* **16** (1998), no. 4 387–405.

[43] E. Davis, *Benchmarks for automated commonsense reasoning: A survey*, *ACM Comput. Surv.* **56** (2023), no. 4.

[44] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, *et. al.*, *Towards artificial general intelligence via a multimodal foundation model*, *Nature Communications* **13** (2022), no. 1 3094.

[45] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, *QA-GNN: Reasoning with language models and knowledge graphs for question answering*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 535–546, Association for Computational Linguistics, 2021.

[46] J. Moore, *Language models understand us, poorly*, in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 214–222, Association for Computational Linguistics, 2022.

[47] Z. Zhao, W. S. Lee, and D. Hsu, *Large language models as commonsense knowledge for large-scale task planning*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[48] S. Murty, O. Paradise, and P. Sharma, *Pseudointelligence: A unifying lens on language model evaluation*, in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 7284–7290, Association for Computational Linguistics, 2023.

[49] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kociský, S. Ruder, D. Yogatama, K. Cao, S. Young, and P. Blunsom, *Mind the gap: Assessing temporal generalization in neural language models*, in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 29348–29363, 2021.

[50] L. Yang, S. Zhang, L. Qin, Y. Li, Y. Wang, H. Liu, J. Wang, X. Xie, and Y. Zhang, *GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective*, in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 12731–12750, Association for Computational Linguistics, 2023.

[51] P. Singhal, J. Forristal, X. Ye, and G. Durrett, *Assessing out-of-domain language model performance from few examples*, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (A. Vlachos and I. Augenstein, eds.), (Dubrovnik, Croatia), pp. 2385–2397, Association for Computational Linguistics, 2023.

[52] F. De Waal, *Are we smart enough to know how smart animals are?* WW Norton & Company, 2016.

[53] C. D. Manning, *Human language understanding & reasoning*, *Daedalus* **151** (2022), no. 2 127–138.

[54] C. Potts, "Is it possible for language models to achieve language understanding?." Medium post, 2020.

[55] H. De Vries, D. Bahdanau, and C. Manning, *Towards ecologically valid research on language user interfaces*, *ArXiv preprint* **abs/2007.14435** (2020).

[56] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. A. Cosgrove, C. D. Manning, C. Re, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. WANG, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. A. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, *Holistic evaluation of language models*, *Transactions on Machine Learning Research* (2023). Featured Certification, Expert Certification.

[57] Q. V. Liao and Z. Xiao, *Rethinking model evaluation as narrowing the socio-technical gap*, *ArXiv preprint* **abs/2306.03100** (2023).

[58] Anthropic, "The claude 3 model family: Opus, sonnet, haiku." Online technical report., 2023.

[59] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et. al.*, *Gemini: a family of highly capable multimodal models*, *ArXiv preprint* **abs/2312.11805** (2023).

[60] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, *et. al.*, *Gemma: Open models based on gemini research and technology*, *ArXiv preprint* **abs/2403.08295** (2024).

[61] A. Zhou, K. Wang, Z. Lu, W. Shi, S. Luo, Z. Qin, S. Lu, A. Jia, L. Song, M. Zhan, and H. Li, *Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[62] K. Gorman and S. Bedrick, *We need to talk about standard splits*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 2786–2791, Association for Computational Linguistics, 2019.

[63] M. Mitchell, A. S. Luccioni, N. Lambert, M. Gerchick, A. McMillan-Major, E. Ozoani, N. Rajani, T. Thrush, Y. Jernite, and D. Kiela, *Measuring data*, *ArXiv preprint* **abs/2212.05129** (2022).

[64] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, E. P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge, *What's in my big data?*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[65] N. Alzahrani, H. A. Alyahya, Y. Alnumay, S. Alrashed, S. Alsubaie, Y. Almushaykeh, F. Mirza, N. Alotaibi, N. Altwairesh, A. Alowisheq, *et. al.*, *When benchmarks are targets: Revealing the sensitivity of large language model leaderboards*, *ArXiv preprint* **abs/2402.01781** (2024).

[66] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme, *Hypothesis only baselines in natural language inference*, in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (M. Nissim, J. Berant, and A. Lenci, eds.), (New Orleans, Louisiana), pp. 180–191, Association for Computational Linguistics, 2018.

[67] T. McCoy, E. Pavlick, and T. Linzen, *Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 3428–3448, Association for Computational Linguistics, 2019.

[68] X. Wang, W. Chen, M. Saxon, and W. Y. Wang, *Counterfactual maximum likelihood estimation for training deep networks*, in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021,*

NeurIPS 2021, December 6-14, 2021, virtual (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 25072–25085, 2021.

[69] M. Saxon, X. Wang, W. Xu, and W. Y. Wang, *PECO: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers*, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (A. Vlachos and I. Augenstein, eds.), (Dubrovnik, Croatia), pp. 3061–3074, Association for Computational Linguistics, 2023.

[70] Z. Peng, Z. Wang, and D. Deng, *Near-duplicate sequence search at scale for large language model memorization evaluation*, *Proc. ACM Manag. Data* **1** (2023), no. 2.

[71] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, *Large language models struggle to learn long-tail knowledge*, in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707, PMLR, 2023.

[72] A. Z. Jacobs and H. Wallach, *Measurement and fairness*, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 375–385, Association for Computing Machinery, 2021.

[73] P. A. M. Casares, B. S. Loe, J. Burden, S. Ó. hÉigeartaigh, and J. Hernández-Orallo, *How general-purpose is a language model? usefulness and safety with human prompters in the wild*, in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 5295–5303, AAAI Press, 2022.

[74] D. Kang, T. Kang, and J. Jang, *Papers with code or without code? Impact of GitHub repository usability on the diffusion of machine learning research*, *Information Processing & Management* **60** (2023), no. 6 103477.

[75] B. Koch, E. Denton, A. Hanna, and J. G. Foster, *Reduced, reused and recycled: The life of a dataset in machine learning research*, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* **1** (2021).

[76] A. M. Hoyle, P. Goel, A. Hian-Cheong, D. Peskov, J. L. Boyd-Graber, and P. Resnik, *Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence*, in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 2018–2033, 2021.

[77] M. Saxon, A. Holtzman, P. West, W. Y. Wang, and N. Saphra, *Benchmarks as microscopes: A call for model metrology*, in *First Conference on Language Modeling*, 2024.

[78] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, *A SICK cure for the evaluation of compositional distributional semantic models*, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Reykjavik, Iceland), pp. 216–223, European Language Resources Association (ELRA), 2014.

[79] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, *A large annotated corpus for learning natural language inference*, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 632–642, Association for Computational Linguistics, 2015.

[80] A. Williams, N. Nangia, and S. Bowman, *A broad-coverage challenge corpus for sentence understanding through inference*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 1112–1122, Association for Computational Linguistics, 2018.

[81] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, *XNLI: Evaluating cross-lingual sentence representations*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 2475–2485, Association for Computational Linguistics, 2018.

[82] Y. Nie, H. Chen, and M. Bansal, *Combining fact extraction and verification with neural semantic matching networks*, in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6859–6866, AAAI Press, 2019.

[83] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, *Adversarial NLI: A new benchmark for natural language understanding*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4885–4901, Association for Computational Linguistics, 2020.

[84] H. Hu, K. Richardson, L. Xu, L. Li, S. Kübler, and L. Moss, *OCNLI: Original Chinese Natural Language Inference*, in *Findings of the Association for Computational*

Linguistics: EMNLP 2020 (T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 3512–3526, Association for Computational Linguistics, 2020.

[85] D. Kaushik, A. Setlur, E. H. Hovy, and Z. C. Lipton, *Explaining the efficacy of counterfactually augmented data*, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[86] Y. Wu, M. Gardner, P. Stenetorp, and P. Dasigi, *Generating data to mitigate spurious correlations in natural language inference datasets*, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), (Dublin, Ireland), pp. 2660–2676, Association for Computational Linguistics, 2022.

[87] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, *Annotation artifacts in natural language inference data*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 107–112, Association for Computational Linguistics, 2018.

[88] G. Zhang, B. Bai, J. Liang, K. Bai, S. Chang, M. Yu, C. Zhu, and T. Zhao, *Selection bias explorations and debias methods for natural language sentence matching datasets*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 4418–4429, Association for Computational Linguistics, 2019.

[89] S. R. Bowman and G. Dahl, *What will it take to fix benchmarking in natural language understanding?*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 4843–4855, Association for Computational Linguistics, 2021.

[90] Z. Chen, Q. Gao, and L. S. Moss, *NeuralLog: Natural language inference with joint neural and logical reasoning*, in *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics* (L.-W. Ku, V. Nastase, and I. Vulić, eds.), (Online), pp. 78–88, Association for Computational Linguistics, 2021.

[91] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, *Entailment as few-shot learner*, *ArXiv preprint* **abs/2104.14690** (2021).

[92] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, *J. Mach. Learn. Res.* **21** (2020) 140:1–140:67.

[93] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, *ByT5: Towards a token-free future with pre-trained byte-to-byte models*, *Transactions of the Association for Computational Linguistics* **10** (2022) 291–306.

[94] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel, *KILT: a benchmark for knowledge intensive language tasks*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 2523–2544, Association for Computational Linguistics, 2021.

[95] B. Wang, S. Wang, Y. Cheng, Z. Gan, R. Jia, B. Li, and J. Liu, *Infobert: Improving robustness of language models from an information theoretic perspective*, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[96] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *ALBERT: A lite BERT for self-supervised learning of language representations*, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020.

[97] L. Xu, H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu, Y. Tian, Q. Dong, W. Liu, B. Shi, Y. Cui, J. Li, J. Zeng, R. Wang, W. Xie, Y. Li, Y. Patterson, Z. Tian, Y. Zhang, H. Zhou, S. Liu, Z. Zhao, Q. Zhao, C. Yue, X. Zhang, Z. Yang, K. Richardson, and Z. Lan, *CLUE: A Chinese language understanding evaluation benchmark*, in *Proceedings of the 28th International Conference on Computational Linguistics* (D. Scott, N. Bel, and C. Zong, eds.), (Barcelona, Spain (Online)), pp. 4762–4772, International Committee on Computational Linguistics, 2020.

[98] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*, *Transactions of the Association for Computational Linguistics* **2** (2014) 67–78.

[99] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, *SemEval-2012 task 6: A pilot on semantic textual similarity*, in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, and D. Yuret, eds.), (Montréal, Canada), pp. 385–393, Association for Computational Linguistics, 2012.

[100] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, *Visual genome: Connecting language and vision using crowdsourced dense image annotations*, *Int. J. Comput. Vision* **123** (2017), no. 1 32–73.

[101] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, *FEVER: a large-scale dataset for fact extraction and VERification*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 809–819, Association for Computational Linguistics, 2018.

[102] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, *HotpotQA: A dataset for diverse, explainable multi-hop question answering*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 2369–2380, Association for Computational Linguistics, 2018.

[103] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, *Learning word vectors for sentiment analysis*, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (D. Lin, Y. Matsumoto, and R. Mihalcea, eds.), (Portland, Oregon, USA), pp. 142–150, Association for Computational Linguistics, 2011.

[104] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, *OpenAI blog* **1** (2019), no. 8 9.

[105] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, *ArXiv preprint* **abs/1907.11692** (2019).

[106] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, *Unsupervised cross-lingual representation learning at scale*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 8440–8451, Association for Computational Linguistics, 2020.

[107] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, 2019.

[108] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, *Transformers: State-of-the-art natural language processing*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Q. Liu and D. Schlangen, eds.), (Online), pp. 38–45, Association for Computational Linguistics, 2020.

[109] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, PMLR, 2017.

[110] W. Huang, H. Liu, and S. R. Bowman, *Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data*, in *Proceedings of the First Workshop on Insights from Negative Results in NLP* (A. Rogers, J. Sedoc, and A. Rumshisky, eds.), (Online), pp. 82–87, Association for Computational Linguistics, 2020.

[111] L. Tu, G. Lalwani, S. Gella, and H. He, *An empirical study on robustness to spurious correlations using pre-trained language models*, *Transactions of the Association for Computational Linguistics* **8** (2020) 621–633.

[112] M. Geva, Y. Goldberg, and J. Berant, *Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets*, 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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 1161–1166, Association for Computational Linguistics, 2019.

[113] J. P. McKenna, S. Choudhary, M. Saxon, G. P. Strimel, and A. Mouchtaris, *Semantic complexity in end-to-end spoken language understanding*, in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020* (H. Meng, B. Xu, and T. F. Zheng, eds.), pp. 4273–4277, ISCA, 2020.

[114] M. Saxon, S. Choudhary, J. P. McKenna, and A. Mouchtaris, *End-to-end spoken language understanding for generalized voice assistants*, in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021* (H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, eds.), pp. 4738–4742, ISCA, 2021.

[115] M. Gardner, W. Merrill, J. Dodge, M. Peters, A. Ross, S. Singh, and N. A. Smith, *Competency problems: On finding and removing artifacts in language data*, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 1801–1813, Association for Computational Linguistics, 2021.

[116] Y. Belinkov, A. Poliak, S. Shieber, B. Van Durme, and A. Rush, *Don't take the premise for granted: Mitigating artifacts in natural language inference*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 877–891, Association for Computational Linguistics, 2019.

[117] N. S. Moosavi, M. de Boer, P. A. Utama, and I. Gurevych, *Improving robustness by augmenting training sentences with predicate-argument structures*, *ArXiv preprint* **abs/2010.12510** (2020).

[118] J. Min, R. T. McCoy, D. Das, E. Pitler, and T. Linzen, *Syntactic data augmentation increases robustness to inference heuristics*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 2339–2352, Association for Computational Linguistics, 2020.

[119] X. Zhou and M. Bansal, *Towards robustifying NLI models against lexical dataset biases*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 8759–8771, Association for Computational Linguistics, 2020.

[120] P. Bhargava, A. Drozd, and A. Rogers, *Generalization in NLI: Ways (not) to go beyond simple heuristics*, in *Proceedings of the Second Workshop on Insights from Negative Results in NLP* (J. Sedoc, A. Rogers, A. Rumshisky, and S. Tafreshi, eds.), (Online and Punta Cana, Dominican Republic), pp. 125–135, Association for Computational Linguistics, 2021.

[121] N. Varshney, P. Banerjee, T. Gokhale, and C. Baral, *Unsupervised natural language inference using PHL triplet generation*, in *Findings of the Association for Computational Linguistics: ACL 2022* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), (Dublin, Ireland), pp. 2003–2016, Association for Computational Linguistics, 2022.

[122] A. Liu, S. Swayamdipta, N. A. Smith, and Y. Choi, *WANLI: Worker and AI collaboration for natural language inference dataset creation*, in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 6826–6847, Association for Computational Linguistics, 2022.

[123] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, *Dataset cartography: Mapping and diagnosing datasets with training*

dynamics, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 9275–9293, Association for Computational Linguistics, 2020.

[124] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, *Zero-shot text-to-image generation*, in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831, PMLR, 2021.

[125] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical text-conditional image generation with clip latents*, *ArXiv preprint* **abs/2204.06125** (2022).

[126] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[127] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, *Photorealistic text-to-image diffusion models with deep language understanding*, in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), 2022.

[128] V. Petsiuk, A. E. Siemenn, S. Surbehera, Z. Chin, K. Tyser, G. Hunter, A. Raghavan, Y. Hicke, B. A. Plummer, O. Kerret, *et. al.*, *Human evaluation of text-to-image models on a multi-task benchmark*, *ArXiv preprint* **abs/2211.12112** (2022).

[129] J. Cho, Y. Hu, J. M. Baldridge, R. Garg, P. Anderson, R. Krishna, M. Bansal, J. Pont-Tuset, and S. Wang, *Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[130] S. Wu, H. Fei, H. Zhang, and T. Chua, *Imagine that! Abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[131] W. Feng, X. He, T. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, *Training-free structured diffusion guidance for compositional text-to-image synthesis*, in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

[132] L. Lian, B. Li, A. Yala, and T. Darrell, *LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models*, *ArXiv preprint* **abs/2305.13655** (2023).

[133] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, *Compositional visual generation with composable diffusion models*, in *European Conference on Computer Vision*, pp. 423–439, Springer, 2022.

[134] J. Zakraoui, M. Saleh, S. Al-Maadeed, and J. M. Jaam, *Improving text-to-image generation with object layout guidance*, *Multimedia Tools and Applications* **80** (2021), no. 18 27423–27443.

[135] W. Feng, W. Zhu, T. Fu, V. Jampani, A. R. Akula, X. He, S. Basu, X. E. Wang, and W. Y. Wang, *LayoutGPT: Compositional visual planning and generation with large language models*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[136] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training GANs*, in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain* (D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, eds.), pp. 2226–2234, 2016.

[137] M. Ku, T. Li, K. Zhang, Y. Lu, X. Fu, W. Zhuang, and W. Chen, *ImagenHub: Standardizing the evaluation of conditional image generation models*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[138] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, *CLIPScore: A reference-free evaluation metric for image captioning*, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 7514–7528, Association for Computational Linguistics, 2021.

[139] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, *TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering*, in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 20349–20360, IEEE, 2023.

[140] Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang, *LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[141] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, *VIEScore: Towards explainable metrics for conditional image synthesis evaluation*, *ArXiv preprint* **abs/2312.14867** (2023).

[142] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, *Deep generative image models using a laplacian pyramid of adversarial networks*, in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 1486–1494, 2015.

[143] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, *Aligning text-to-image models using human feedback*, *ArXiv preprint* **abs/2302.12192** (2023).

[144] W. Zhu, X. Wang, P. Narayana, K. Sone, S. Basu, and W. Y. Wang, *Towards understanding sample variance in visually grounded language generation: Evaluations and observations*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 8806–8811, Association for Computational Linguistics, 2020.

[145] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5967–5976, IEEE Computer Society, 2017.

[146] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, *SDEdit: Guided image synthesis and editing with stochastic differential equations*, in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.

[147] C. Spearman, *The proof and measurement of association between two things*, *The American Journal of Psychology* **15** (1904), no. 1 72–101.

[148] A. N. Kolmogorov, *Sulla determinazione empirica di una legge didistribuzione (on the empirical determination of a distribution)*, *Giorn Dell'inst Ital Degli Att* **4** (1933) 89–91.

[149] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, *T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[150] M. Saxon and W. Y. Wang, *Multilingual conceptual coverage in text-to-image models*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 4831–4848, Association for Computational Linguistics, 2023.

[151] M. Saxon, Y. Luo, S. Levy, C. Baral, Y. Yang, and W. Y. Wang, *Lost in translation? Translation errors and challenges for fair assessment of text-to-image models on multilingual concepts*, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)* (K. Duh, H. Gomez, and S. Bethard, eds.), (Mexico City, Mexico), pp. 572–582, Association for Computational Linguistics, 2024.

[152] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, *Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale*, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1493–1504, Association for Computing Machinery.

[153] M. Hodosh, P. Young, and J. Hockenmaier, *Framing image description as a ranking task: Data, models and evaluation metrics*, in *IJCAI 2015 - Proceedings of the 24th International Joint Conference on Artificial Intelligence* (M. Wooldridge and Q. Yang, eds.), IJCAI International Joint Conference on Artificial Intelligence, pp. 4188–4192, International Joint Conferences on Artificial Intelligence, 2015. 24th International Joint Conference on Artificial Intelligence, IJCAI 2015 ; Conference date: 25-07-2015 Through 31-07-2015.

[154] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, *Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649, IEEE Computer Society, 2015.

[155] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, *Microsoft COCO Captions: Data collection and evaluation server*, 2015.

[156] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, *et. al.*, *Scaling autoregressive models for content-rich text-to-image generation*, *ArXiv preprint* **abs/2206.10789** (2022).

[157] J. Cho, A. Zala, and M. Bansal, *DALL-EVAL: probing the reasoning skills and social biases of text-to-image generation models*, in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3020–3031, IEEE, 2023.

[158] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, *Pick-a-pic: An open dataset of user preferences for text-to-image generation*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[159] M. Yarom, Y. Bitton, S. Changpinyo, R. Aharoni, J. Herzig, O. Lang, E. Ofek, and I. Szpektor, *What you see is what you read? Improving text-image alignment evaluation*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[160] J. W. Pratt, J. D. Gibbons, J. W. Pratt, and J. D. Gibbons, *Kolmogorov-Smirnov two-sample tests*, *Concepts of nonparametric theory* (1981) 318–344.

[161] V. W. Berger and Y. Zhou, *Kolmogorov–Smirnov test: Overview*, *Wiley statsref: Statistics reference online* (2014).

[162] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, *Learning transferable visual models from natural language supervision*, in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021.

[163] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, *SDXL: improving latent diffusion models for high-resolution image synthesis*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[164] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[165] Y. Zha, Y. Yang, R. Li, and Z. Hu, *AlignScore: Evaluating factual consistency with a unified alignment function*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 11328–11348, Association for Computational Linguistics, 2023.

[166] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, *Scaling up visual and vision-language representation learning with noisy text supervision*, in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916, PMLR, 2021.

[167] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si, *mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections*, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 7241–7259, Association for Computational Linguistics, 2022.

[168] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Jiang, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, *mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality*, 2023.

[169] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et. al.*, *Llama: Open and efficient foundation language models*, *ArXiv preprint* **abs/2302.13971** (2023).

[170] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[171] H. Liu, C. Li, Q. Wu, and Y. J. Lee, *Visual instruction tuning*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[172] H. Liu, C. Li, Y. Li, and Y. J. Lee, *Improved baselines with visual instruction tuning*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296, IEEE, 2024.

[173] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR, 2022.

[174] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, *InstructBLIP: Towards general-purpose vision-language models with instruction tuning*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[175] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar, *Introducing our multimodal models (Fuyu-8b)*, 2023.

[176] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, *Scaling laws for neural language models*, *ArXiv preprint* **abs/2001.08361** (2020).

[177] Z. Dong, T. Tang, L. Li, and W. X. Zhao, *A survey on long text modeling with transformers*, *ArXiv preprint* **abs/2302.14502** (2023).

[178] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, *et. al.*, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, *ArXiv preprint* **abs/2403.05530** (2024).

[179] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, *et. al.*, *Mixtral of experts*, *ArXiv preprint* **abs/2401.04088** (2024).

[180] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, *Mantis: Interleaved multi-image instruction tuning*, *ArXiv preprint* **abs/2405.01483** (2024).

[181] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding, and J. Tang, *Cogagent: A visual language model for GUI agents*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14281–14290, IEEE, 2024.

[182] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, *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* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett,

eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, PMLR, 2023.

[183] G. Zhang, Y. Zhang, K. Zhang, and V. Tresp, *Can vision-language models be a good guesser? exploring vlms for times and location reasoning*, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 636–645, 2024.

[184] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, *Git: A generative image-to-text transformer for vision and language*, *ArXiv preprint* **abs/2205.14100** (2022).

[185] V.-Q. Nguyen, M. Suganuma, and T. Okatani, *Grit: Faster and better image captioning transformer using dual visual features*, in *European Conference on Computer Vision*, pp. 167–184, Springer, 2022.

[186] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. V. Thapliyal, J. Bradbury, and W. Kuo, *Pali: A jointly-scaled multilingual language-image model*, in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

[187] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, *Image as a foreign language: BEIT pretraining for vision and vision-language tasks*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 19175–19186, IEEE, 2023.

[188] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, *mplug-2: A modularized multi-modal foundation model across text, image and video*, in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 38728–38748, PMLR, 2023.

[189] H.-T. Chen, F. Xu, S. A. Arora, and E. Choi, *Understanding retrieval augmentation for long-form question answering*, *ArXiv preprint* **abs/2310.12150** (2023).

[190] J. Phang, Y. Zhao, and P. Liu, *Investigating efficiently extending transformers for long input summarization*, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 3946–3961, Association for Computational Linguistics, 2023.

[191] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela,

Retrieval-augmented generation for knowledge-intensive NLP tasks, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[192] P. Xu, W. Ping, X. Wu, L. McAfee, C. Zhu, Z. Liu, S. Subramanian, E. Bakhturina, M. Shoeybi, and B. Catanzaro, *Retrieval meets long context large language models*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[193] Y. Fan, J. Gu, K. Zhou, Q. Yan, S. Jiang, C.-C. Kuo, X. Guan, and X. E. Wang, *Muffin or chihuahua? Challenging large vision-language models with multipanel VQA*, *ArXiv preprint* **abs/2401.15847** (2024).

[194] S. Ghosh, C. K. R. Evuru, S. Kumar, U. Tyagi, O. Nieto, Z. Jin, and D. Manocha, *VDGD: Mitigating LVLM hallucinations in cognitive prompts by bridging the visual perception gap*, *ArXiv preprint* **abs/2405.15683** (2024).

[195] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, *Multi-modal hallucination control by visual information grounding*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14303–14312, IEEE, 2024.

[196] D. Song, S. Chen, G. H. Chen, F. Yu, X. Wan, and B. Wang, *MileBench: Benchmarking MLLMs in long context*, *ArXiv preprint* **abs/2404.18532** (2024).

[197] W. Wang, S. Zhang, Y. Ren, Y. Duan, T. Li, S. Liu, M. Hu, Z. Chen, K. Zhang, L. Lu, X. Zhu, P. Luo, Y. Qiao, J. Dai, W. Shao, and W. Wang, *Needle in A multimodal haystack*, in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[198] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, *ELI5: Long form question answering*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 3558–3567, Association for Computational Linguistics, 2019.

[199] L. Huang, S. Cao, N. Parulian, H. Ji, and L. Wang, *Efficient attentions for long document summarization*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur,

I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 1419–1436, Association for Computational Linguistics, 2021.

[200] A. Dilawari and M. U. G. Khan, *ASoVS: Abstractive summarization of video sequences*, *IEEE Access* **7** (2019) 29253–29263.

[201] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, *OK-VQA: A visual question answering benchmark requiring external knowledge*, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204, Computer Vision Foundation / IEEE, 2019.

[202] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, and F. Zhao, *Are we on the right way for evaluating large vision-language models?*, in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[203] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998), no. 11 2278–2324.

[204] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *BERTScore: Evaluating text generation with BERT*, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020.

[205] C.-Y. Lin, *ROUGE: A package for automatic evaluation of summaries*, in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, 2004.

[206] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, *Seed-bench: Benchmarking multimodal LLMs with generative comprehension*, *ArXiv preprint* **abs/2307.16125** (2023).

[207] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, *et. al.*, *MMbench: Is your multi-modal model an all-around player?*, *ArXiv preprint* **abs/2307.06281** (2023).

[208] X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, *MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567, IEEE, 2024.

[209] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K. Chang, M. Galley, and J. Gao, *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[210] P. Lu, S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan, *Learn to explain: Multimodal reasoning via thought chains for science question answering*, in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), 2022.

[211] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, *A diagram is worth a dozen images*, in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251, Springer, 2016.

[212] Moondream, "tiny vision language model." `https://github.com/vikhyat/moondream`, 2024.

[213] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*, 2024.

[214] Google, "Paligemma." `https://ai.google.dev/gemma/docs/paligemma`, 2024.

[215] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov, *Generating images from captions with attention*, in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2016.

[216] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, *Generative adversarial text to image synthesis*, in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (M. Balcan and K. Q. Weinberger, eds.), vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1060–1069, JMLR.org, 2016.

[217] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 7871–7880, Association for Computational Linguistics, 2020.

[218] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2014.

[219] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, *Communications of the ACM* **63** (2020), no. 11 139–144.

[220] M. Kwon, J. Jeong, and Y. Uh, *Diffusion models already have A semantic latent space*, in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

[221] B. Hutchinson, J. Baldridge, and V. Prabhakaran, *Underspecification in scene description-to-depiction tasks*, in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, eds.), (Online only), pp. 1172–1184, Association for Computational Linguistics, 2022.

[222] V. Prabhu and A. Birhane, *Large datasets: A pyrrhic win for computer vision*, *ArXiv preprint* **abs/2006.16923** (2020).

[223] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, *et. al.*, *The pile: An 800gb dataset of diverse text for language modeling*, *ArXiv preprint* **abs/2101.00027** (2021).

[224] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, *LAION-400m: Open dataset of CLIP-filtered 400 million image-text pairs*, *ArXiv preprint* **abs/2111.02114** (2021).

[225] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, *Model cards for model reporting*, in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

[226] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, *Datasheets for datasets*, *Communications of the ACM* **64** (2021), no. 12 86–92.

[227] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas, and R. Ghosh, *Dall·e mini*, 2021.

[228] J. Howard and S. Ruder, *Universal language model fine-tuning for text classification*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (I. Gurevych and Y. Miyao, eds.), (Melbourne, Australia), pp. 328–339, Association for Computational Linguistics, 2018.

[229] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, *Dense passage retrieval for open-domain question answering*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 6769–6781, Association for Computational Linguistics, 2020.

[230] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping*, *ArXiv preprint* **abs/2002.06305** (2020).

[231] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, *RACE: Large-scale ReAding comprehension dataset from examinations*, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (M. Palmer, R. Hwa, and S. Riedel, eds.), (Copenhagen, Denmark), pp. 785–794, Association for Computational Linguistics, 2017.

[232] P. Rajpurkar, R. Jia, and P. Liang, *Know what you don't know: Unanswerable questions for SQuAD*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (I. Gurevych and Y. Miyao, eds.), (Melbourne, Australia), pp. 784–789, Association for Computational Linguistics, 2018.

[233] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, *The woman worked as a babysitter: On biases in language generation*, 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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3407–3412, Association for Computational Linguistics, 2019.

[234] C. Clark, M. Yatskar, and L. Zettlemoyer, *Don't take the easy way out: Ensemble based methods for avoiding known dataset biases*, 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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 4069–4082, Association for Computational Linguistics, 2019.

[235] H. Cho, C. Sankar, C. Lin, K. R. Sadagopan, S. Shayandeh, A. Celikyilmaz, J. May, and A. Beirami, *CheckDST: Measuring real-world generalization of dialogue state tracking performance*, *ArXiv preprint* **abs/2112.08321** (2021).

[236] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, *Beyond accuracy: Behavioral testing of NLP models with CheckList*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4902–4912, Association for Computational Linguistics, 2020.

[237] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language models are few-shot learners*, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[238] W. Xu, Y.-L. Tuan, Y. Lu, M. Saxon, L. Li, and W. Y. Wang, *Not all errors are equal: Learning text generation metrics using stratified error synthesis*, in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 6559–6574, Association for Computational Linguistics, 2022.

[239] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, *et. al.*, *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*, *ArXiv preprint* **abs/2206.04615** (2022).

[240] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, and J. Wei, *Challenging BIG-bench tasks and whether chain-of-thought can solve them*, in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 13003–13051, Association for Computational Linguistics, 2023.

[241] Z. Chen, G. Liu, B.-W. Zhang, Q. Yang, and L. Wu, *AltCLIP: Altering the language encoder in CLIP for extended language capabilities*, in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 8666–8682, Association for Computational Linguistics, 2023.

[242] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang, *Cogview: Mastering text-to-image generation via transformers*, in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 19822–19835, 2021.

[243] A. Krizhevsky, G. Hinton, *et. al.*, *Learning multiple layers of features from tiny images*, .

[244] R. Navigli and S. P. Ponzetto, *BabelNet: Building a very large multilingual semantic network*, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (J. Hajič, S. Carberry, S. Clark, and J. Nivre, eds.), (Uppsala, Sweden), pp. 216–225, Association for Computational Linguistics, 2010.

[245] S. Engelmann, C. Ullstein, O. Papakyriakopoulos, and J. Grossklags, *What people think AI should infer from faces*, in *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, (New York, NY, USA), p. 128–141, Association for Computing Machinery, 2022.

[246] N. Reimers and I. Gurevych, *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*, 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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, 2019.

[247] M. Ando, J. Okamoto, and S. Ishizaki, *Extraction of associative attributes from nouns and quantitative expression of prototype concept*, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (M. González Rodríguez and C. P. Suarez Araujo, eds.), (Las Palmas, Canary Islands - Spain), European Language Resources Association (ELRA), 2002.

[248] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi, *Gender bias in machine translation*, *Transactions of the Association for Computational Linguistics* **9** (2021) 845–874.

[249] A. Irvine, J. Morgan, M. Carpuat, H. Daumé III, and D. Munteanu, *Measuring machine translation errors in new domains*, *Transactions of the Association for Computational Linguistics* **1** (2013) 429–440.

[250] J. Hitschler, S. Schamoni, and S. Riezler, *Multimodal pivots for image caption translation*, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 2399–2409, Association for Computational Linguistics, 2016.

[251] S. M. Singh, L. Sanayai Meetei, T. D. Singh, and S. Bandyopadhyay, *Multiple captions embellished multilingual multi-modal neural machine translation*, in *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)* (T. Doren Singh, C. España i Bonet, S. Bandyopadhyay, and J. van Genabith, eds.), (Online (Virtual Mode)), pp. 2–11, INCOMA Ltd., 2021.

[252] M. Patel, T. Gokhale, C. Baral, and Y. Yang, *ConceptBed: Evaluating concept learning abilities of text-to-image diffusion models*, in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada* (M. J. Wooldridge, J. G. Dy, and S. Natarajan, eds.), pp. 14554–14562, AAAI Press, 2024.

[253] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, *Don't just assume; look and answer: Overcoming priors for visual question answering*, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4971–4980, IEEE Computer Society, 2018.

[254] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, *VQA: visual question answering*, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433, IEEE Computer Society, 2015.

[255] C. Y. Cui, A. Khandelwal, Y. Artzi, N. Snavely, and H. Averbuch-Elor, *Who's Waldo? Linking people across text and images*, in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1354–1364, IEEE, 2021.

[256] K. Ye and A. Kovashka, *A case study of the shortcut effects in visual commonsense reasoning*, in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 3181–3189, AAAI Press, 2021.

[257] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, *From recognition to cognition: Visual commonsense reasoning*, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6720–6731, Computer Vision Foundation / IEEE, 2019.

[258] T. S. Kuhn, *The structure of scientific revolutions*. University of Chicago press, 1962.

[259] J. Maynez, P. Agrawal, and S. Gehrmann, *Benchmarking large language model capabilities for conditional generation*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 9194–9213, Association for Computational Linguistics, 2023.

[260] X. Xu, Z. Qi, J. Ma, H. Zhang, Y. Shan, and X. Qie, *BTS: A bi-lingual benchmark for text segmentation in the wild*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19152–19162, 2022.

[261] W. Zhang, M. Aljunied, C. Gao, Y. K. Chia, and L. Bing, *M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models*, in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[262] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, and M. Shmatova, *Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet*, in *Proceedings of the Eighth Conference on Machine Translation* (P. Koehn, B. Haddow, T. Kocmi, and C. Monz, eds.), (Singapore), pp. 1–42, Association for Computational Linguistics, 2023.

[263] O. Bojar, C. Federmann, B. Haddow, P. Koehn, M. Post, and L. Specia, *Ten years of WMT evaluation campaigns: Lessons learnt*, in *LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pp. 27–34, 2016.

[264] S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, eds., *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, (Singapore), Association for Computational Linguistics, 2023.

[265] S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou, *The GEM benchmark: Natural language generation, its evaluation and metrics*, in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)* (A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, eds.), (Online), pp. 96–120, Association for Computational Linguistics, 2021.

[266] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, *Chain-of-thought prompting elicits reasoning in large language models*, in *Advances in Neural Information Processing Systems 35: Annual Conference on*

Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), 2022.

[267] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi, *Encouraging divergent thinking in large language models through multi-agent debate*, *ArXiv preprint* **abs/2305.19118** (2023).

[268] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, and W. Y. Wang, *Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies*, *Transactions of the Association for Computational Linguistics* **12** (2024) 484–506.

[269] Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang, *Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models*, 2024.

[270] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, *et. al.*, *Openai o1 system card*, *ArXiv preprint* **abs/2412.16720** (2024).