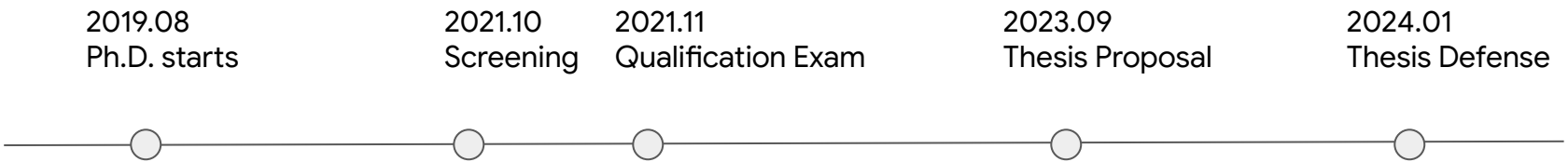
Emphasizing the Importance of Data and Evaluation in the Era of Large Language Models

Jiao Sun

jiaosun@usc.edu

University of Southern California

TimeLine



- CHI 2022 Best Paper Honorable Mention
- ACL 2021 Best Paper Nomination
- Amazon ML Fellow
- 2023 EECS Rising Star

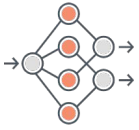
Inspecting Each Stage of Generation

Trustworthiness

Are generation models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks

ACL 2023 — Dialect-robust Evaluation of Generated Text

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

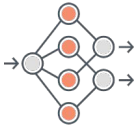
Inspecting Each Stage of Generation

Trustworthiness

Are generation models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2024 — **Evaluating Large Language Models on Controlled Generation Tasks**

ACL 2023 — **Dialect-robust Evaluation of Generated Text**

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — **DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback**

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

"**ChatGPT** is scary good. We are not far from dangerously strong AI."

"Time it took to reach 1 million users:
Netflix - 3.5 years
Facebook - 10 months
Spotify - 5 months
Instagram - 2.5 months
ChatGPT - 5 days"

TechCrunch

App Store and Play Store are flooded with dubious ChatGPT apps

As Macrumors noted, an app named "ChatGPT Chat GPT AI With GPT-3," has managed to reach the top charts in the productivity category in...

NBC News

ChatGPT used by mental health tech app in AI experiment with users

Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?

Siobhan O'Connor [a] [1] ... ChatGPT [b]

JI Translate "I am fixing my hardware while watching a guinea pig show" to Taiwanese Chinese

豚鼠 (Mainland Mandarin) ❌

"我在看豚鼠表演的同時修理我的硬體" 天竺鼠 (Taiwanese Mandarin) ✅

ChatGPT does **not** handle dialects properly.

JI **You**
generate an image showing "2 leaves 2 wallets"

ChatGPT

Here is the image you requested.

ChatGPT does **not** handle multiple objects properly.

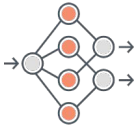
Inspecting Each Stage of Generation

Trustworthiness

Are large language models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of text generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks

ACL 2023 — Dialect-robust Evaluation of Generated Text

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

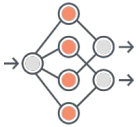
Inspecting Each Stage of Generation

Trustworthiness

Are large language models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of text generation?

- CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages
- EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

- EMNLP 2024 — **Evaluating Large Language Models on Controlled Generation Tasks**
- ACL 2023 — Dialect-robust Evaluation of Generated Text

- EMNLP 2021 — AESOP: Paraphrase Generation with Adaptive Syntactic Control
- EMNLP 2022 — Context Situated Pun Generation
- CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

- ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia
- NeurIPS 2023 — LIMA: Less is More for Alignment
- EMNLP 2022 — Investigate the Benefits of Free-Form Rationales
- EMNLP 2022 — ExPUNations: Augmenting Puns with Keywords and Explanations

Evaluating Large Language Models on Controlled Generation

Are LLMs better than fine-tuned smaller models at controllability on **generation tasks**?

Task	**Control**	**Benchmark**
Content Generation	sentiment, topic, keywords	Amazon Review CommonGen M2D2
Story Generation	prefix	RoC, Writing Prompts
Rationale Generation	correct answer	CoS-E, ECQA
Numerical Planning NEW	prefix, number of words, end words	NPB NEW
Paraphrase Generation	semantics, syntax	ParaNMT, QQP-Pos

Numerical Planning Benchmark

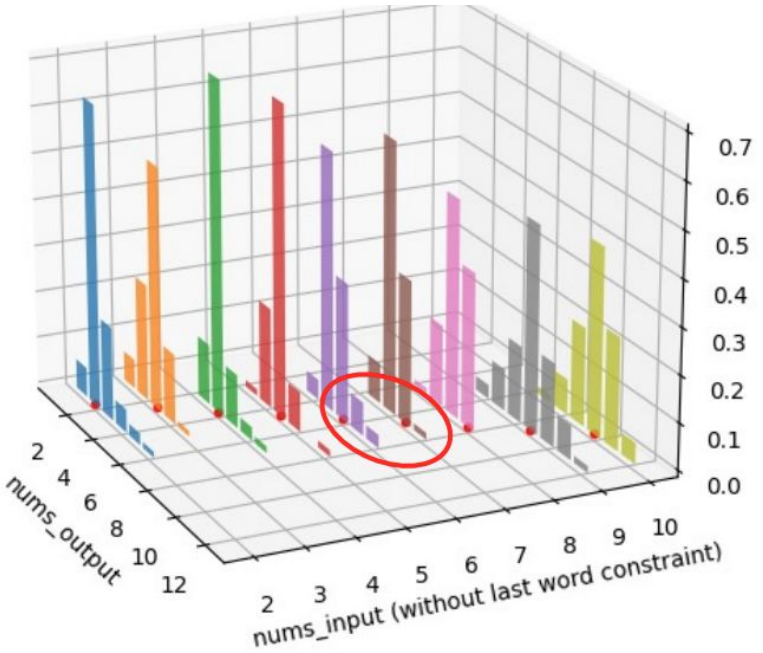
Task Description

- Generate a piece of text that contains a predefined number of **words/sentences/paragraphs** with or without **prefix (start) and suffix (ending)** as constraints

Granularity	Task Illustration
Word/Syllable	Generate a sentence using exactly 5 words/syllables.
	Complete sentence "This is a story" using exactly 5 words/syllables.
	Complete sentence "This is a story" using exactly 5 words/syllables, including the last word as "town".
Sentence	Generate a paragraph with 5 sentences, ...
Paragraph	Generate an article with 5 paragraphs, ...

Success Rates for Word Counting Task

Model	SR - count ↑	SR - last word ↑	SR - both ↑	MSE - count ↓
GPT-2 (fine-tuned)	0.64	0.86	0.60	1.62
Alpaca-7b$_{zs}$	0.17	0.31	0.09	9.19
Alpaca-7b$_{ICL}$	0.14	0.34	0.07	9.76
Vicuna$_{zs}$	0.08	0.09	0.03	27.68
Vicuna$_{ICL}$	0.13	0.30	0.04	13.43
Falcon$_{zs}$	0.13	0.42	0.08	11.60
Falcon-7b$_{ICL}$	0.11	0.34	0.03	13.72
ChatGPT	**0.41**	0.74	**0.36**	**3.64**
ChatGPT$_{ICL}$	0.37	**0.78**	0.34	4.95



zs: zero shot, **ICL**: in-context learning

LLMs < Smaller LMs

LLMs on Content Generation and Story Generation

Model	Topic ↑	Sentiment ↑	Keyword ↑
Diffusion-LM	68.9	83.7	93.2
GPT-2 (1.5B, fine-tuned)	63.4	76.5	88.9
T5 (3B, fine-tuned)	67.3	83.9	94.8
LLaMA-7B$_{zs}$	45.3	58.4	83.5
LLaMA-7B$_{ICL}$	63.5	85.1	93.0
Alpaca-7B$_{zs}$	58.9	78.4	91.2
Alpaca-7B$_{ICL}$	65.2	86.9	94.8
Vicuna-7B$_{zs}$	61.0	80.5	91.6
Vicuna-7B$_{ICL}$	65.8	87.4	94.3
Falcon-7B$_{zs}$	61.9	81.0	92.1
Falcon-7B$_{ICL}$	66.0	87.7	94.2
ChatGPT$_{zs}$	66.4	84.5	97.3
ChatGPT$_{ICL}$	**88.4**	**90.3**	**98.1**

LLMs >> Smaller LMs

LM	Method	rep-2↓	rep-3↓	rep-4↓	diversity↑	coherence↑
			ROC			
	Human	1.74	0.32	0.04	0.97	0.48
GPT-2-XL	Nucleus	1.80	0.35	0.12	0.97	0.33
	Typical	2.06	0.4	0.16	0.97	0.33
	η-sampling	**0**	**0**	**0**	**1.00**	0.34
	SimCTG	3.10	0.46	0.23	0.96	0.32
	Look-back	7.24	0.92	0.14	0.92	0.47
LLM	Vicuna	2.36	0.45	0.15	0.97	0.60
	Falcon	2.52	1.87	1.86	0.94	**0.69**
	ChatGPT	1.18	0.10	0.02	0.98	0.52
			Writing Promts			
	Human	15.61	3.78	1.24	0.80	0.31
GPT-2-XL	Nucleus	5.40	2.41	1.72	0.91	0.34
	Typical	3.60	1.51	1.10	0.94	0.30
	η-sampling	6.17	2.88	2.16	0.89	0.35
	SimCTG	**2.84**	**0.36**	**0.19**	**0.97**	0.31
	Look-back	7.94	1.25	0.33	0.91	0.52
LLM	Vicuna	8.27	2.59	1.14	0.88	0.49
	Falcon	11.20	7.79	6.94	0.76	**0.53**
	ChatGPT	5.99	1.15	0.35	0.92	0.52

LLMs > Smaller LMs

LLMs on Rationale Generation

On Commonsense QA tasks, how much can **adding rationales to the input** can help boost the QA accuracy?

I→O	0.87	
I+R_{CoS-E} →O	0.92	
I+R_{ECQA} →O	**0.99**	
Model	**Leakage** ↑	**Non-Leakage** ↑
I+$R_{Alpaca-7B}$ →O	0.91	0.86
I+$R_{LLaMA-7B}$ →O	0.87	0.79
I+$R_{Vicuna-7B}$ →O	0.95	0.74
I+$R_{Falcon-7B}$ →O	0.83	0.65
I+$R_{ChatGPT}$ →O	**0.98**	**0.93**

LLMs ≈ humans > Smaller LMs

LLMs on Paraphrase Generation

Semantics Preservation

Syntax Conformation

	BLEU↑	METEOR↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	TED-R↓ (H=2)	TED-E↓ (H=2)

LLMs << Smaller LMs

Evaluating Large Language Models on Controlled Generation

Are LLMs better than fine-tuned smaller models at controllability on generation tasks?

Task	Control	Benchmark
Content Generation	sentiment, topic, keywords	Amazon Review CommonGen M2D2
Story Generation	prefix	RoC, Writing Prompts
Rationale Generation	correct answer	CoS-E, ECQA
Numerical Planning	prefix, number of words, end words	NPB
Paraphrase Generation	semantics, syntax	ParaNMT, QQP-Pos

Evaluating Large Language Models on Controlled Generation

Are LLMs better than fine-tuned smaller models at controllability on generation tasks?

Task	Control	Benchmark
Content Generation	sentiment, topic, keywords	Amazon Review CommonGen M2D2
Story Generation	prefix	RoC, Writing Prompts
Rationale Generation	correct answer	CoS-E, ECQA
Numerical Planning	prefix, number of words, end words	NPB
Paraphrase Generation	semantics, syntax	ParaNMT, QQP-Pos

Are Automatic Evaluation Metrics Always Reliable?

“**ChatGPT** is scary good. We are not far from dangerously strong AI.”

“Time it took to reach 1 million users:
Netflix - 3.5 years
Facebook - 10 months
Spotify - 5 months
Instagram - 2.5 months
ChatGPT - 5 days”

TechCrunch

App Store and Play Store are flooded with dubious ChatGPT apps

As Macrumors noted, an app named “ChatGPT Chat GPT AI With GPT-3,” has managed to reach the top charts in the productivity category in...

NBC News

ChatGPT used by mental health tech app in AI experiment with users

Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?

Siobhan O'Connor [a] [1] ... ChatGPT [b]

JI Translate "I am fixing my hardware while watching a guinea pig show" to Taiwanese Chinese

豚鼠 (Mainland Mandarin) ❌

"我在看豚鼠表演的同時修理我的硬體" 天竺鼠 (Taiwanese Mandarin) ✅

ChatGPT does **not** handle dialects properly.

JI You
generate an image showing "2 leaves 2 wallets"

ChatGPT

Here is the image you requested.

ChatGPT does **not** multiple objects properly.

Why is Dialect-Robust Evaluation Important?

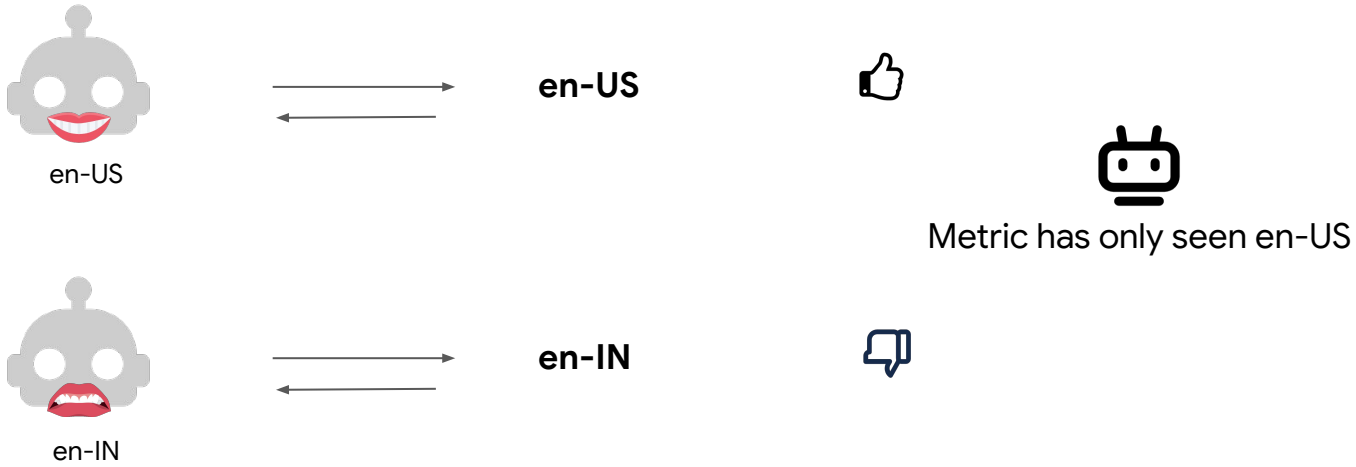
en-US as recently as April there was a big fight

Metric has only seen en-US

en-IN recently only in April there was a big fight

Why is Dialect-Robust Evaluation Important?

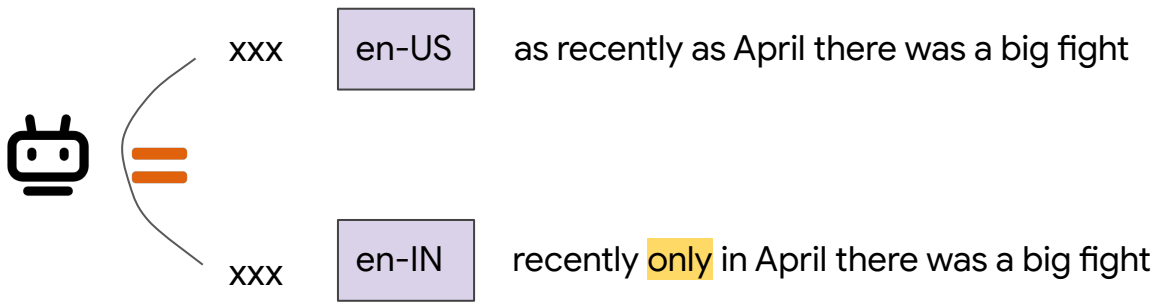
Imagine a generation task where outputting diverse dialects of English is desired



Definition of Dialect Robustness

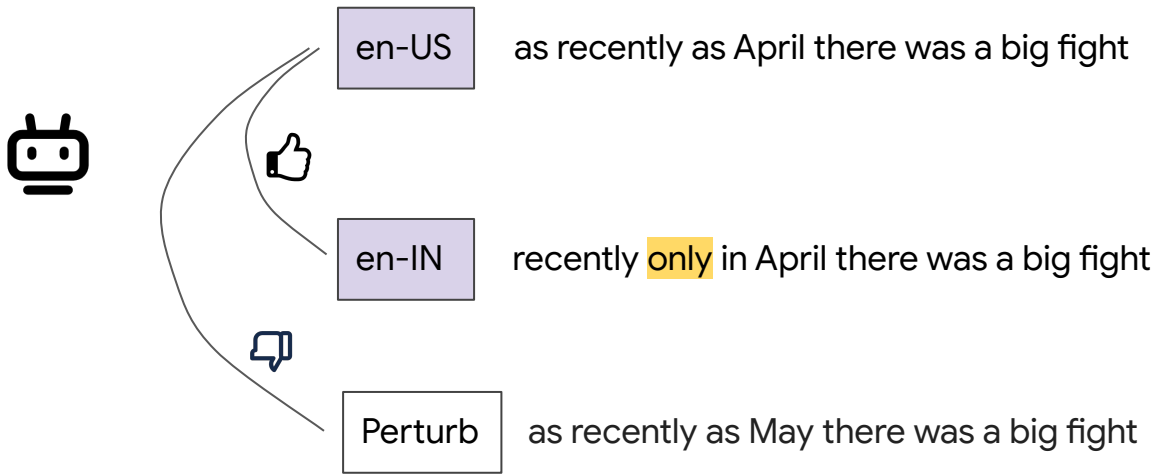
semantically-equivalent texts but in two different dialects (d1, d2), the metric m should assign the two the **same** score m(d1) = m(d2)

TOO STRICT



Relaxed Dialect Robustness

semantically-equivalent dialects (d1, d2)
ϕ is a **semantic perturbation** function
m(d1, d2) **>** m(ϕ(d1), d2)

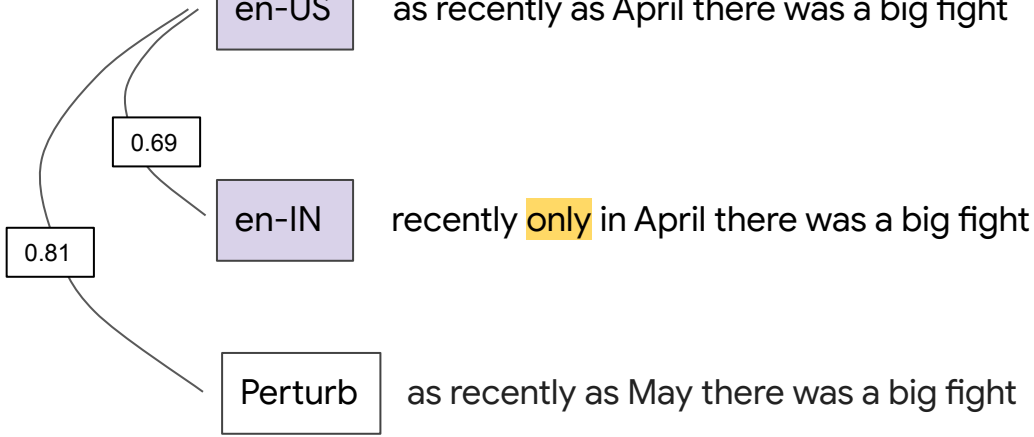


Relaxed Dialect Robustness

semantically-equivalent dialects (d1, d2)
ϕ is a **semantic perturbation** function
m(d1, d2) **>** m(ϕ(d1), d2)

For the same reference, BLEURT thinks perturb is semantically more similar than dialects!

BLEURT



Let's Quantify the Dialect Robustness!

Lexical	BLEU, chrF
Neural	BLEURT, Prism, COMET, YiSi

Comparing the dialect robustness is challenging!

Different scales of all evaluation metrics

e.g., 0.8 BLEURT points v.s. 20 BLEU points

Two Statistical Methods

Dialect-Robust Metric

How much do dialect rewrites outscore semantic perturbations?

Dialect - Semantic Perturb > 0 

How often do dialect rewrites outscore semantic perturbations?

Dialect > Semantic Perturb 

Dataset

en-US as recently as April there was a big fight

en-IN recently only in April there was a big fight

Perturb as recently as May there was a big fight

Token-level dialect rewrite (NAACL 21)*

pt-BR a que horas o restaurante abre para o pequeno-almoço

pt-PT a que horas o restaurante abre para o café da manhã

Perturb a que horas o restaurante abre para o almoço

zh-CN 鳄梨是加利福尼亚州的州果。

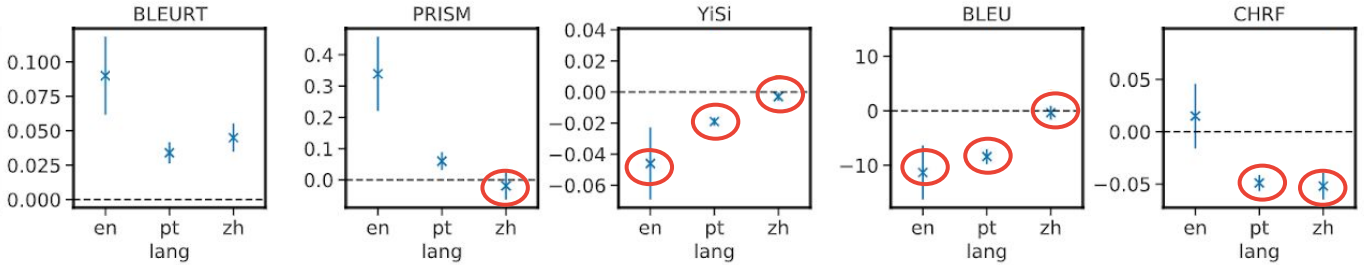
zh-TW 酪梨为加州官方认定的代表水果。

Perturb 鳄梨是佛罗里达州的官方水果。

Sentence-level Dialect Rewrite, FRMT (arxiv 22)+

Are evaluation metrics robust to dialects?

How much do dialect rewrites improve the metric value over semantic perturbations? Dialect - Semantic Perturb



Dialect - Semantic Perturb < 0, metric not robust to dialects

Existing metrics are not robust to dialects, except BLEURT!

Are evaluation metrics robust to dialects?

How often do dialect rewrites score higher than semantic perturbations?

Dialect > Semantic Perturb

	BLEURT	PRISM	YiSi	BLEU	chrF
EN	0.53	0.51	0.53	0.49	0.46
PT	**0.59**	0.53	0.36	0.35	0.35
ZH	**0.59**	0.47	0.46	0.35	0.36

Rates that are greater than a random chance (0.5) and significant!

Existing metrics are not robust to dialects, except BLEURT!

Improve Dialect Robustness by Pretraining

Introduce **NANO**

- NANO is **a pretraining strategy** aims to add dialectal information into training

The **Goals** of NANO

- Improve dialect robustness of a learnt metric
- Not degrade performance on standard metric benchmarks (be a good metric 😁)

NANO: Dialect Acceptability Pretraining

mC4: corpus for mT5 Pretraining

recently only in April there was a big fight

LangID model

Region in **URL**

en

.in

10 langs, 95 dialects

en-US, en-IN, en-GB, ..., en-any

zh-CN, zh-TW, zh-any ....

cs-AT, cs-CZ, cs-any, ...

de-BE, de-NL, de-any, ...

ja-JP, ja-any

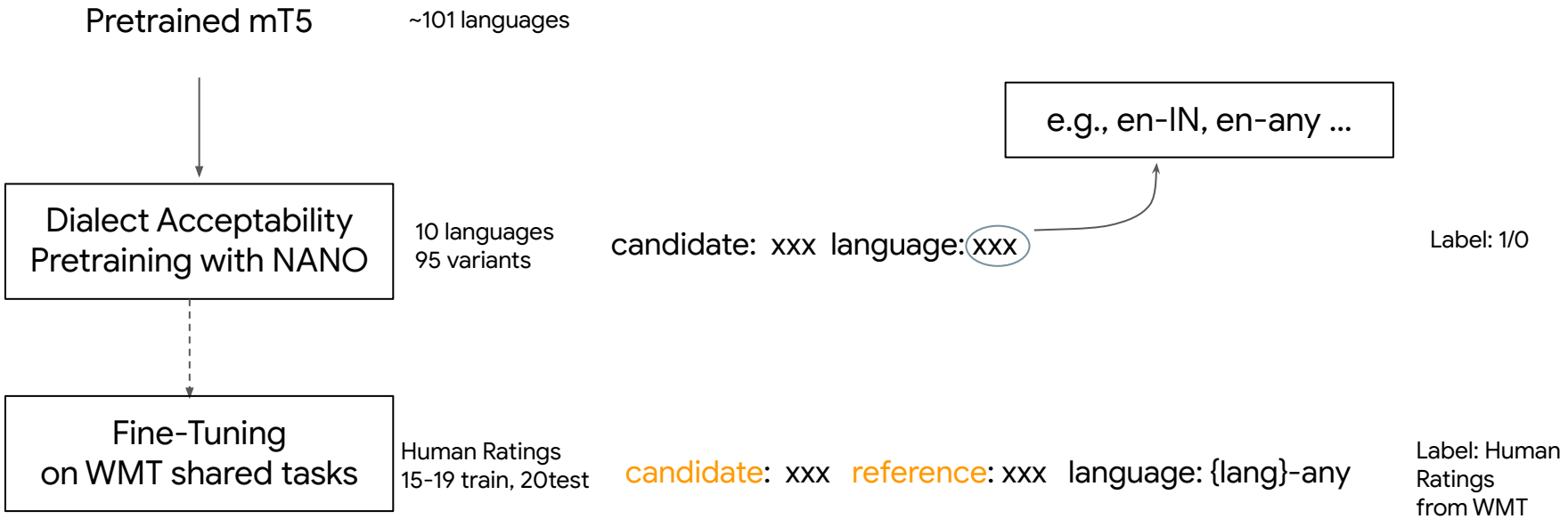
...

Creating Examples:

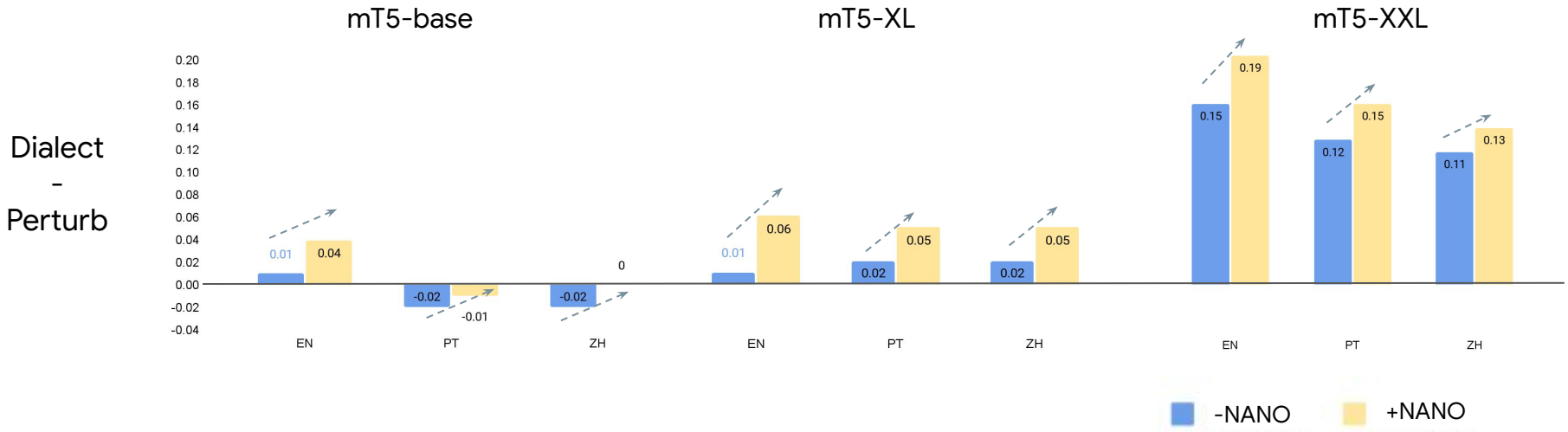
candidate: recently only in April there was a big fight **language:** en-IN **label**: 1

candidate: as recently as April there was a big fight **language**: pt-any **label**: 0

Adapting to WMT Shared Tasks



NANO Improves the Dialect Robustness



Consistent increase of dialect-perturb score across all the languages and all model sizes

NANO improves the dialect robustness!

NANO Improves the Dialect Robustness

mT5-XL mT5-XXL

Dialect > Perturb

	BLEURT	PRISM	YiSi	BLEU	chrF
EN	0.53	0.51	0.53	0.49	0.46
PT	**0.59**	0.53	0.36	0.35	0.35
ZH	**0.59**	0.47	0.46	0.35	0.36

NANO improves the dialect robustness!

More Results Worth Checking Out!

➔ NANO not only improves the dialect robustness, but also metrics' performance on the WMT shared tasks (i.e., from 49.2 to 54.2 with mT5 XL).

➔ NANO generalizes well to all settings: within-language assessment and quality estimation

- Increase the dialect robustness
- Increase performance on WMT benchmark

Takeaways

- We formalize **Dialect Robustness** in the context of NLG evaluation
- We propose a suite of **statistical methods** to test the dialect robustness
- We show that distilling dialectal information from existing corpus (**NANO**) can help improve the dialect robustness

Revisit NANO for Improving Dialect Robustness

Conclusion: Distilling dialectal information from existing corpus (i.e., mT5) can help improve the dialect robustness of the evaluation metric/model

💡 **Idea**

Can we improve large models solely from better data utilization?

Evidence: Better Rationales -> Better Interpretability + Model Acc

Data Quality

Better **data** for better utility of text generation?

EMNLP 2022 Investigate the Benefits of Free-Form Rationales

Question	Where would you find a monkey in the wild?
Options	zoo, barrel, research laboratory, captivity, **thailand**
CoS-E	thailand find a monkey in the wild
ECQA	All the other options are incorrect as they are not a wild place. In thailand, monkeys can be found in the wild.

Crowdsourced

Human-Perceived Usefulness of Rationales

Do crowdsourced rationales aid human interpretability?

Are rationales simply **leak**ing the correct answer?

Do rationales contain additional **background knowledge** necessary to answer the question?

❓ Choose one to explain to a child: $\mathbf{R}_{\text{crowd}}$ $\mathbf{R}_{\text{construct}}$ neither either

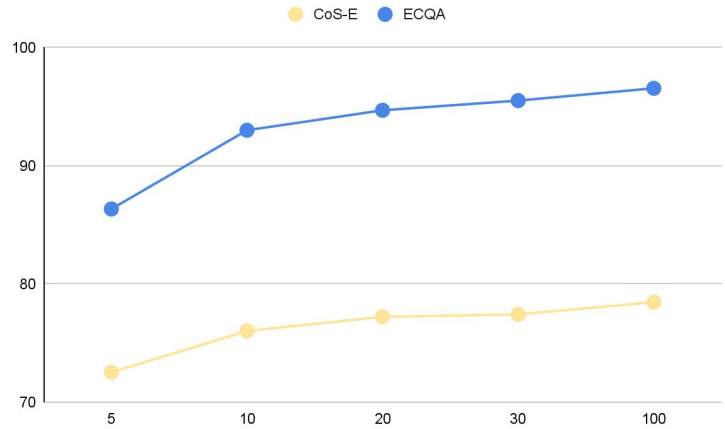
	$\mathbf{R}_{\text{crowd}}$	$\mathbf{R}_{\text{construct}}$	**neither**	**either**
CoS-E	3.0%	5.0%	**92.0%**	0.0%
ECQA	73.0%	9.0%	14.0%	4.0%

Rationales can aid human interpretability! But only the good-quality ones!

Can Rationales Help Improve the Model Accuracy?

Without Rationales: model accuracy is 57.0

Training: we add rationales from ECQA or CoS-E



Models trained with better rationales -> higher accuracy

Evidence: Better Rationales -> Better Interpretability + Model Acc

💡 Idea

Can we improve large models solely from better data utilization?

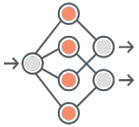
Inspecting Each Stage of Generation

Trustworthiness

Are large language models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of text generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2022 — Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts

ACL 2023 — Dialect-robust Evaluation of Generated Text

EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

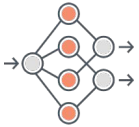
Inspecting Each Stage of Generation

Trustworthiness

Are large language models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models, and where to get good data**?

Data Quality

Better **data** for better utility of text generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2022 — Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts

ACL 2023 — Dialect-robust Evaluation of Generated Text

EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

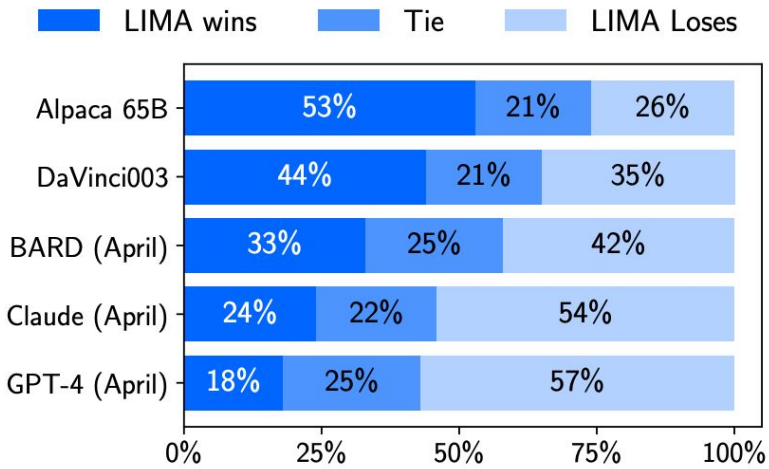
LIMA: Less is More for Alignment (Data Quality > Quantity)

❗ **Only limited instruction tuning data (~1000) is necessary to teach models to produce high quality output**

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

Table 1: Sources of training prompts (inputs) and responses (outputs), and test prompts. The total amount of training data is roughly 750,000 tokens, split over exactly 1,000 sequences.

Evaluation



Human preference evaluation, comparing LIMA to five different baselines across 300 test prompts

“**ChatGPT** is scary good. We are not far from dangerously strong AI.”

“Time it took to reach 1 million users:
Netflix - 3.5 years
Facebook - 10 months
Spotify - 5 months
Instagram - 2.5 months
ChatGPT - 5 days”

TechCrunch

App Store and Play Store are flooded with dubious ChatGPT apps

As Macrumors noted, an app named “ChatGPT Chat GPT AI With GPT-3,” has managed to reach the top charts in the productivity category in...

NBC News

ChatGPT used by mental health tech app in AI experiment with users

Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?

Siobhan O'Connor [a] [1] ... ChatGPT [b]

JI Translate "I am fixing my hardware while watching a guinea pig show" to Taiwanese Chinese

豚鼠 (Mainland Mandarin) ✗

"我在看豚鼠表演的同時修理我的硬體" 天竺鼠 (Taiwanese Mandarin) ✅

ChatGPT does **not** handle dialects properly.

JI **You**
generate an image showing "2 leaves 2 wallets"

ChatGPT

Here is the image you requested.

ChatGPT does **not** multiple objects properly.

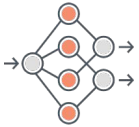
Inspecting Each Stage of Practical Text Generation

Trustworthiness

Are large language models free of gender biases?

Evaluation

How to **evaluate** models automatically and fairly?



Modeling

How can we build better generation **models**?

Data Quality

Better **data** for better utility of text generation?

CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

EMNLP 2022 — Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts

ACL 2023 — Dialect-robust Evaluation of Generated Text

EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks

NeurIPS 2023 — LIMA: Less is More for Alignment

CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

EMNLP 2022 — Investigate the Benefits of Free-Form Rationales

Text-to-Image Generation Models Do NOT Align with Human Input Well

Take SDXL as an example

An apple sitting on a chair made of marshmallow, in a park, at sunrise

A white cat with black ears and markings

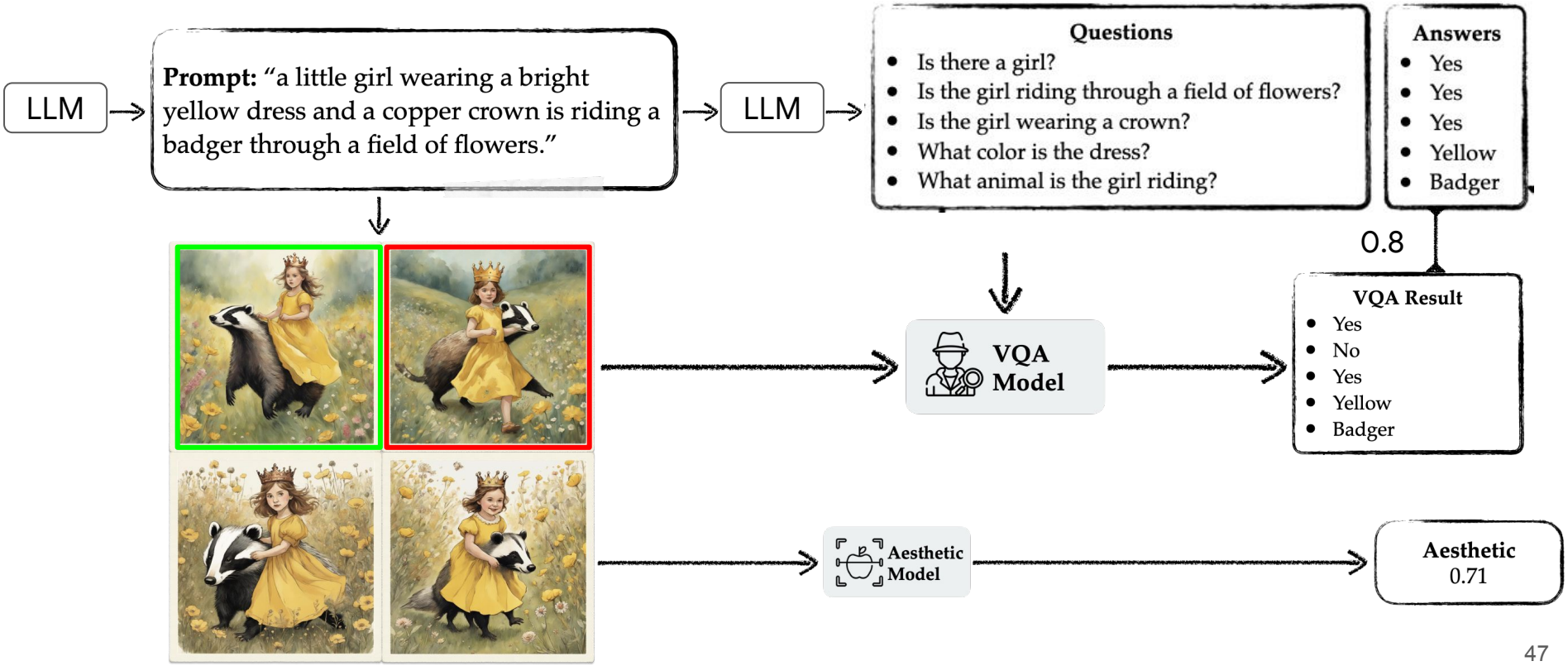
One wooden heart and one marble heart

Topographical letters contour made of a layered paper, muted pastel colors

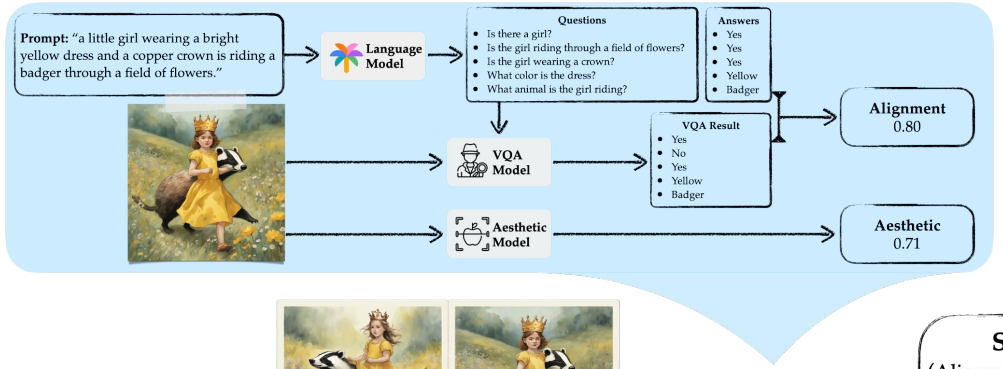
Goal

Improve the text-to-image generation alignment while maintaining the aesthetic appeal

Training Data Preparation and Rejection Sampling



DreamSync



Prompt: "a little girl wearing a bright yellow dress and a copper crown is riding a badger through a field of flowers."

T2I Model

Scoring

Scores
(Alignment, Aesthetic)

(1.00, 0.80)	(0.80, 0.71)
(0.80,0.73)	(1.00, 0.68)

Select Best

LoRA Finetune

Evaluation

Goal: Improve the text-to-image generation alignment while maintaining the aesthetic appeal

- We use LLM-generated QA pairs vs. VQA result to measure the text faithfulness

    -> Mean & absolute scores

- We use aesthetics model to measure the visual appearance

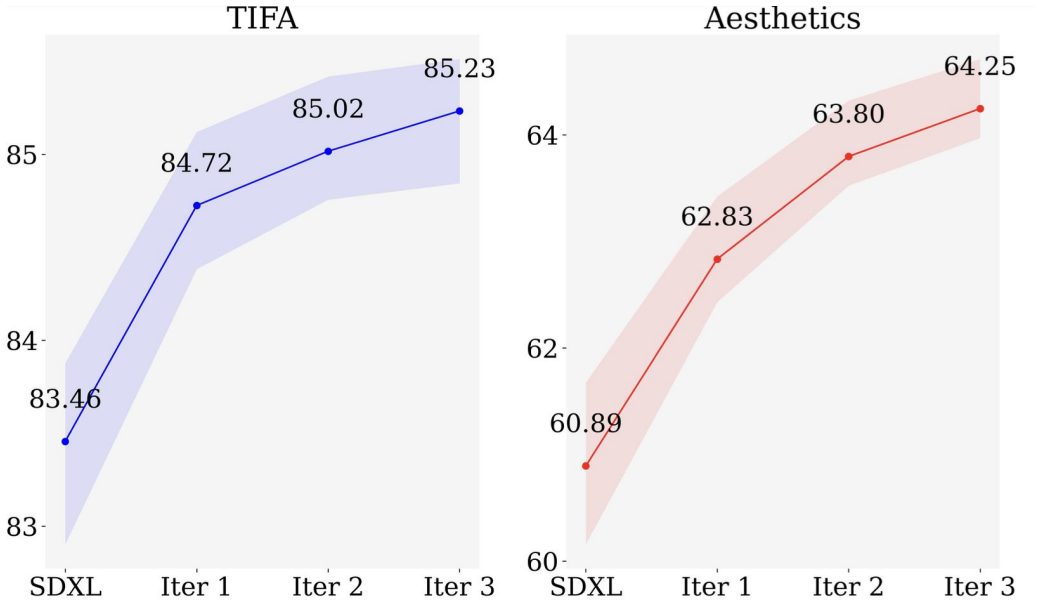
Quantitative Results

datasets

Model	Alignment		Text Faithfulness			Visual Appeal
			TIFA		DSG1K	
			Mean	*Absolute*		
SD v1.4		No alignment	76.6	33.6	72.0	44.6
	Training-Free	SynGen	76.8 (+0.2)	34.1 (+0.5)	71.2 (–0.8)	42.4 (–2.2)
		StructureDiffusion	76.5 (–0.1)	33.6 (+0.0)	71.9 (–0.1)	41.5 (–3.1)
	RL	DPOK	76.4 (–0.2)	33.8 (+0.2)	70.3 (–1.7)	**46.5** (+1.9)
		DDPO	76.7 (+0.1)	34.4 (+0.8)	70.0 (–2.0)	43.5 (–1.1)
		DreamSync (ours)	**77.6** (+1.0)	**35.3** (+1.7)	**73.2** (+1.2)	44.9 (+0.3)

methods

Quantitative Results



DreamSync improves faithfulness and aesthetics iteratively

Let’s see some pictures!

Text-to-Image Generation Models Do NOT Align with Human Input Well

Take SDXL as an example

An apple sitting on a chair made of marshmallow, in a park, at sunrise

A white cat with black ears and markings

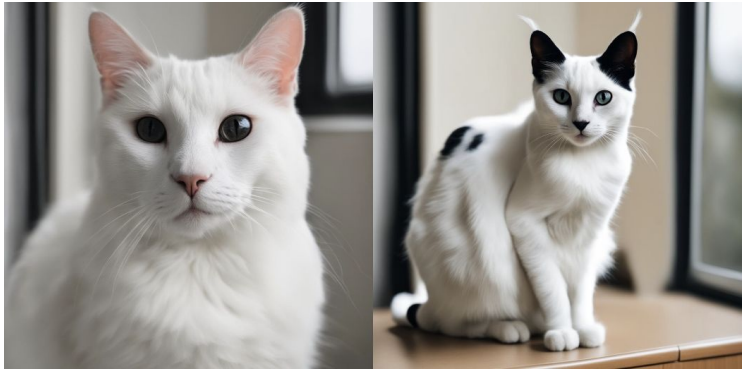
One wooden heart and one marble heart

Topographical letters contour made of a layered paper, muted pastel colors

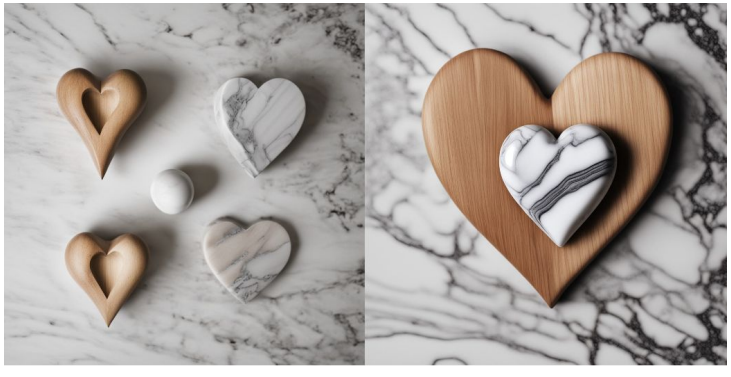
Qualitative Examples of DreamSync Improving Text-Image Alignment

An apple sitting on a chair made of marshmallow, in a park, at sunrise

A white cat with black ears and markings

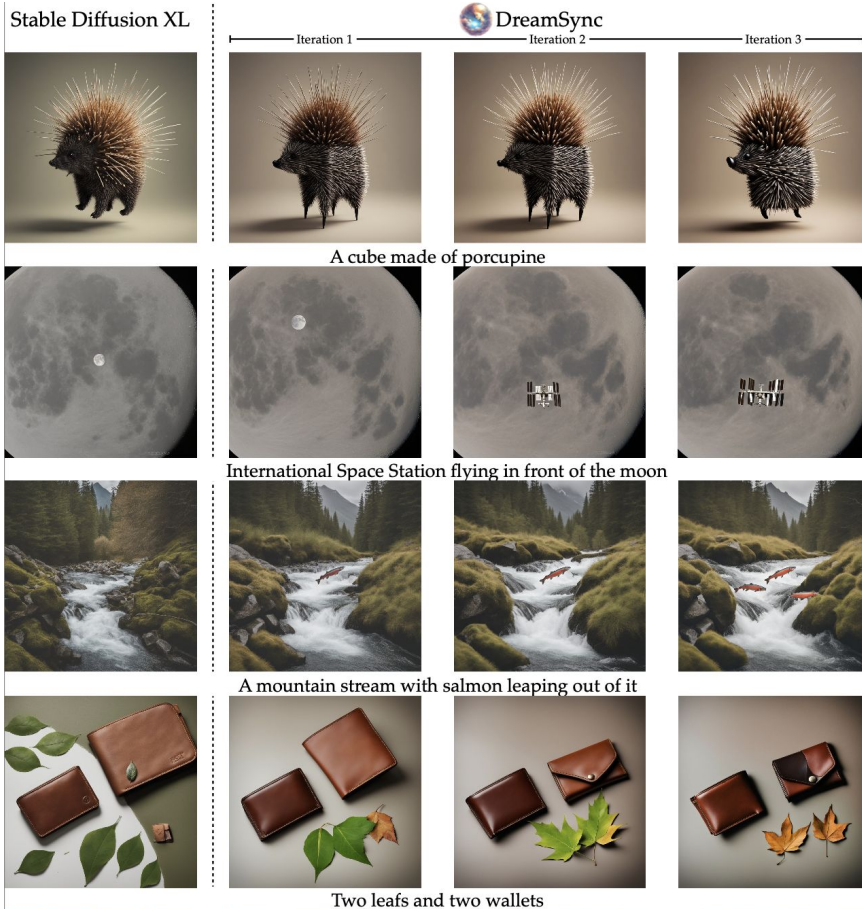


One wooden heart and one marble heart



Topographical letters contour made of a layered paper, muted pastel colors

Qualitative Examples of DreamSync Improving Iteratively



Conclusion

- We introduce **DreamSync, a model-agnostic framework** to improve text-to-image synthesis with feedback from image understanding models

- With two text-to-image models as backbone, we show DreamSync improves both their **alignment and aesthetics** on two benchmarks

Inspecting Each Stage of Practical Text Generation

Trustworthiness

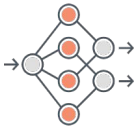
Are large language models free of gender biases?

Greeting Card Messages & Recommendation Letters → Not yet :(

Evaluation

How to **evaluate** models automatically and fairly?

Dialect-Robust Evaluation & Evaluating Control Gen → Fair and robust evaluation of generated text



Modeling

How can we build better generation **models**?

Dreamsync & LIMA ↘

Data Quality

Better **data** for better utility of text generation?

Event Bias in Wikipedia & Free-Form Rationales ↗

Better generative models requires better data quality

Inspecting Each Stage of Practical Text Generation

Trustworthiness

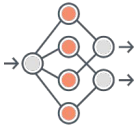
Are large language models free of gender biases?

- CHI 2022 — Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages
- EMNLP 2024 — Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

Evaluation

How to **evaluate** models automatically and fairly?

- EMNLP 2022 — Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts
- ACL 2023 — Dialect-robust Evaluation of Generated Text
- EMNLP 2024 — Evaluating Large Language Models on Controlled Generation Tasks



Modeling

How can we build better generation **models**?

- NeurIPS 2023 — LIMA: Less is More for Alignment
- CVPR 2024 Submission — DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback
- EMNLP 2021 — AESOP: Paraphrase generation with adaptive syntactic control
- EMNLP 2022 — Context-situated pun generation

Data Quality

Better **data** for better utility of text generation?

- ACL 2021 — Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia
- EMNLP 2022 — Investigate the Benefits of Free-Form Rationales
- EMNLP 2022 — ExPUNations: Augmenting puns with keywords and explanations

Ucla

Bloomberg

IBM

CMU