Rigorous measurement in text-to-image systems

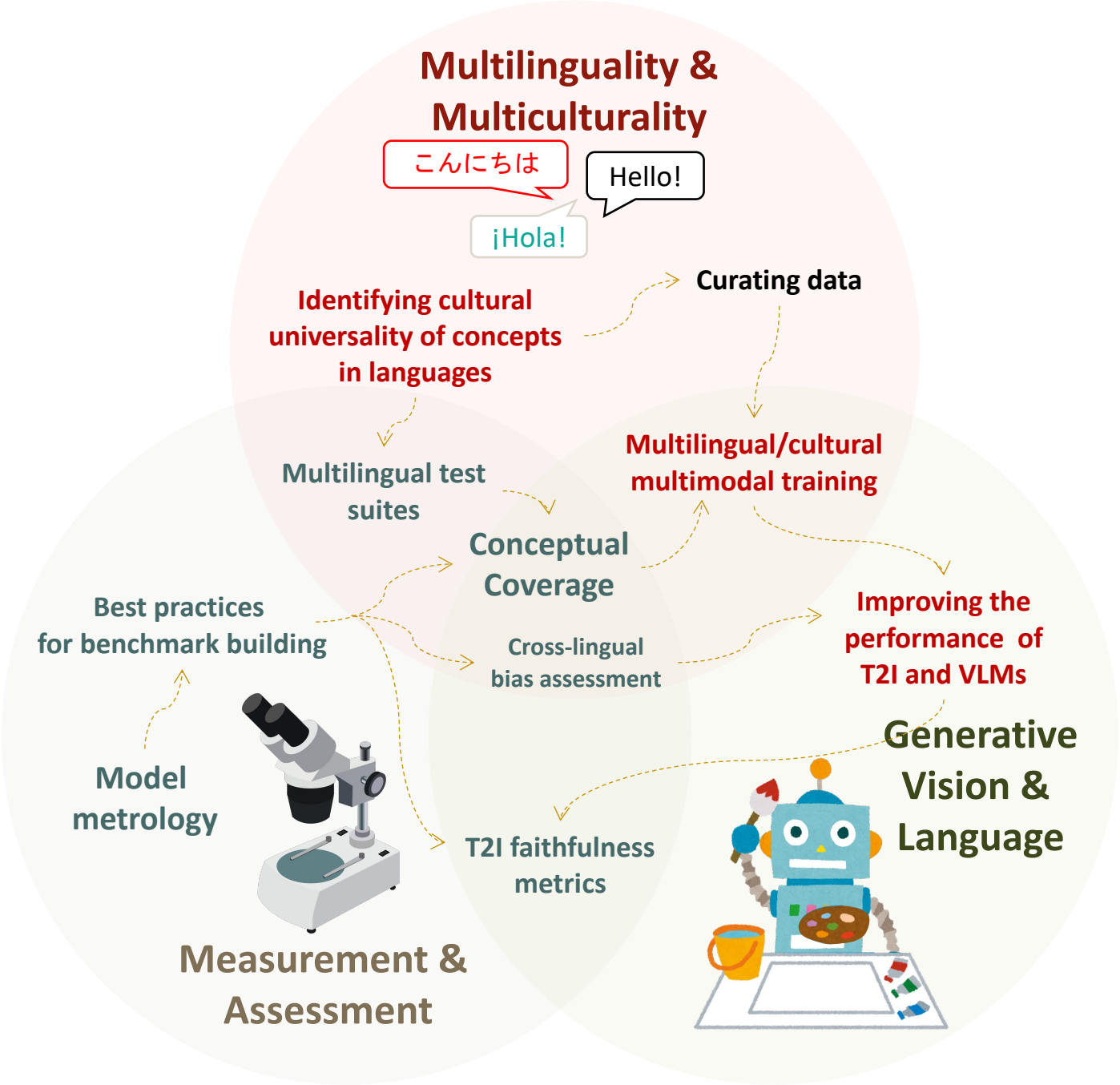
Michael Saxon
Computer Science
University of California, Santa Barbara
https://saxon.me/

UC SANTA BARBARA

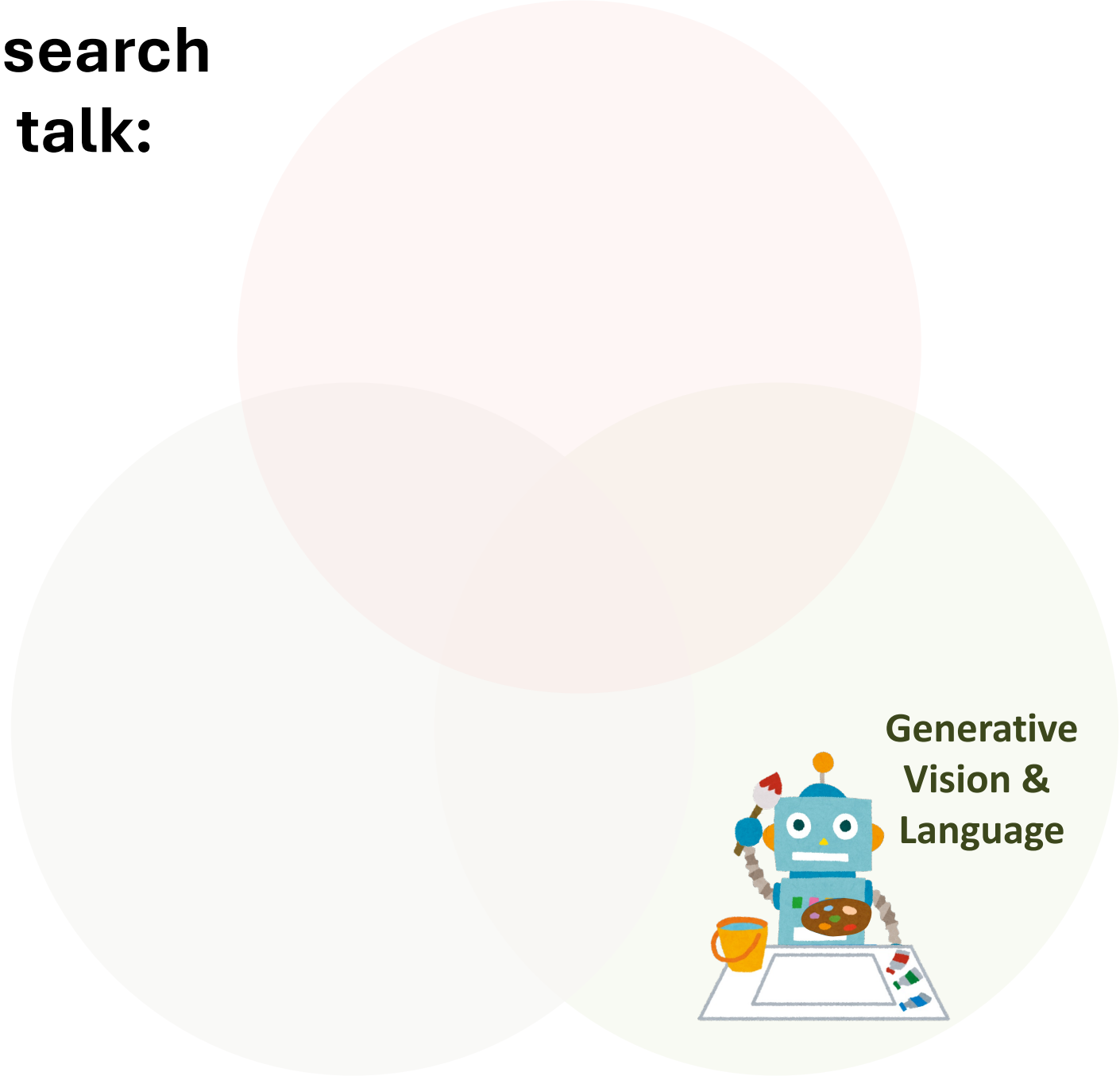
Rigorous measurement in text-to-image systems *(and AI more broadly?)*

Michael Saxon
Computer Science
University of California, Santa Barbara
https://saxon.me/

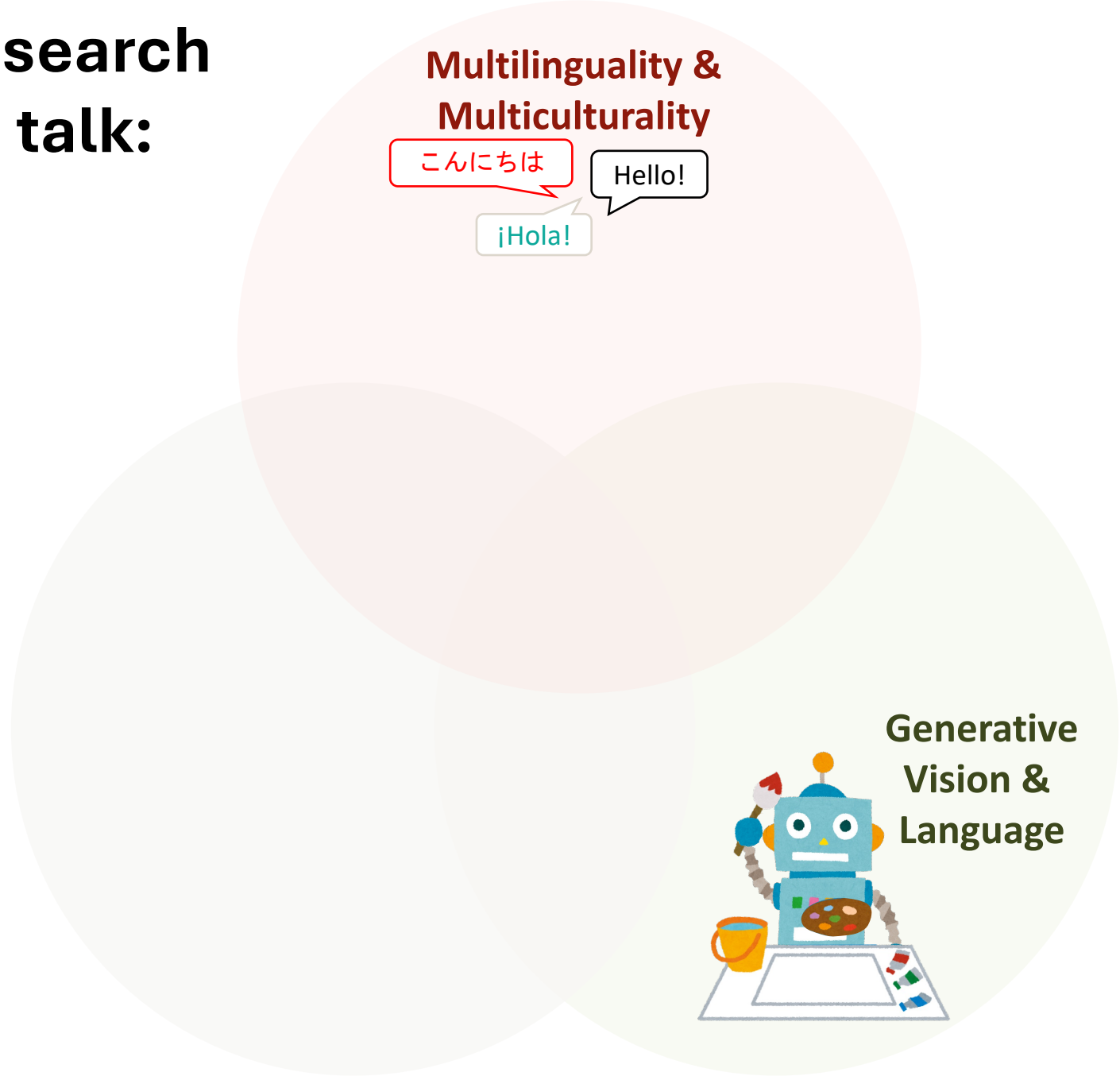
UC SANTA BARBARA


Multilinguality & Multiculturality
こんにちは
Hello!
¡Hola!
Identifying cultural universality of concepts in languages
Curating data
Multilingual test suites
Multilingual/cultural multimodal training
Conceptual Coverage
Best practices for benchmark building
Cross-lingual bias assessment
Improving the performance of T2I and VLMs
Model metrology
Generative Vision & Language
T2I faithfulness metrics
Measurement & Assessment

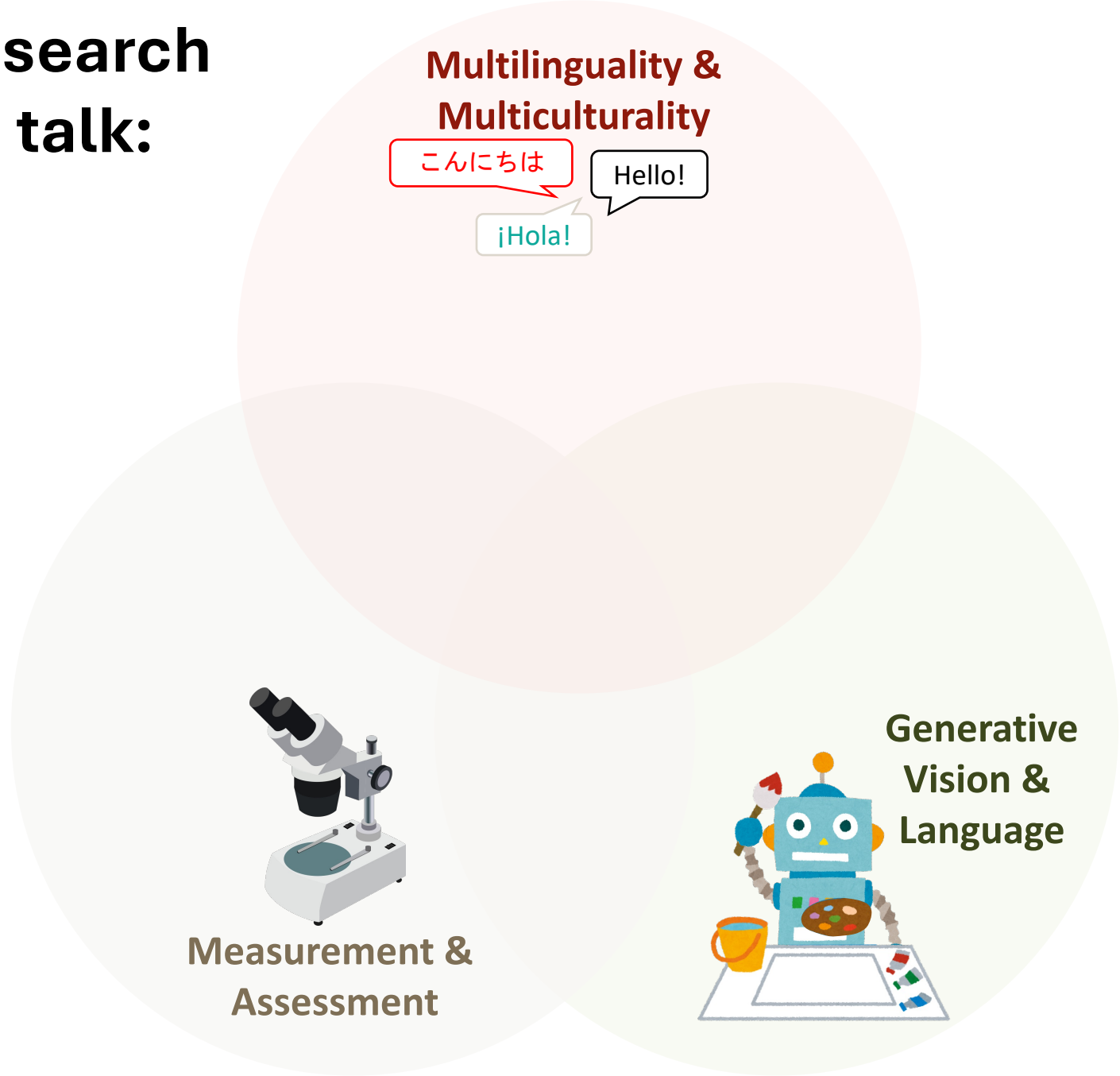
Three core research areas for this talk:



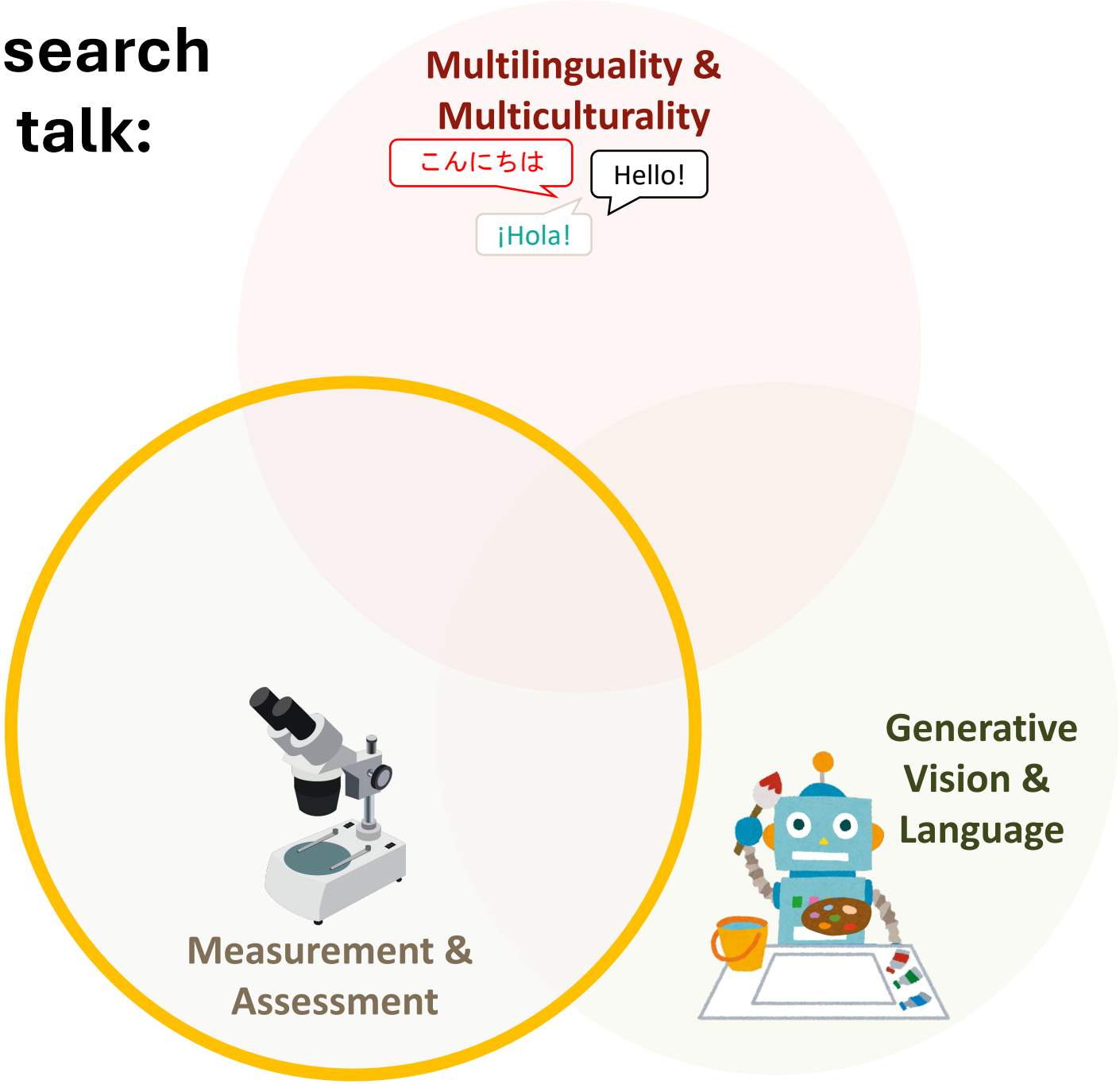
Three core research areas for this talk:



Three core research areas for this talk:



Three core research areas for this talk:



Measurement & Assessment

What’s missing in AI science?

The current state of AI science

- Not good!
- We're talking about fundamental issues like general intelligence and risks
- Answering these questions matters: **public is watching**

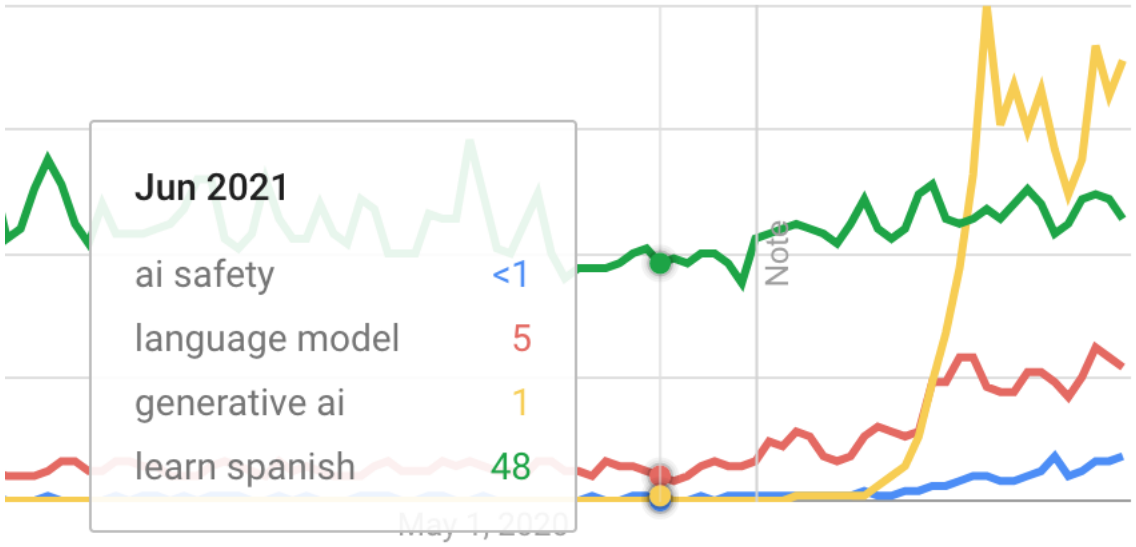
OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to



The current state of AI science

- **Fundamental disagreement between experts** on questions that matter!
- Do models "understand" language? Will current methods ever?

Yann LeCun: Machine learning is great. But the idea that somehow we're going to just scale up the techniques that we have and get to human-level AI? No. We're missing something big to get machines to learn efficiently, like humans and animals do. We don't know what it is yet.

The current state of AI science

- **Fundamental disagreement between experts** on questions that matter!
- Do models "understand" language? Will current methods ever?

The current state of AI science

- **Fundamental disagreement between experts** on questions that matter!
- Do models "understand" language? Will current methods ever?

Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

"A system trained on form alone has a-priori no way to learn meaning"

The current state of AI science

- **Fundamental disagreement between experts** on questions that matter!
- Do models "understand" language? Will current methods ever?

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

"GPT-4 can solve novel and difficult tasks…without special prompting"

The current state of AI science

- **Sparks of AGI: GPT-4 can draw unicorns!!!**

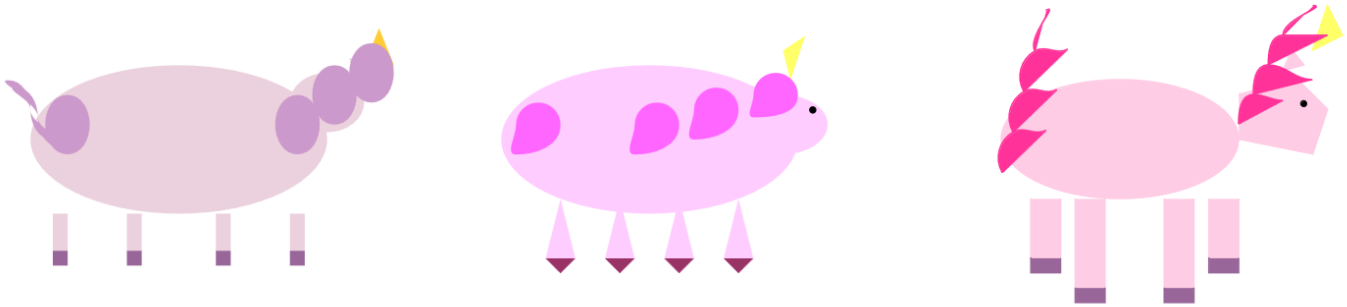


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt "Draw a unicorn in TikZ". We can see a clear evolution in the sophistication of GPT-4's drawings.

"This is another challenging task that requires combining **visual imagination** and coding skills"

The current state of AI science

- **We might be underestimating just how long the tail is…**

"The duck pond": showcase of TikZ-drawn animals/ducks

Asked 6 years, 8 months ago Modified 10 months ago Viewed 18k times

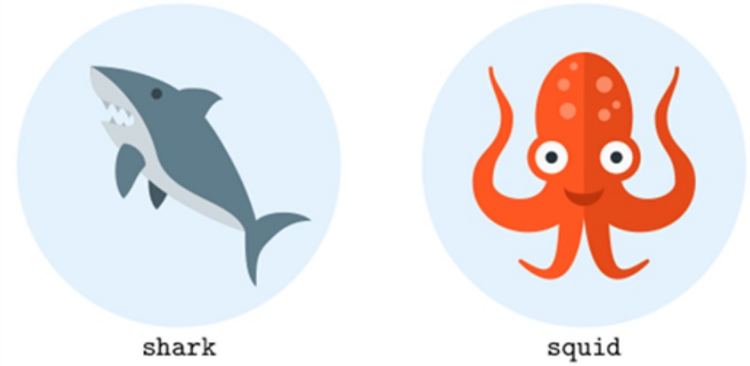
81

We have tons of nice TikZ-drawn pictures on this site. Among them some great pictures of animals like cfr's cat code. But they are distributed all over the site.

As we definitely miss a place to collect them, this post asks about nice examples of animals/pets drawn with TikZ. Especially those pictures about **ducks** are welcome.

The current state of AI science

- **We might be underestimating just how long the tail is…**




```
\documentclass[tikz,border=5]{standalone}
\usepackage[prefix=]{xcolor-material}
\pgfmathdeclarerandomlist{material}{%
{Red}{Blue}{Green}}
\tikzset{%
  half clip/.code={
    \clip (0, -256) rectangle (256, 256);
  },
  color/.code=\colorlet{fill color}{#1},
  color alias/.code args={#1 as #2}{\colorlet{#1}{#2}},
  colors alias/.style={color alias/.list/.expanded={#1}},
  execute/.code={#1},
  on left/.style={.. on left/.style={#1}},
  on right/.style={.. on right/.style={#1}},
  split/.style args={#1 and #2}{
    on left ={color alias=fill color as #1},
    on right={color alias=fill color as #2, half clip}
  }
```

The current state of AI science

- **Let's think step by step!**

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei **Xuezhi Wang** **Dale Schuurmans** **Maarten Bosma**
Brian Ichter **Fei Xia** **Ed H. Chi** **Quoc V. Le** **Denny Zhou**

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

"although chain of thought emulates the thought processes of human reasoners, this does not answer whether the neural network is actually 'reasoning,'"

The current state of AI science

- **Let's think long-tail….**

Yanai Elazar
@yanaiela

Let's think "step by step"!

Another tidbit I like about data and prompts that miraculously
Searching for this phrase resulted in this website (among othe
geteasysolution.com, containing many math step-by-step sol
How common are they? Quite.

Makes you think.

What number is 1000 more than 2728 - solution with step by step explanation

Below you can find the full step by step solution for you problem. We hope it will be very helpful for you and it will help you to understand the solving process.

If it's not what You are looking for, type integer numbers into the boxes below and see the solution.

What number is [ ] more than [ ] Solve

What is 1000 more than 2728?

We are looking for a new number which is 1000 more than 2728.
We will get the new number by adding 1000 to 2728.
We write it down as:
2728+1000=3728

And finally the solution for: What number is 1000 more than 2728?
is **3728**

The current state of AI science

- Le

Yar
@ya

Let's thinl

Another t
Searching
geteasysc
How com

Makes yo

Corpora

☑ C4 ☑ mC4-en ☑ OSCAR ☑ RedPajama ☑ LAION-2B-en ☑ Dolma

Domain to Lookup

www.geteasysolution.com

Search

Domain	Corpus	Rank	Tokens	% of All Tokens
www.geteasysolution.com	Dolma	151022	3,549	0.000081%
www.geteasysolution.com	OSCAR	277233	224,965	0.000045%
www.geteasysolution.com	C4	473082	49,859	0.000032%
www.geteasysolution.com	RedPajama	472159	49,859	0.000023%
www.geteasysolution.com	mC4-en	1658921	156,174	0.0000056%

step by step

ul for you and it will

tion.

The current state of AI science

- To summarize, a huge amount of noise is being made over these extremely fundamental high-level ideas
- Everyone (experts included) bring prior opinions to the party
- Many "discoveries" are effectively **tribal knowledge**
  - We know that *X* might work well, but we really don't know why
- Attempts to build theory from those discoveries rarely survive new discoveries and competing interpretations
- **Fundamental impasse?**

AI “science” or alchemy?

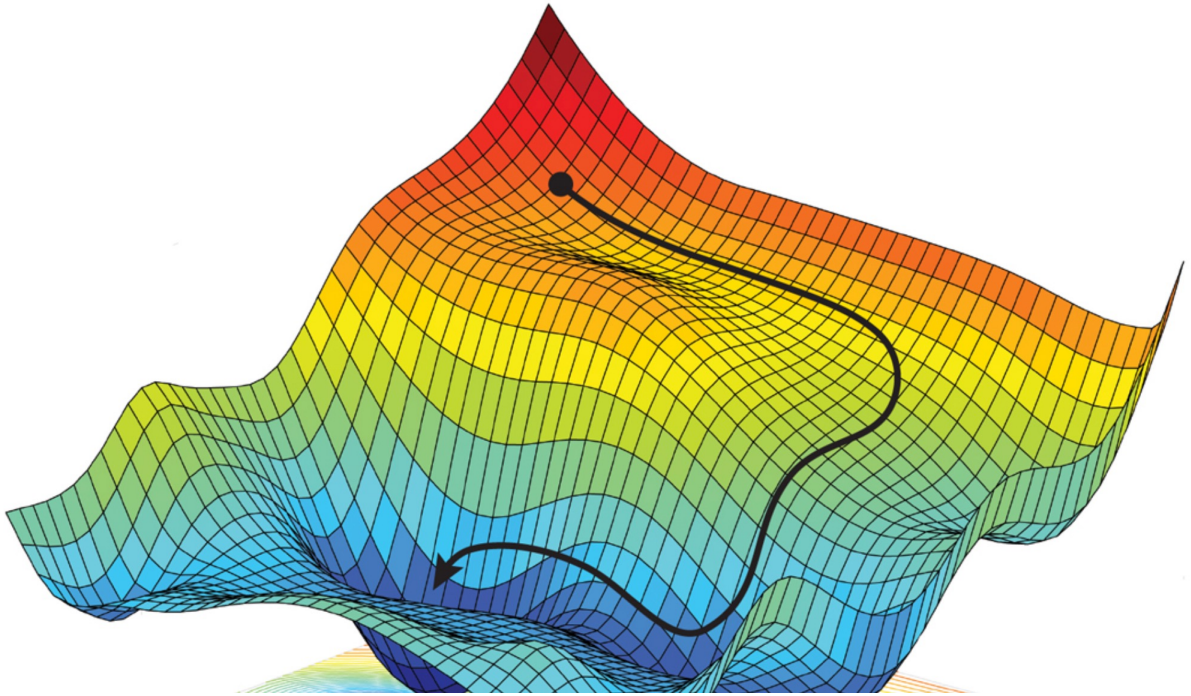
Without deep understanding of the basic tools needed to build and train new algorithms, he says, researchers creating AIs resort to hearsay, like medieval alchemists. “People gravitate around cargo-cult practices,” relying on “folklore and magic spells,” adds François Chollet, a computer scientist at Google in Mountain View, California. For example, he says, they adopt pet methods

AI “science” or alchemy?

AI researchers allege that machine learning is alchemy

Study cites ways to bolster scientific foundations of artificial intelligence

3 MAY 2018 · BY MATTHEW HUTSON



AI “science” or alchemy?

ALCHEMY AND ARTIFICIAL INTELLIGENCE

Hubert L. Dreyfus

December 1965

What can we do about this?

Let's look at how "real" sciences have advanced

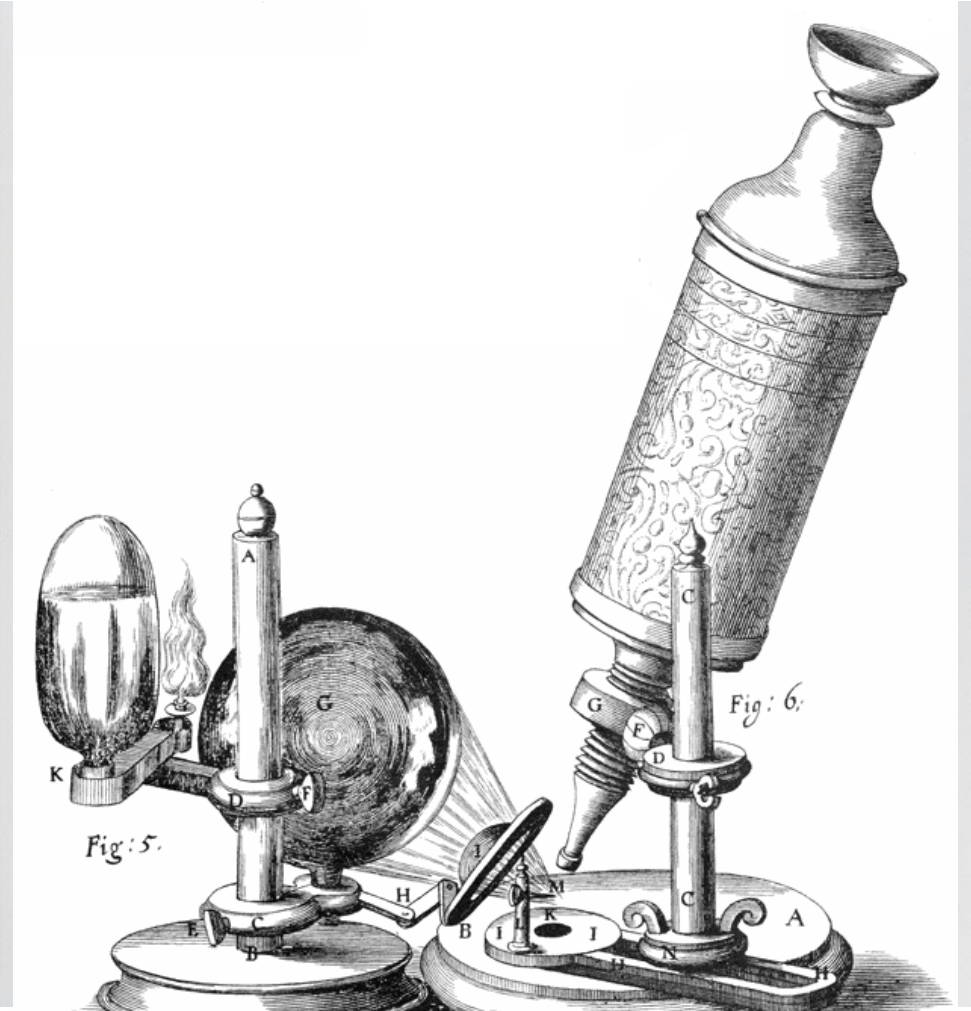
- Tight coupling of **advances in measurement** and **advances in theory**
- Galileo's telescope among the earliest built in 1609
- Enabled him to observe sun-spots, moons of Jupiter
- This advance in measurement helped confirm that the Earth orbits the Sun

More measurement and theory connections

- By reversing his telescope Galileo produced one of the earliest microscopes

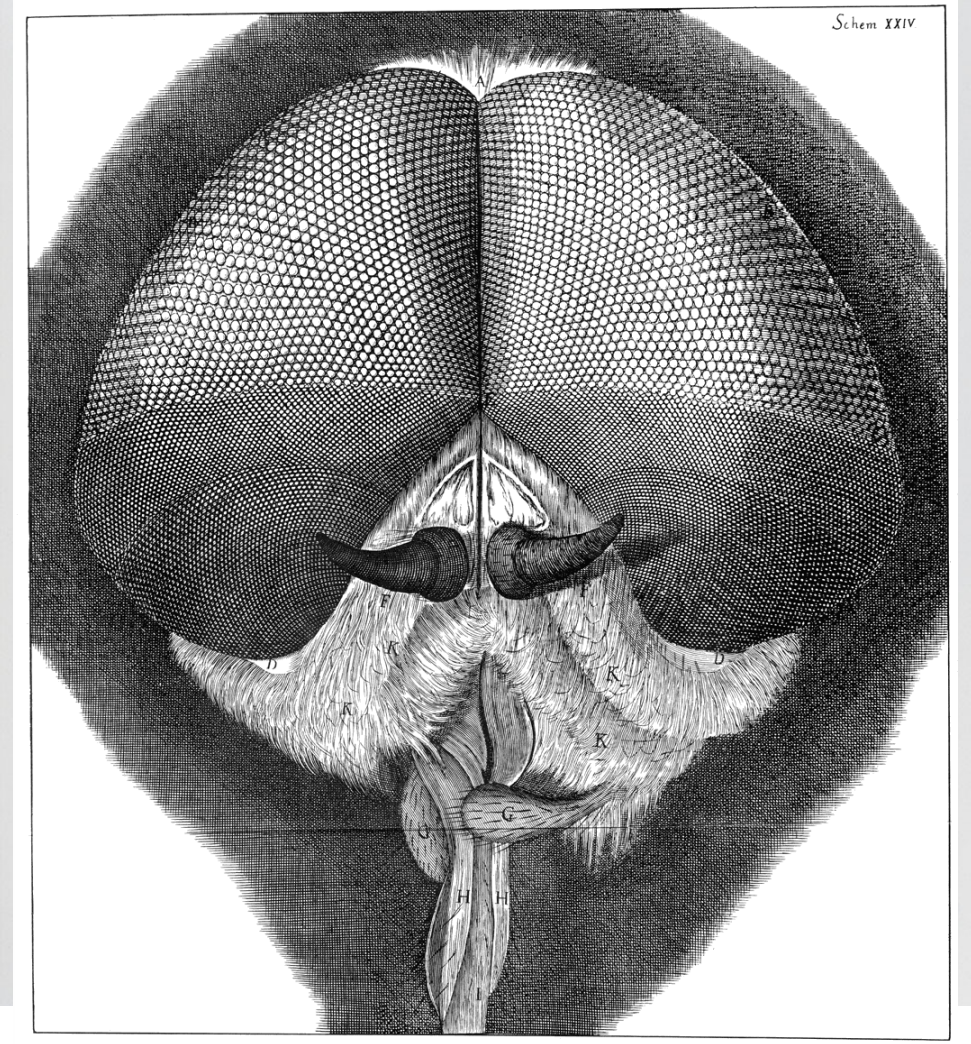
More measurement and theory connections

- By reversing his telescope Galileo produced one of the earliest microscopes
- Polymath Robert Hooke advanced on the design in 1665, birthing microbiology



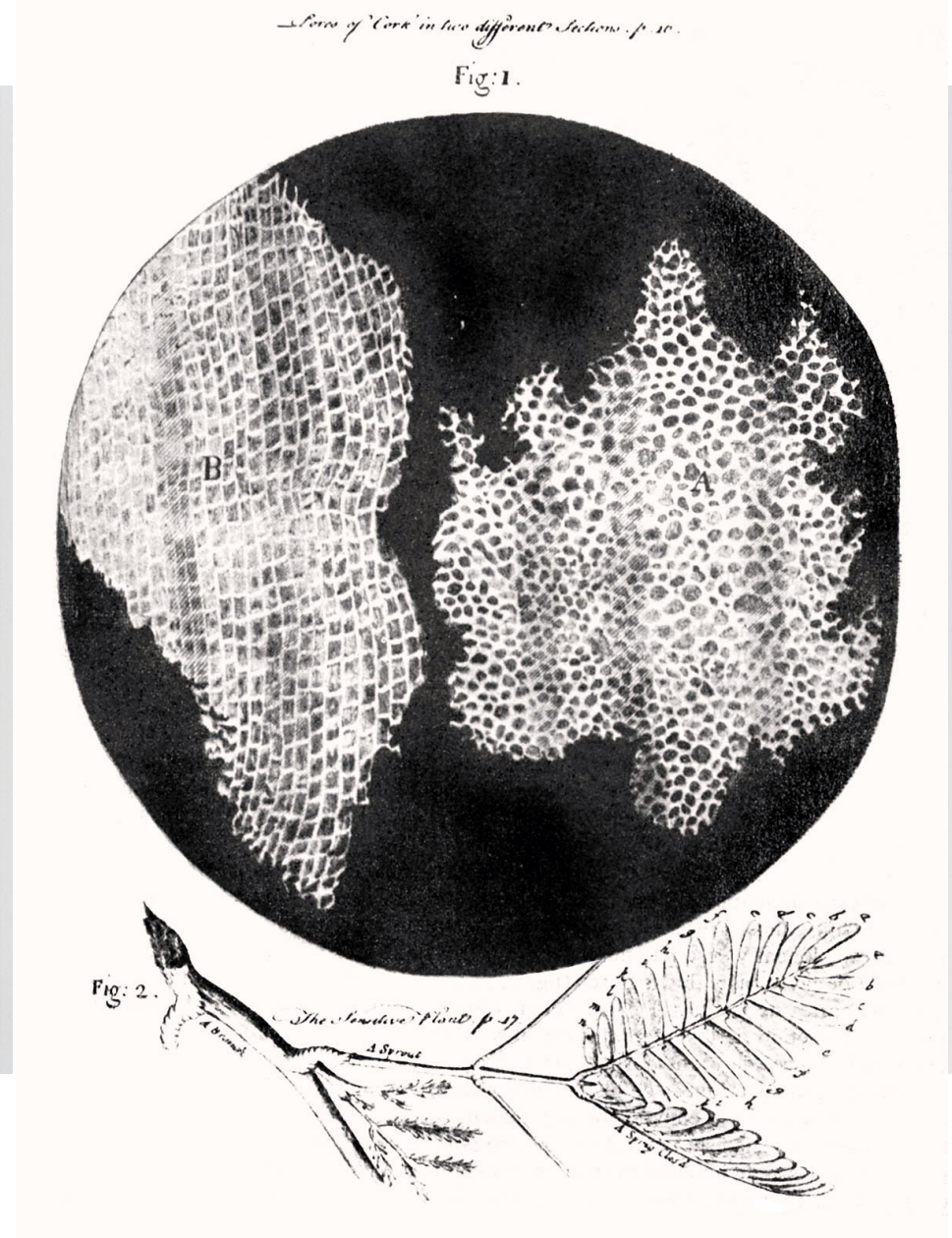
More measurement and theory connections

- By reversing his telescope Galileo produced one of the earliest microscopes
- Polymath Robert Hooke advanced on the design in 1665, birthing microbiology
- Hooke was the first to document tiny insects



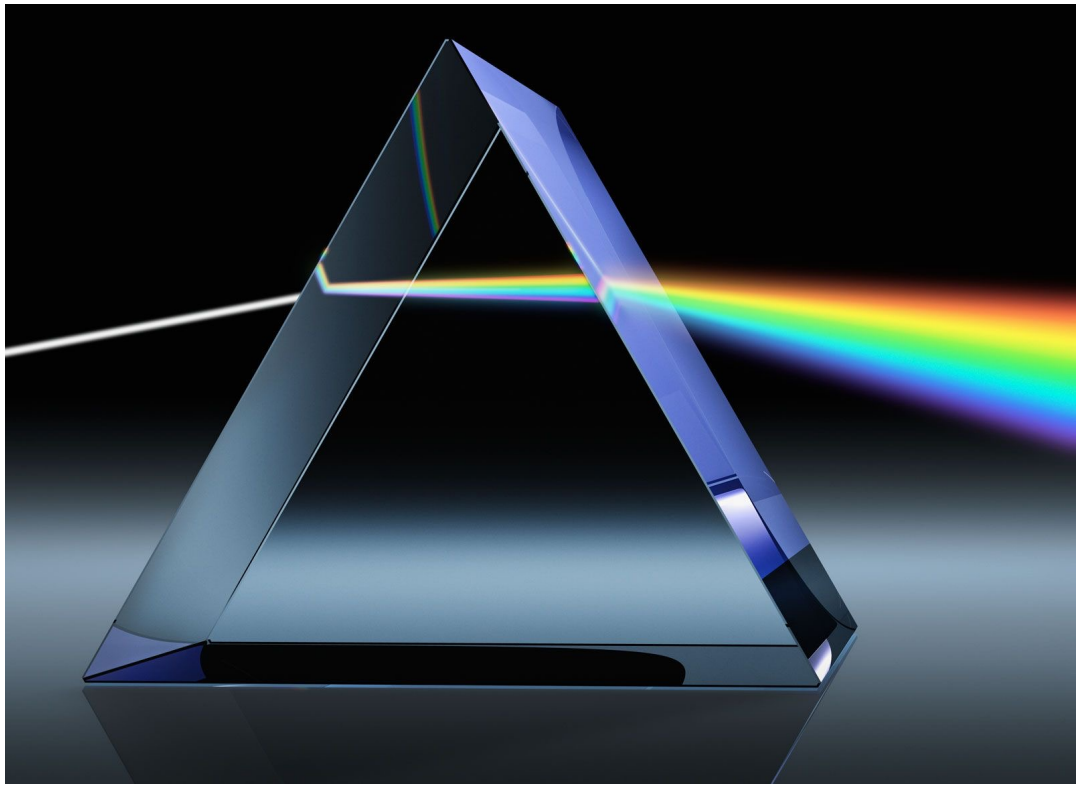
More measurement and theory connections

- By reversing his telescope Galileo produced one of the earliest microscopes
- Polymath Robert Hooke advanced on the design in 1665, birthing microbiology
- Hooke was the first to document tiny insects
- In his book Micrographia Hooke first saw and named cells



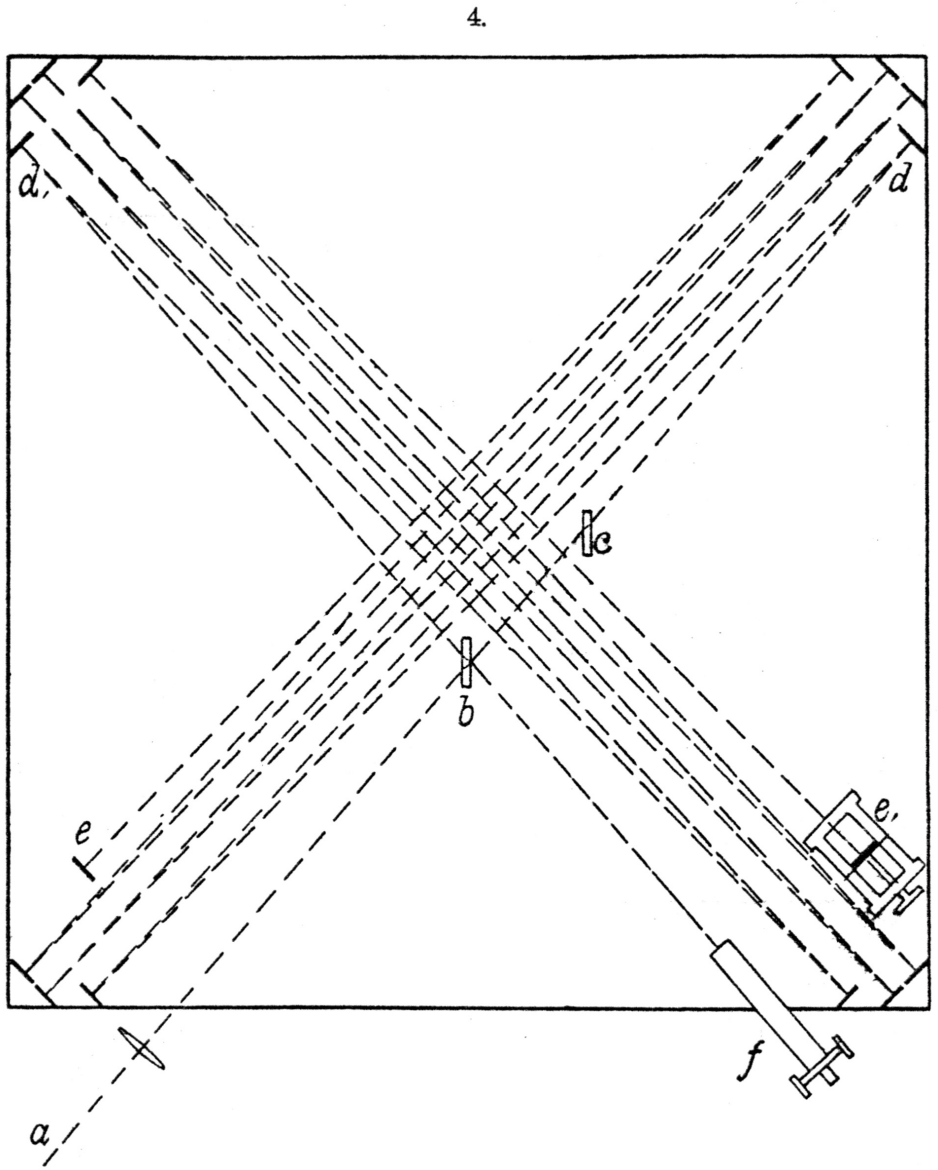
Theoretical progress via measurement advancement

- Last example!
- Wave theories of light were long believed based on behaviors resembling physical waves
- All other waves travel in physical media; physicists hypothesized *aether* as light's medium
- An attempt to *confirm* aether actually **disproved** it! Michelson-Morley experiment



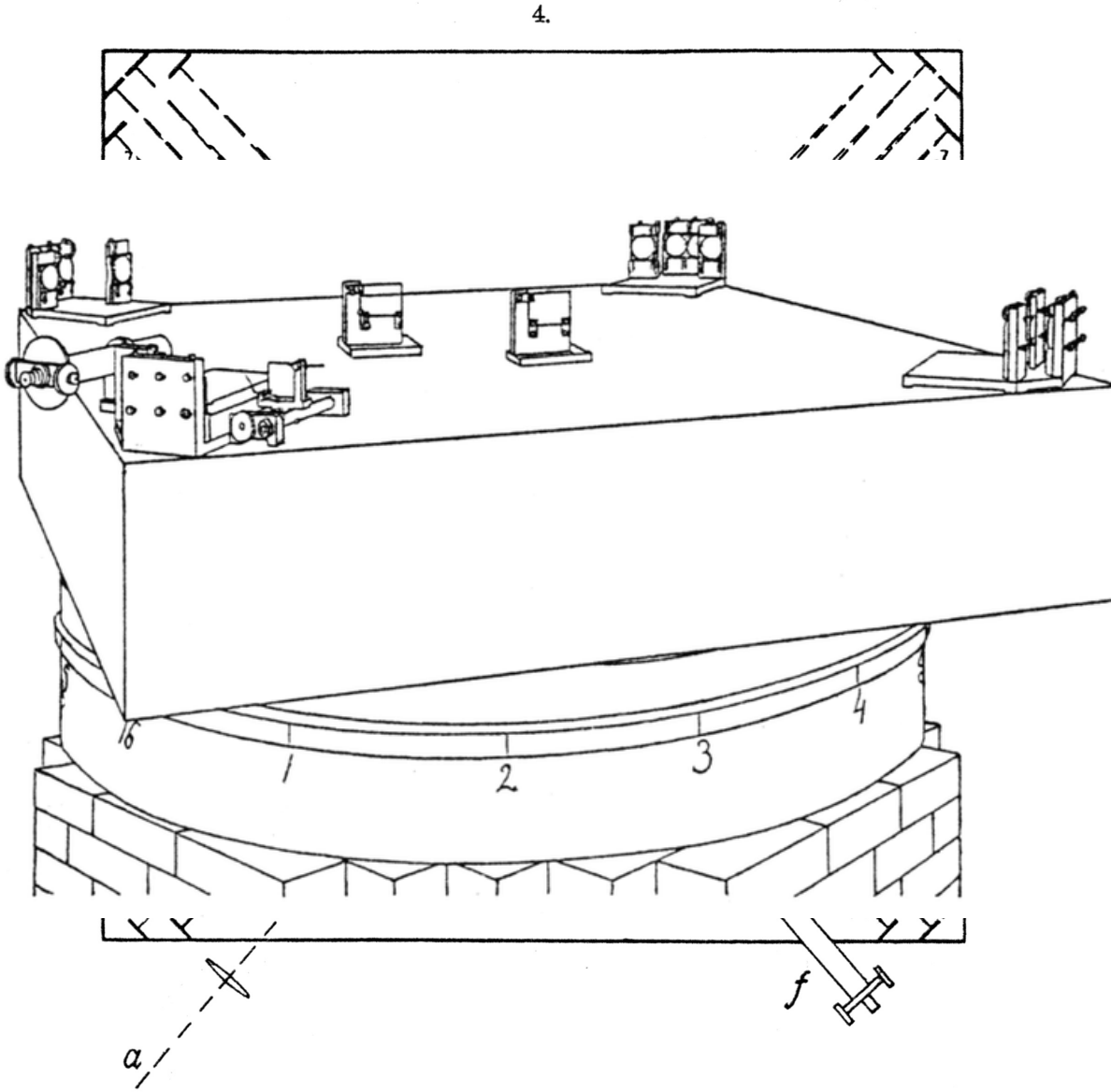
Advances in measurement: Michelson-Morely

- Attempt to measure the speed of light through in diff directions
- "Aether wind" would require diff speeds in diff directions
- Testing this required a **precise measurement apparatus**



Advances in measurement: Michelson-Morely

- Attempt to measure the speed of light through in diff directions
- "Aether wind" would require diff speeds in diff directions
- Testing this required a **precise measurement apparatus**
- Precision optics only possible in 1887, prior setups inadequate
- Found same speed in all dirs!
- Surprising find drove theory forward; relativity reconciled it



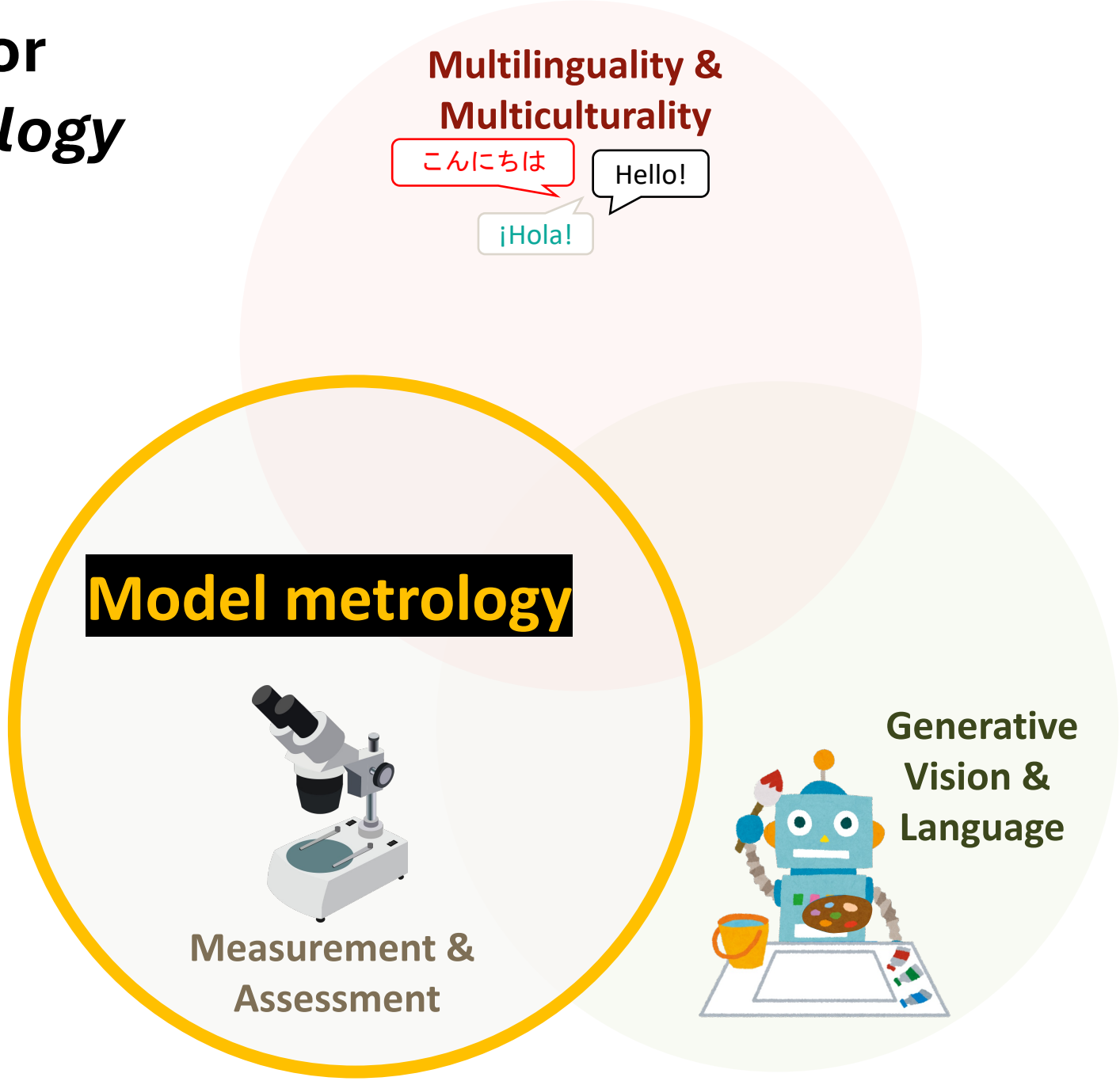
The point?

- Galileo, Hooke gentleman polymaths who built their own tools and made their own discoveries
- Tool sophistication needs required specialization, Michelson and Morely didn't make their own mirrors
- Nowadays, telescopes and microscopes are built by dedicated engineering teams

The point?

- Mass-market commodity tools now exist for
- Measurement in manufacturing
- Measurement in biology
- Specialist engineers continue to push frontiers
- The field of research in measurement tools is called **metrology**

Advocating for *Model metrology*



Rigorous measurement for AI systems? (opinion)

- **Measurement side** of AI science needs focus!
- We are now at **maybe** the Galileo-telescope-stage
- Many issues are present in AI that aren't in science
- Disparate observations aren't clearly connected
- Clear **constraints** in our measurement tools
  - Generalized capability assessment is futile, contested!
  - Constrained capabilities more measurable?
- **Calibration** of measures is needed!
- **Construct validity** is needed!

What would measurement tools for AI look like?

- Not sure!
- Benchmarks are promising, but have issues:
- When they saturate, they just become useless
  - We don't learn much about system capabilities **outside the benchmark's world**
  - Models get perfect scores on GSM8K; still can't add
- Ideal benchmarks should be:
  - **Dynamic**ally generated (to avoid memorized saturation)
  - **Calibrated** so that real saturation tells us something useful about

Ideal benchmarking for metrology

- To get **construct validity**:
- Go **constrained** and/or **work with domain experts**
- Generalized reasoning is hard to measure, will it generalize?
- Maybe reasoning wrt customer service policies is more measurable!
- Could be enabled via adversarial agents, etc

Air Canada chatbot promised a discount. Now the airline has to pay it.

Air Canada argued the chatbot was a separate legal entity 'responsible for its own actions,' a Canadian tribunal said

By Kyle Melnick

February 18, 2024 at 8:35 p.m. EST

Getting there as a community

- Encourage model vendors to use **many constrained metrics** instead of flawed "generalized" measures
- Building business-relevant metrics might incentivize this
- Get domain-experts and users to bring their constraints
- Focus research on how to reliably build dynamic **benchmark-generating processes** that constraints can be plugged in to
- Build a **community of research and practice** focused specifically on developing measurement tools to share best practices, insights, and findings

Putting my money where my mouth is

Attempting to build such a benchmark

Text-to-image (T2I) models have grown increasingly popular...

... but they have 3 big issues!

Issue 1: We don't know how they are so "creative"

- In these example images, DALL-E 2 (OpenAI) is able to creatively combine some set of *known concepts* into novel images
  - Astronaut, horse, soup, planet, basketball, cat, bear… + *styles*
  - All prompted from purely plain text!
- We don't **understand how they do this!**
- **What mechanism enables creative combination of concepts?**

Issue 2: We don't know what knowledge they even have

- It seems believable that in order to make, "***a dog made of fire standing on the moon***" the model must "know" what "a dog," "the moon," "fire," etc… is!

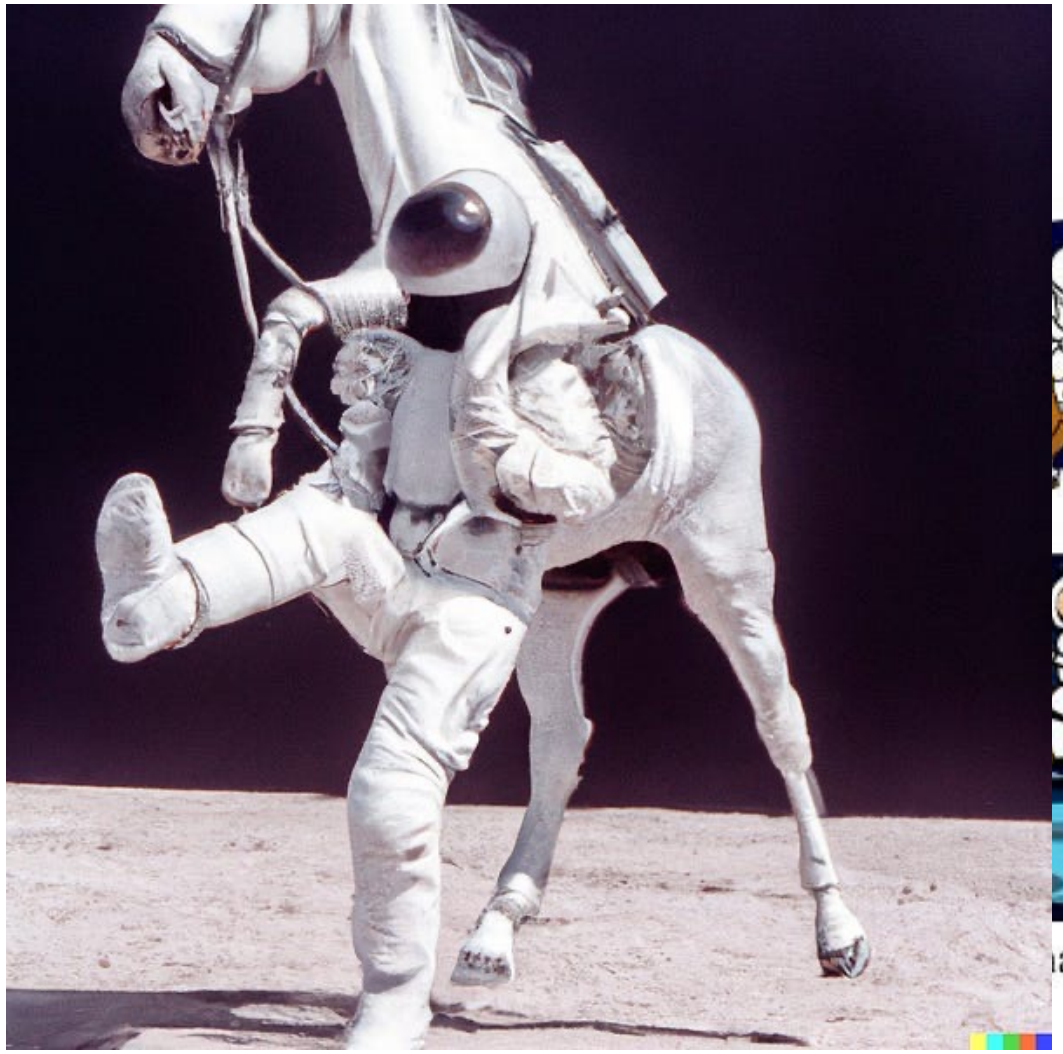
- **But what is the nature of this requisite conceptual knowledge? How can we test for it, and how can we add it if it's missing?**

Horse riding astronaut?

- Commonly discussed "hard prompt" for T2I
- Gary Marcus in 2022: "evidence of no reasoning"

Horse riding astronaut?

- Commonly discussed "hard prompt" for T2I
- Gary Marcus in 2022: "evidence of no reasoning"
- Joscha Bach in 2022: "I told it to think harder"
- What knowledge was gained in this exchange?



Issue 3: We can’t predict the ways they will fail in deployment

A real-world example from the multilingual setting….

For example: *what is a dog to DALL-E Mini?*

🇺🇸

For example: *what is a dog to DALL-E Mini?*

“big dog”

DALL-E Mini

🇺🇸

For example: *what is a dog to DALL-E Mini?*

“perro grande”

For example: *what is a dog to DALL-E Mini?*

“perro grande”

DALL-E Mini

For example: *what is a dog to DALL-E Mini?*

“anjing besar”

🇮🇩

For example: *what is a dog to DALL-E Mini?*

“anjing besar”

DALL-E Mini

🇮🇩

For example: *what is a dog to DALL-E Mini?*

“大きな犬”

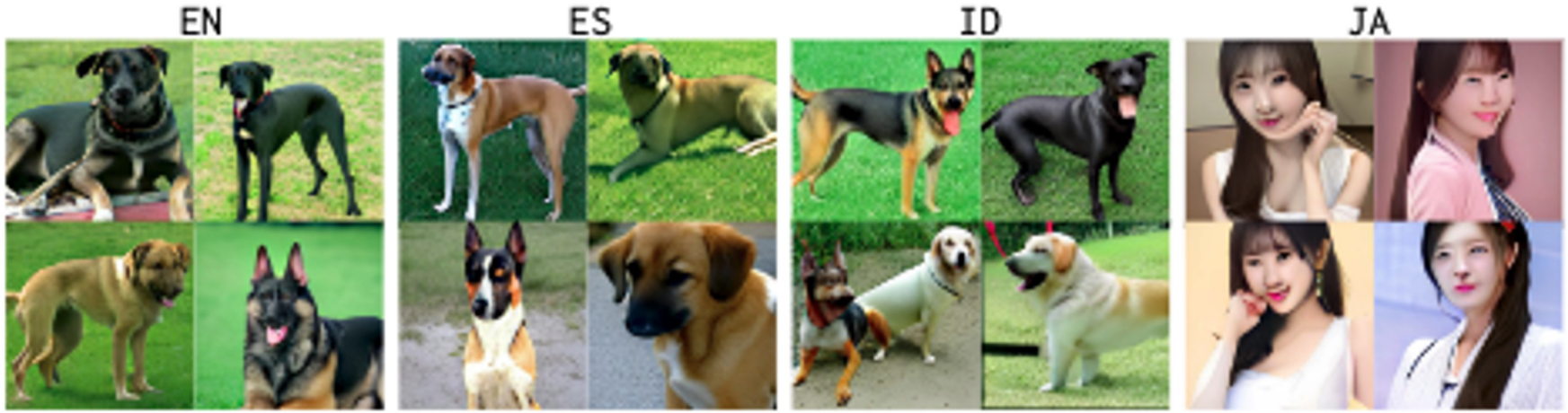
For example: *what is a dog to DALL-E Mini?*

“大きな犬”

DALL-E Mini

🇯🇵

What the heck!?

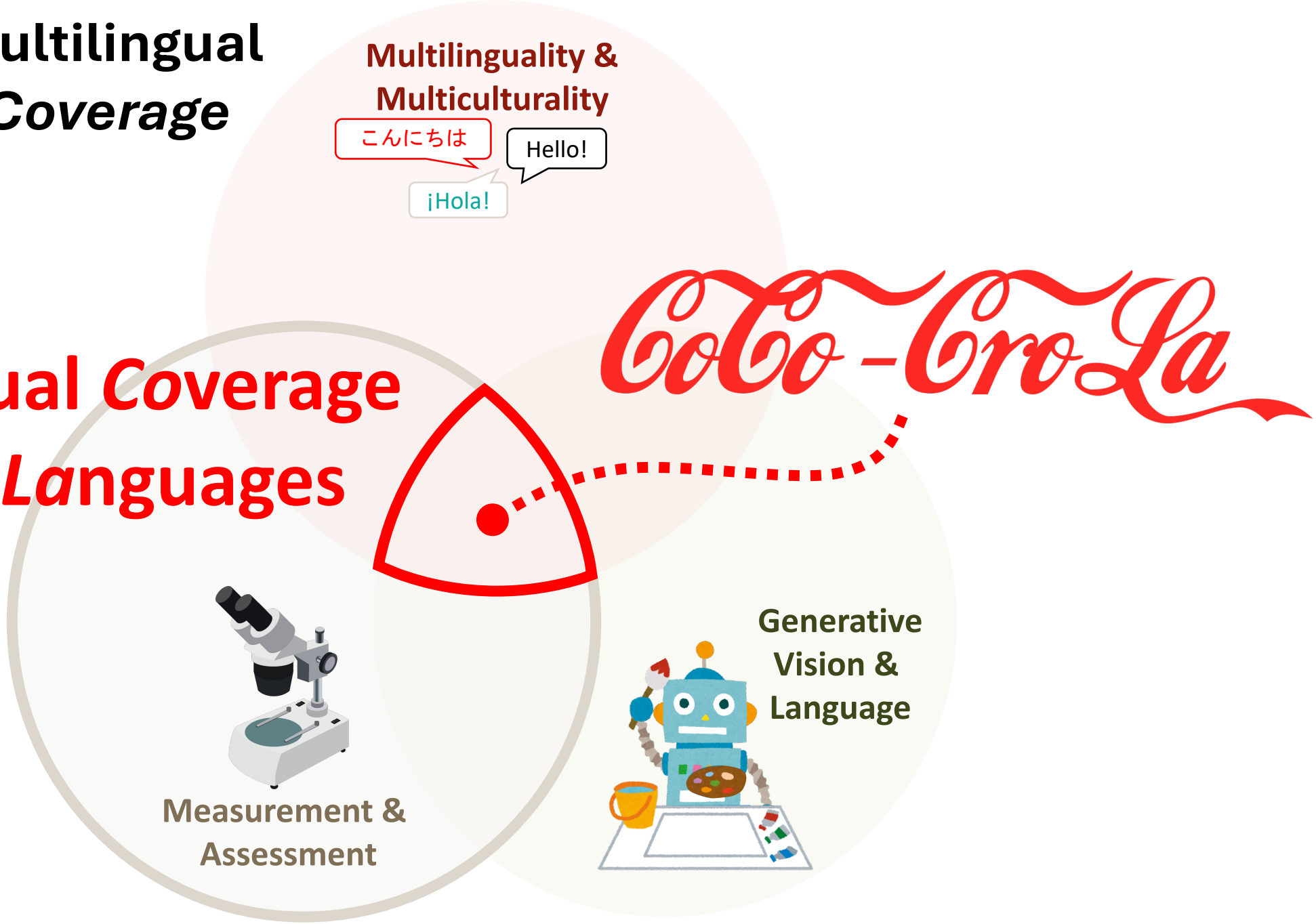


- “Dog”->”woman” collision in JA is **unpredictable** and **offensive**.
- How can we find more of these? How can we mitigate them?

We progress toward solving all 3 issues by studying T2I models through the lens of:

Multilingual Conceptual Coverage

Inspecting Multilingual *Conceptual Coverage*



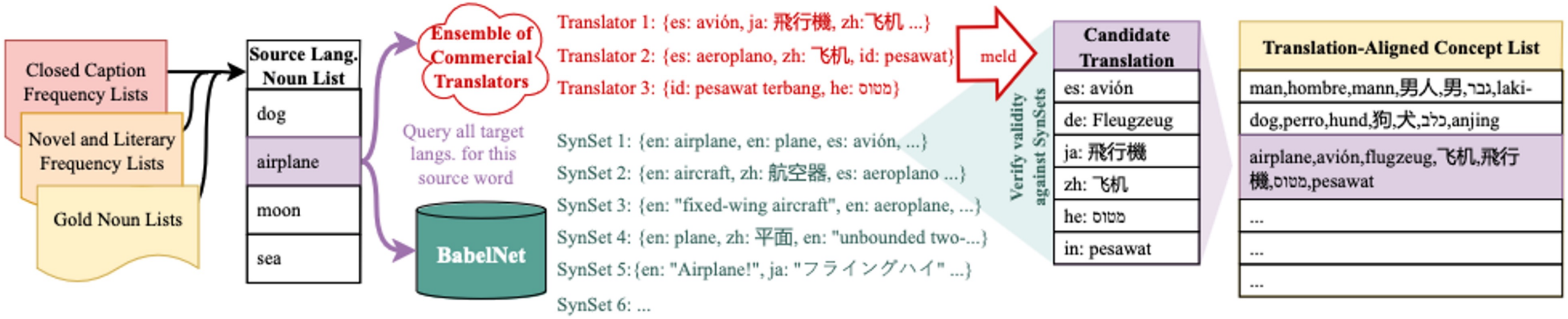
This benchmarking domain checks many boxes

- Despite being intended for English-only use, **most models trained on web-scale T2I datasets have multilingual** capabilities anyway
- We should **ensure these models actually work** well in all languages (construct validity)
- This is a **constrained** setting too (knowledge of tangible concepts)
- We introduce a **benchmark generating process** for assessing multilingual T2I knowledge

Technique

Collecting the concept list, metrics to analyze coverage, models tested

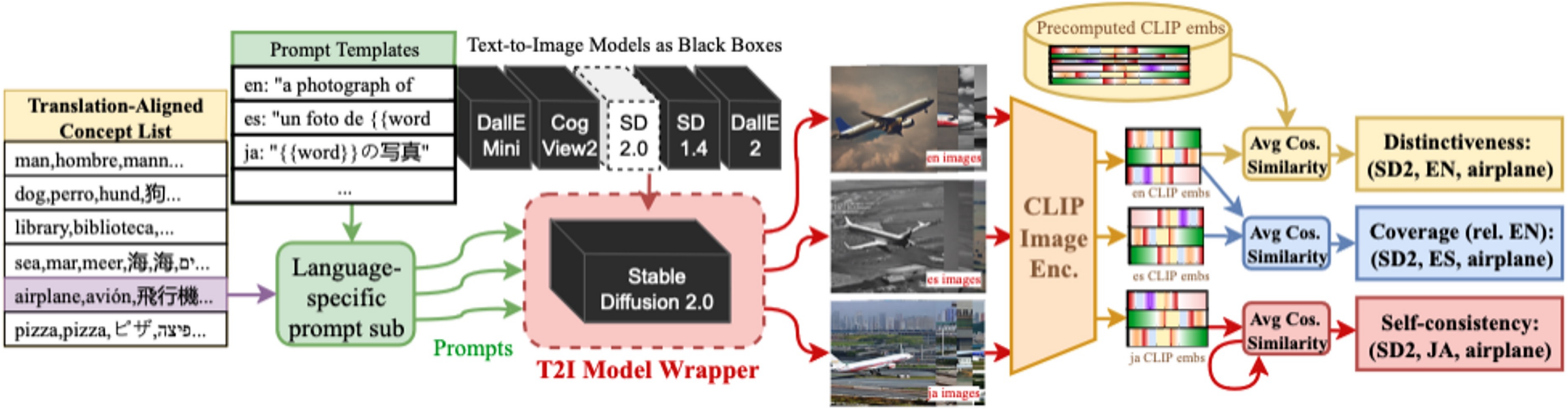
Multilingual Concept List



- We collect 1,000 common English nouns
- We translate each into Spanish, German, Chinese, Japanese, Hebrew, and Indonesian with MT
- To ensure* quality translations we use 4 commercial translators in ensemble + BabelNet

Metrics: How do we know a model "knows" a concept?

- We consider a model to "possess" a concept if it is able to **distinctively**, **correctly**, and **consistently** generate images of a tangible concept (e.g., dog, airplane, man, computer, fire) when prompted.
- We **treat each model as a black box** and assess conceptual possession by **analyzing the population of images it generates**

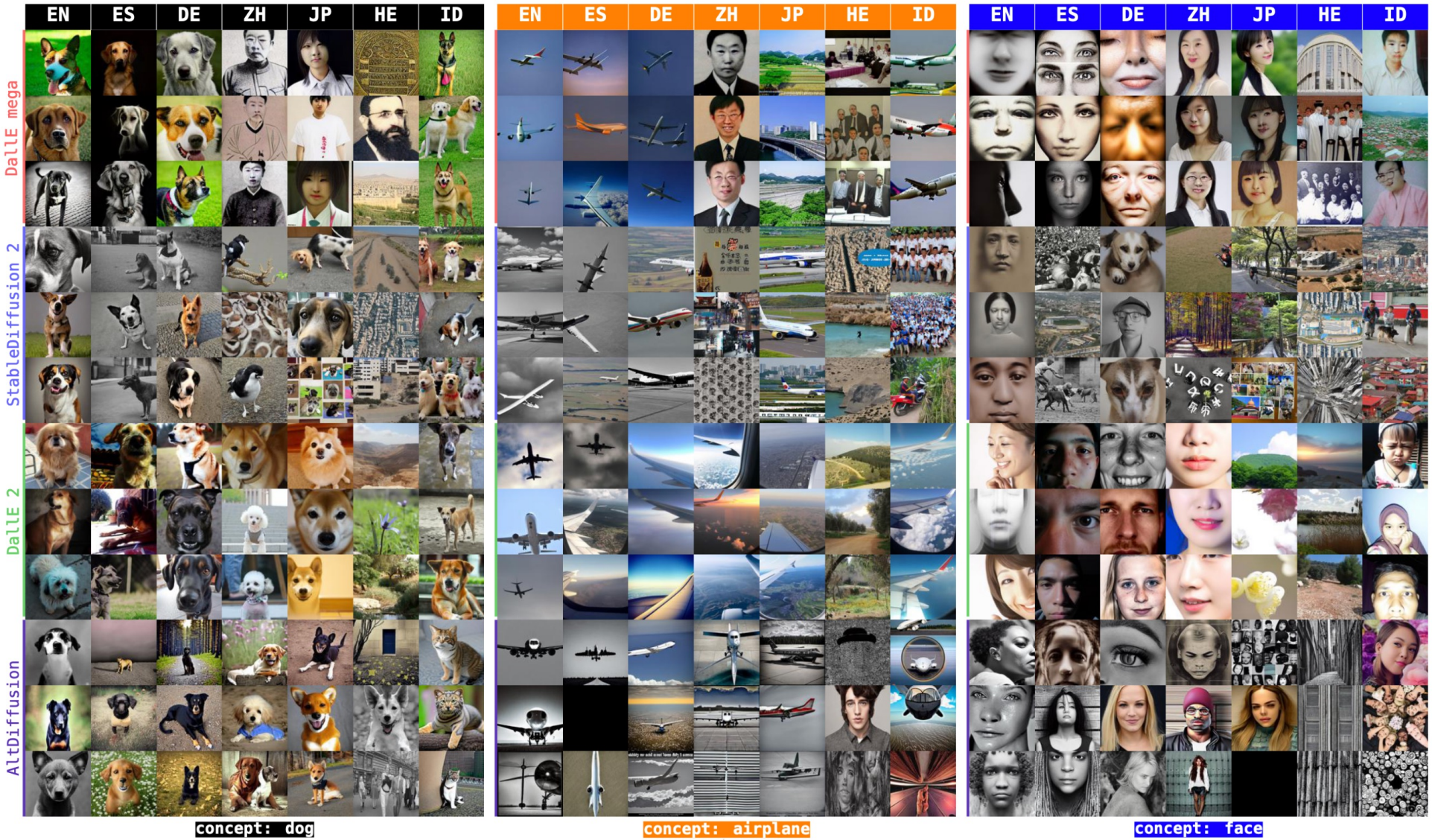


Models Tested

- We assess the multilingual performance of a variety of commercial and open-source T2I models
- **DALL-E mini/mega (Craiyon.com)**
- **StableDiffusion 1.1, 1.2, 1.4, 2.0 (Stability.ai)**
- **CogView 2 (academic)**
- **DALL-E 2 (OpenAI)**
- **AltDiffusion m9 (academic)**

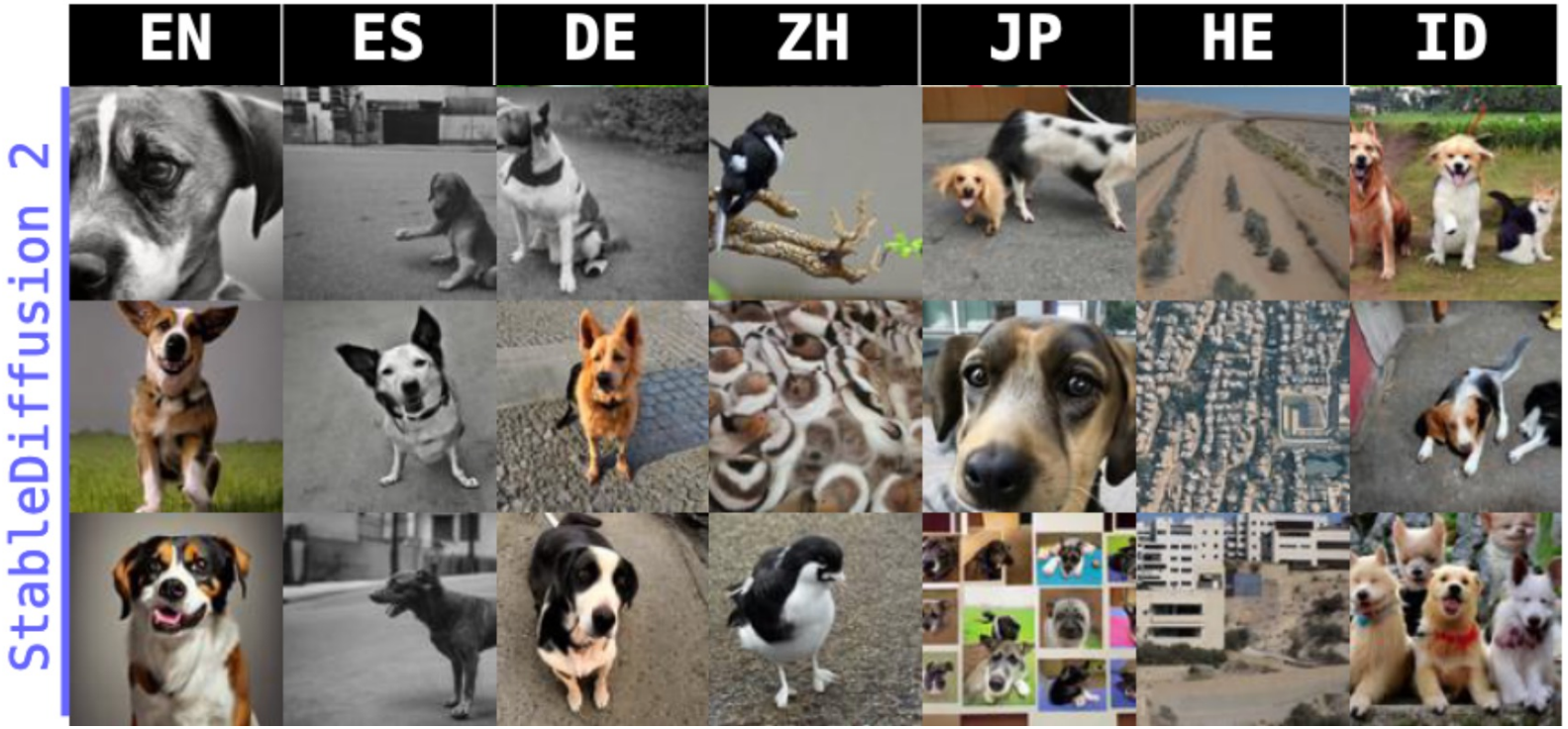
Output Examples

Example: Dog, Airplane, Face

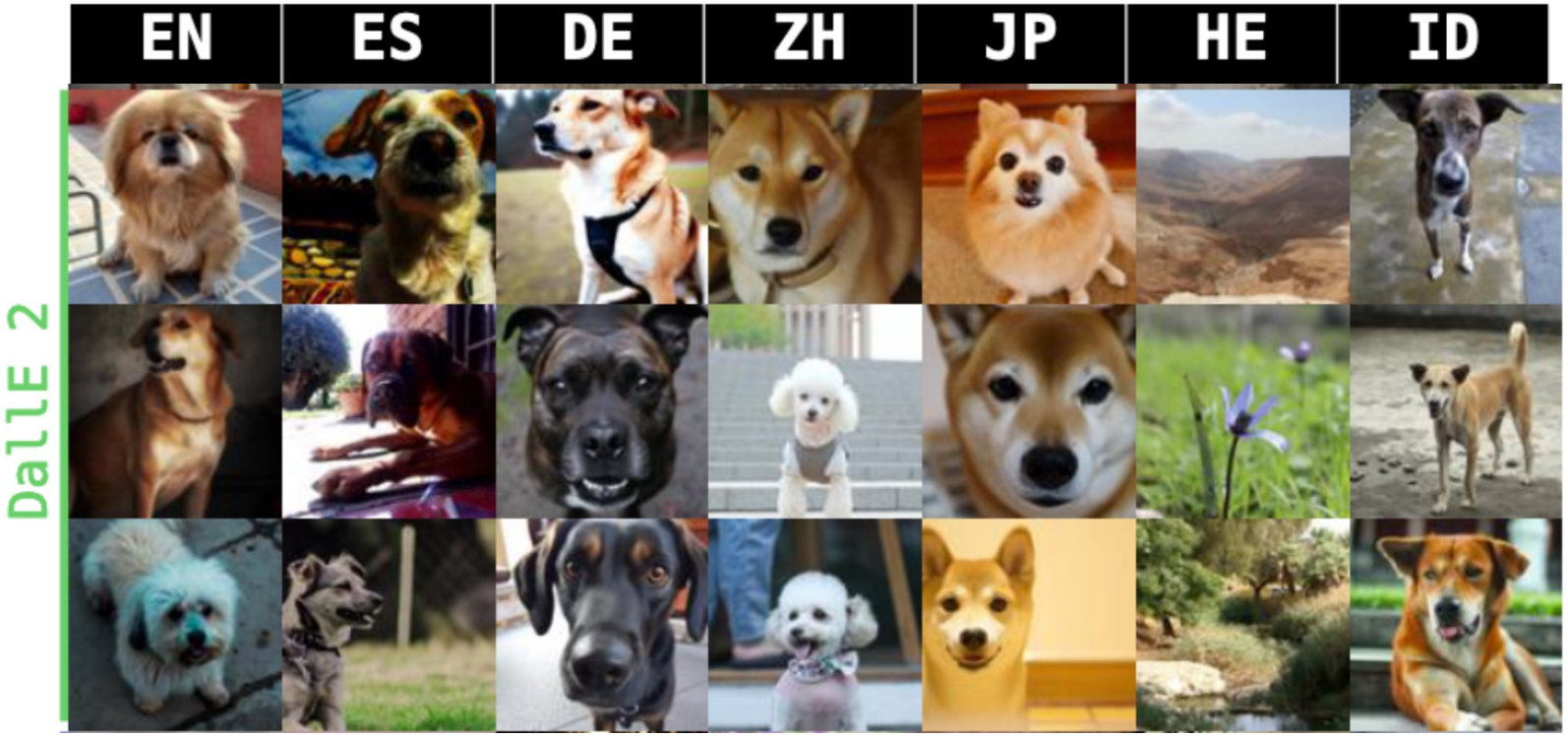


Example: Dog, DALL-E Mega

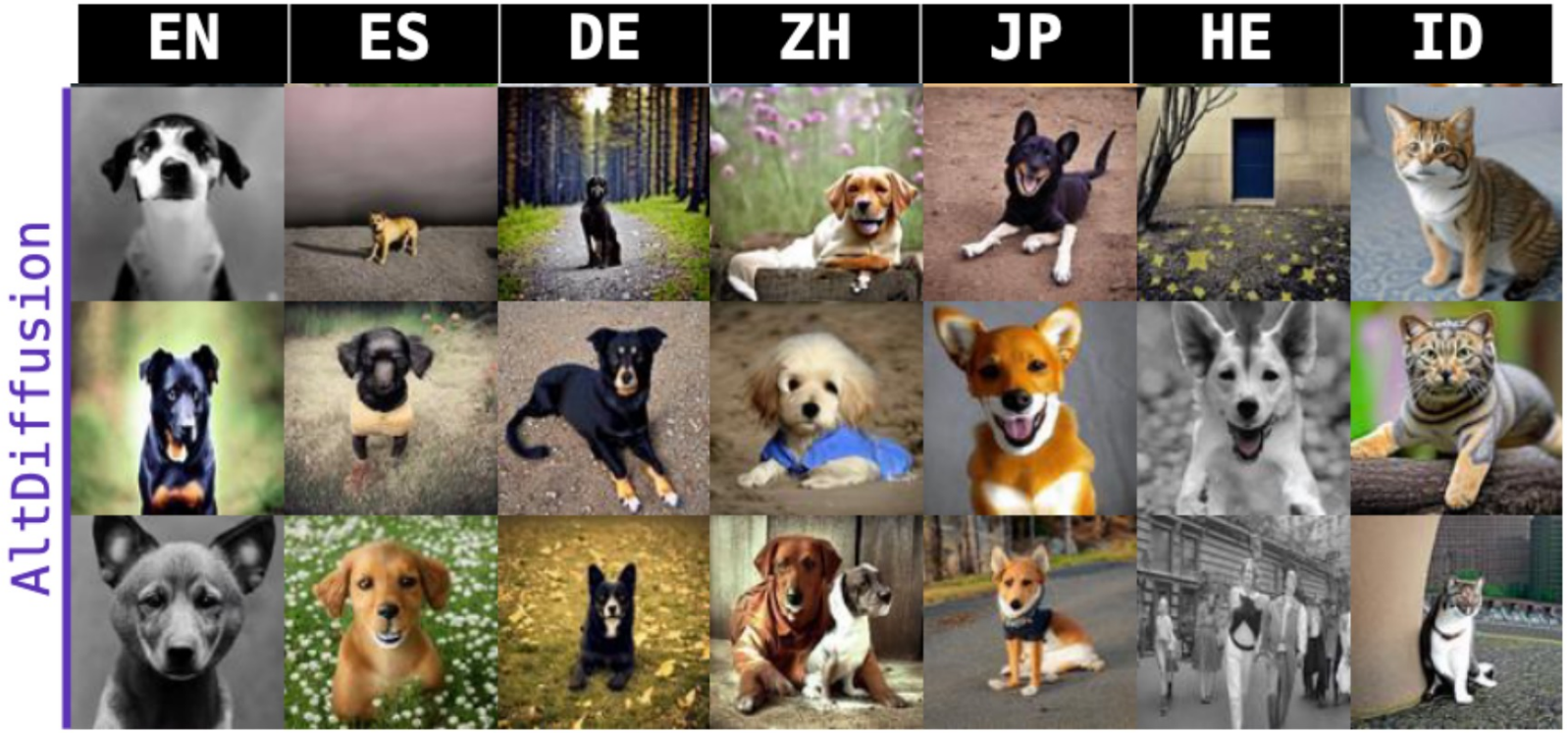
Example: Dog, Stable Diffusion 2



Example: Dog, DALL-E 2



Example: Dog, AltDiffusion



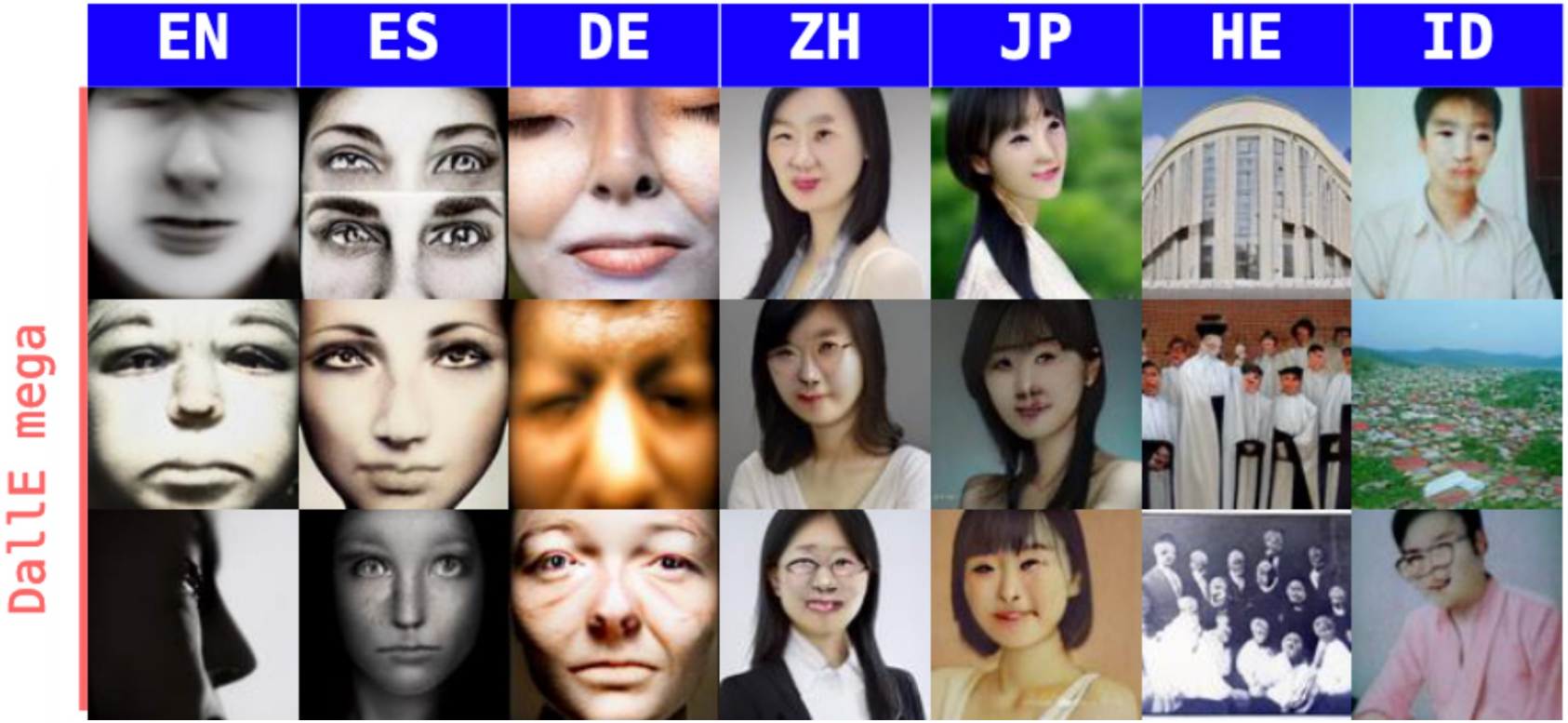
Example: Airplane, DALL-E Mega

Example: Airplane, Stable Diffusion 2

Example: Airplane, DALL-E 2

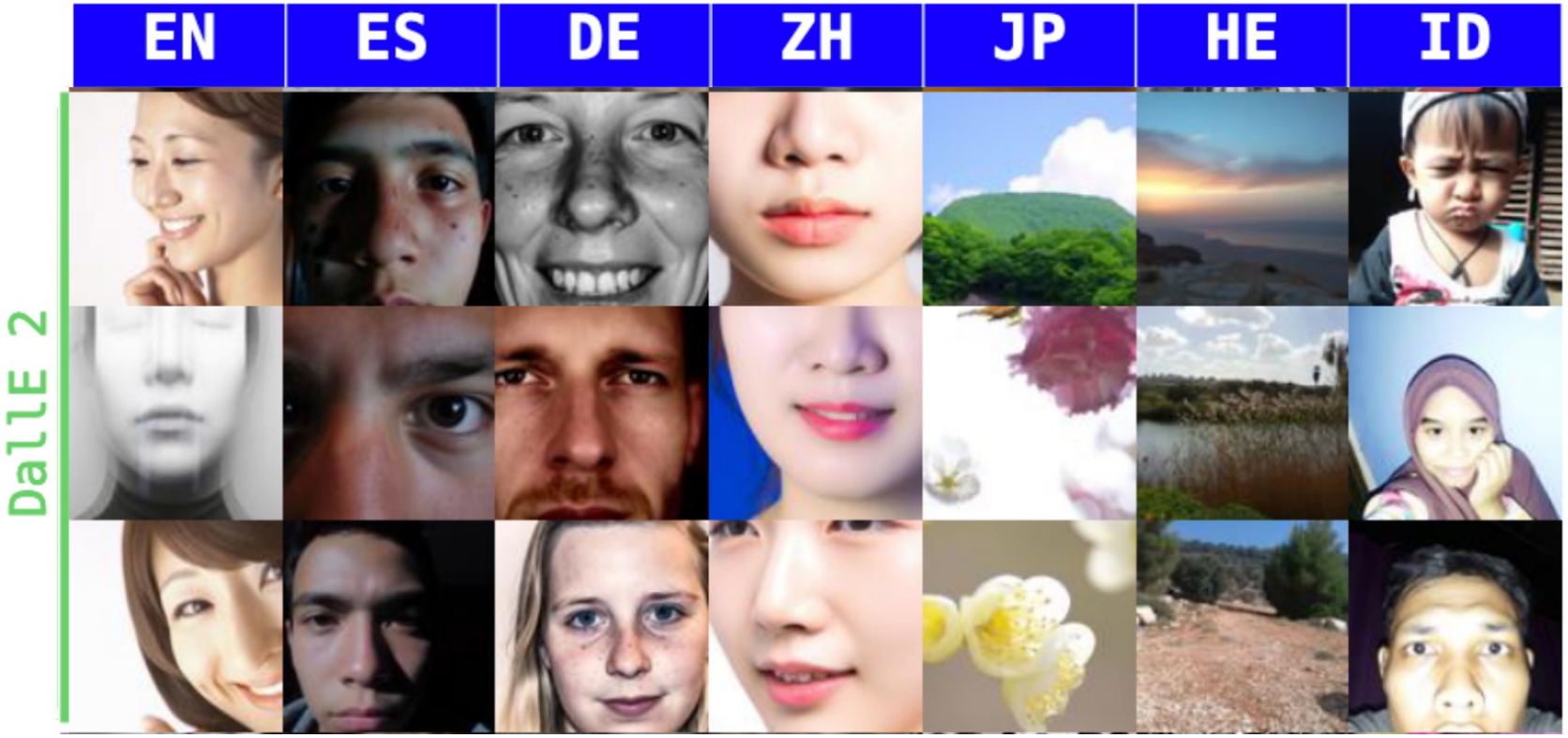
Example: Airplane, AltDiffusion

Example: Face, DALL-E Mega



Example: Face, Stable Diffusion 2

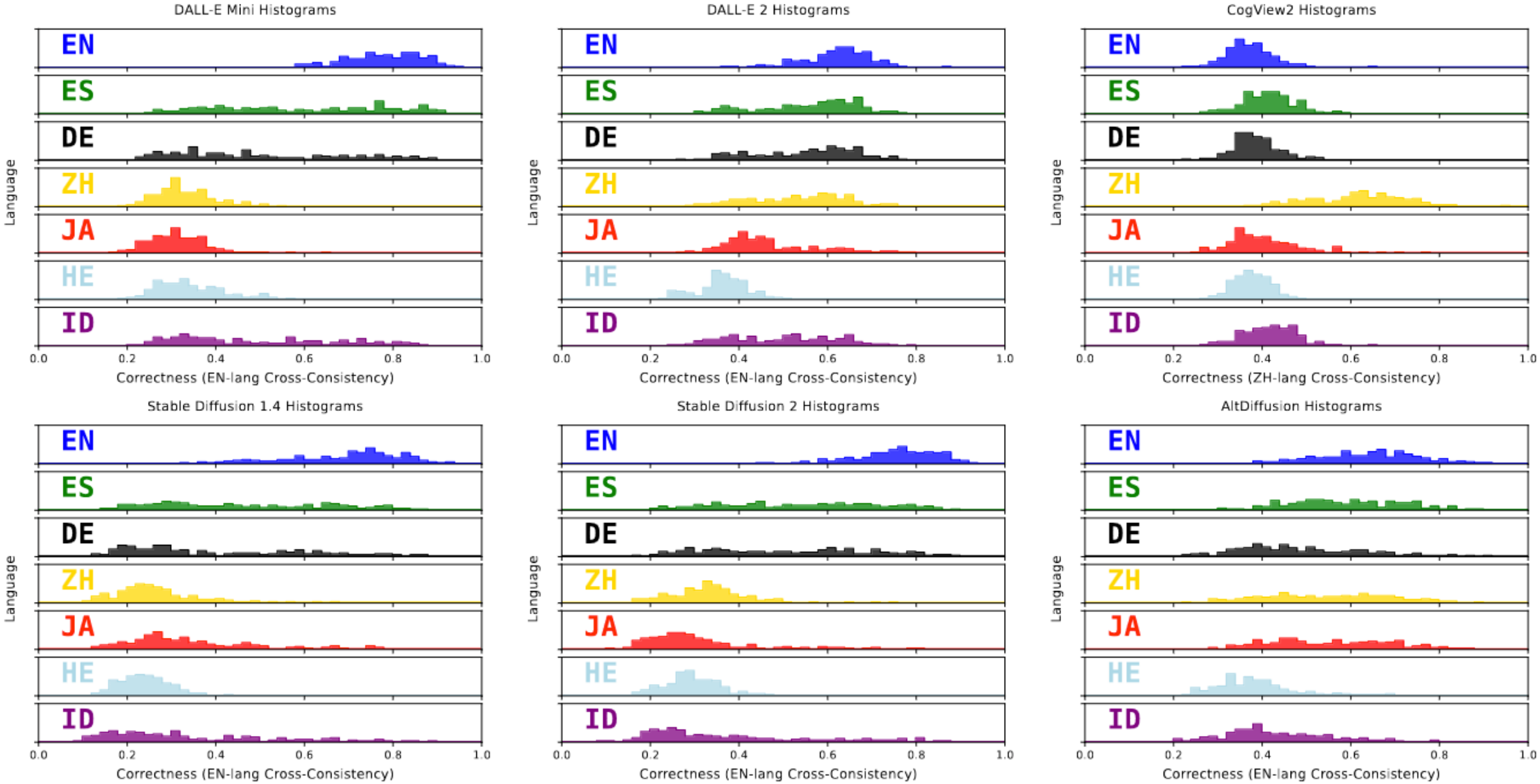
Example: Face, DALL-E 2



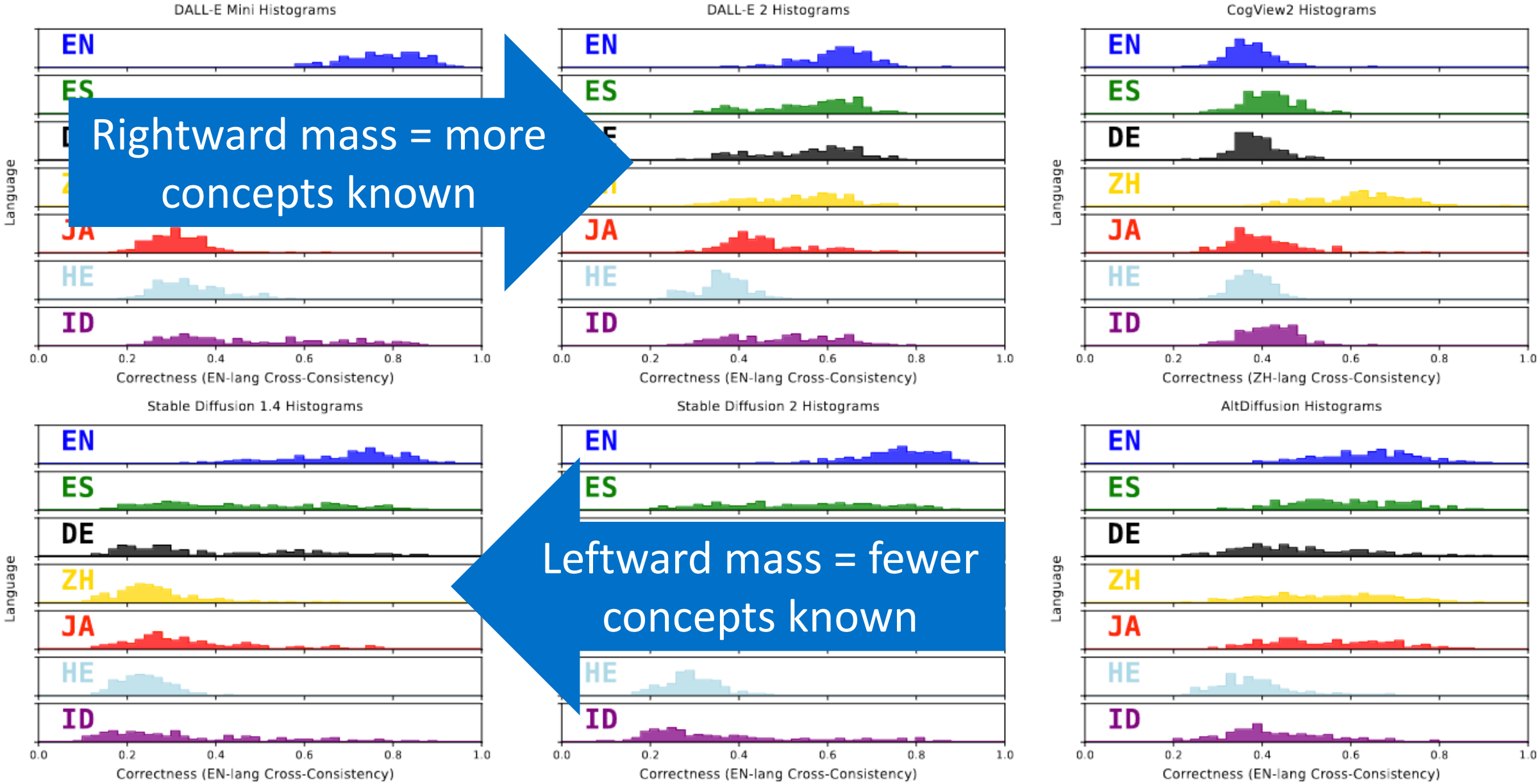
Example: Face, AltDiffusion

Quantitative Results

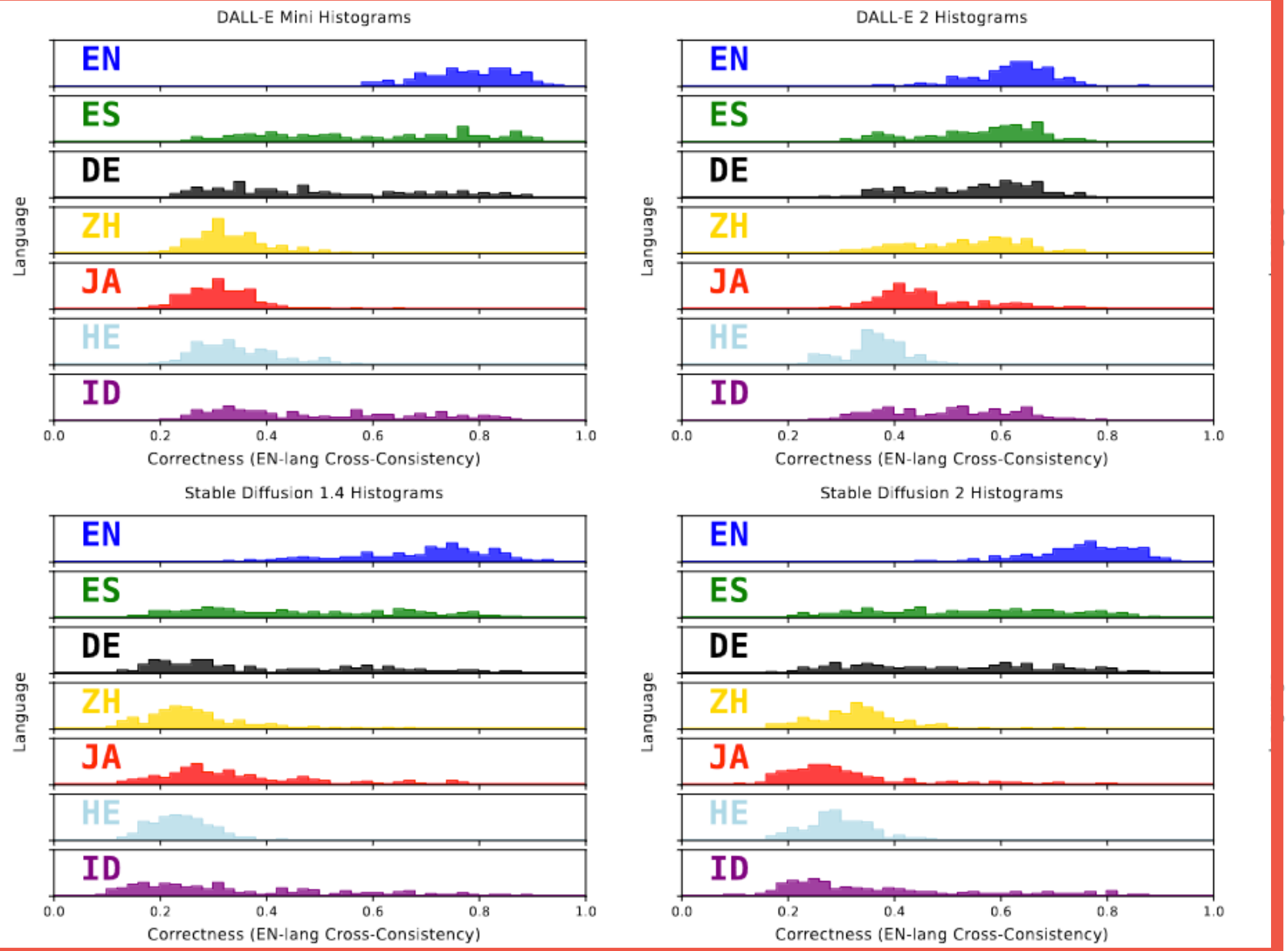
Comparing Conceptual Possession



Comparing Conceptual Possession

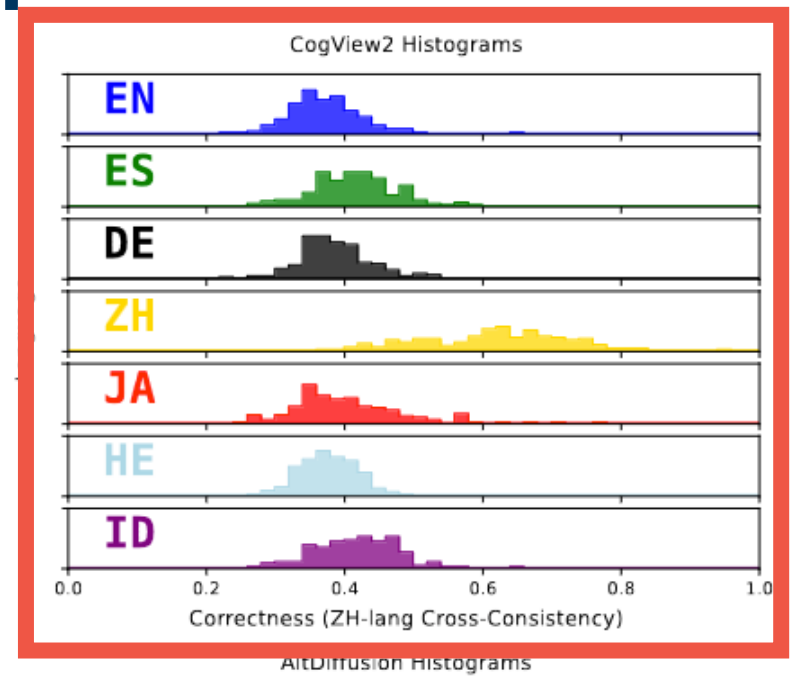


Comparing Conceptual Possession



Models trained primarily on English tend to favor languages that use Latin alphabet (ES, DE, ID)

Comparing Conceptual Possession

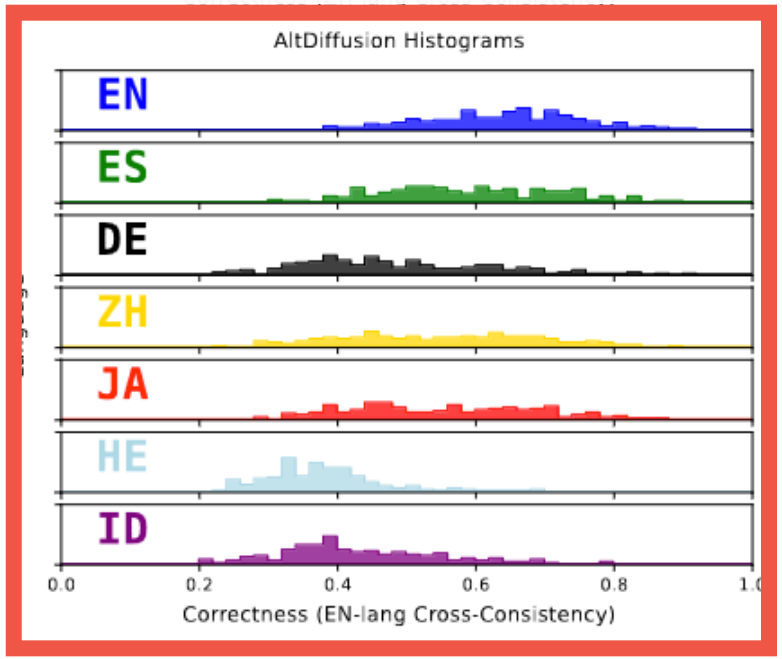


CogView2 (trained on Chinese only) does bad on all other languages except slightly better on JA (some exact-character-match words)

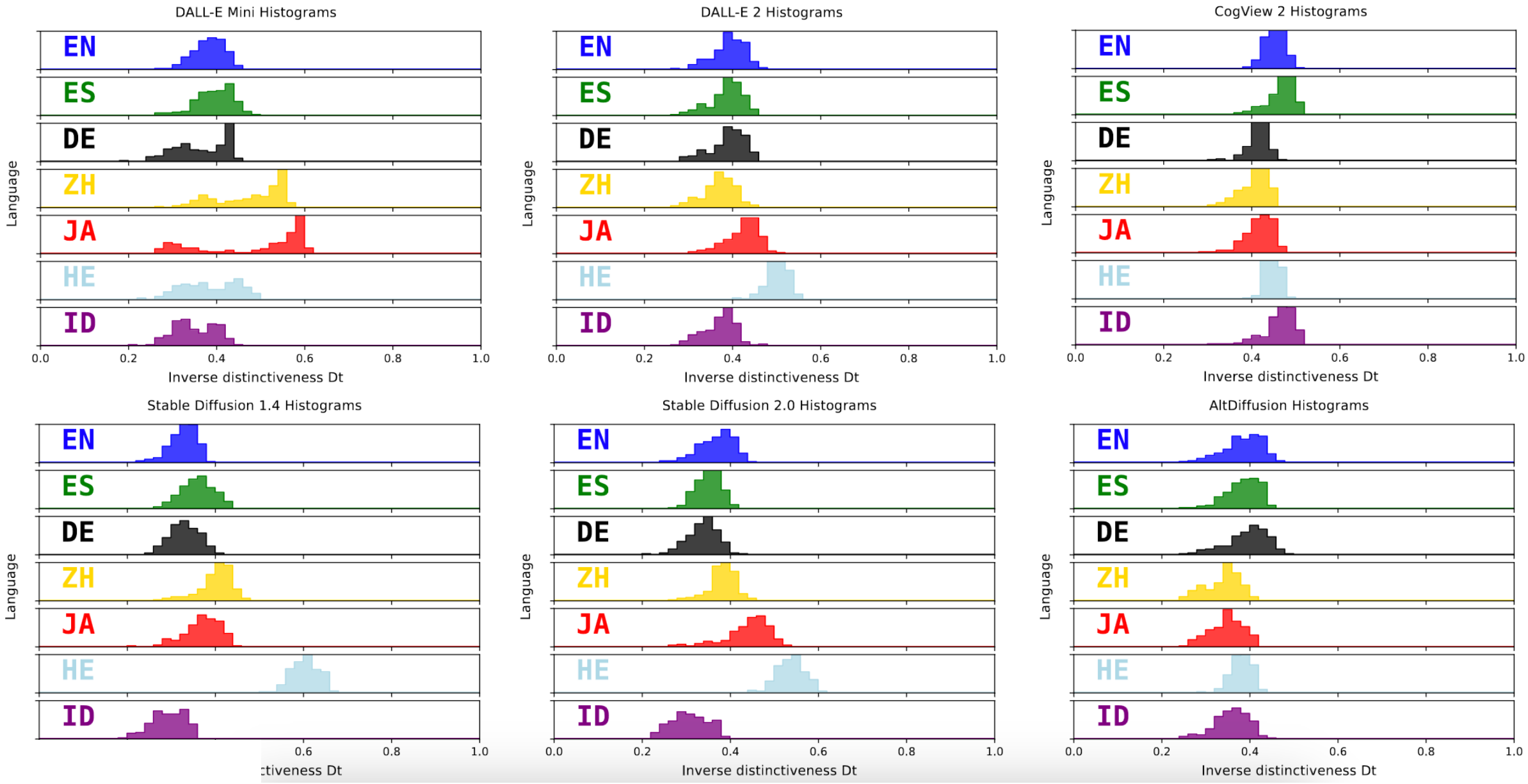
Comparing Conceptual Possession

AltDiffusion is the only model that received multilingual training; it performs the most consistently across all languages

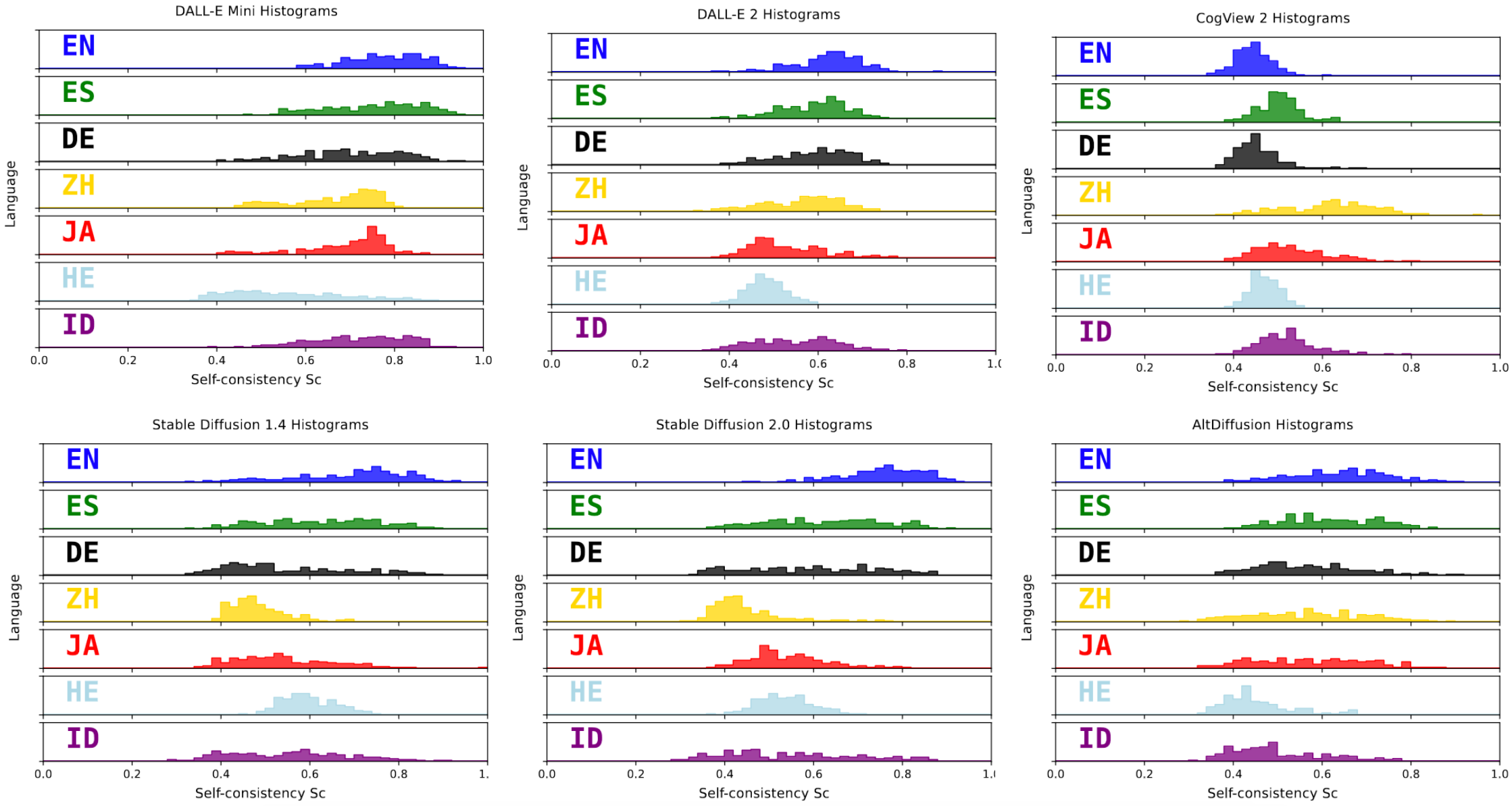
But multilingual performance comes at a cost; it's worse in English than the others (more leftward mass for EN histogram)



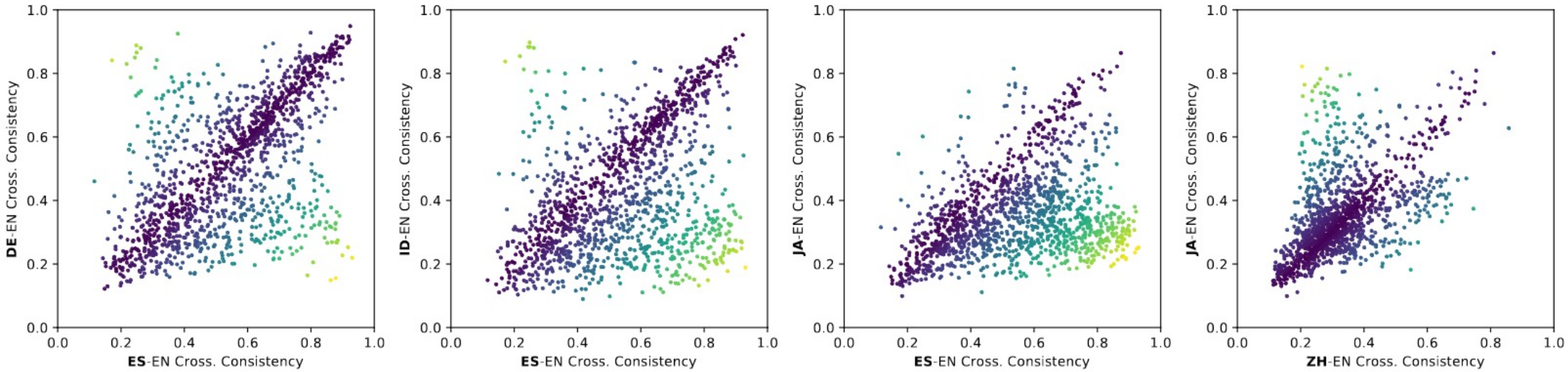
Comparing *Inverse* Distinctiveness



Comparing Concept Self-Consistency



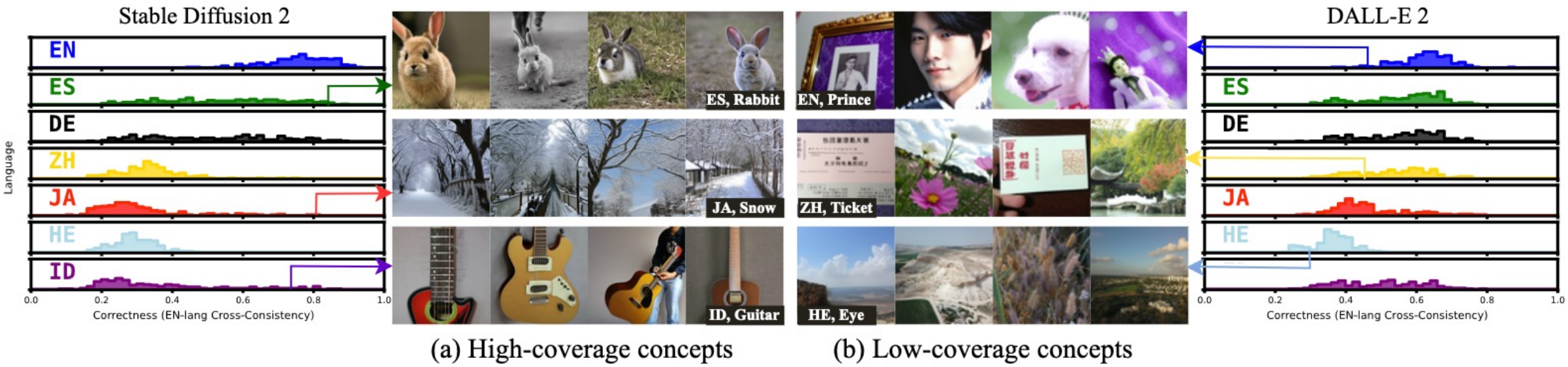
Language cross-correlations



- In general we find **some correlation** between script-sharing languages on (concept, model) level
- Eg *If a given model possesses a given concept in ES, it is likely to have it in DE and ID*
- ***Still there are plenty of exceptions***

Qualitative results

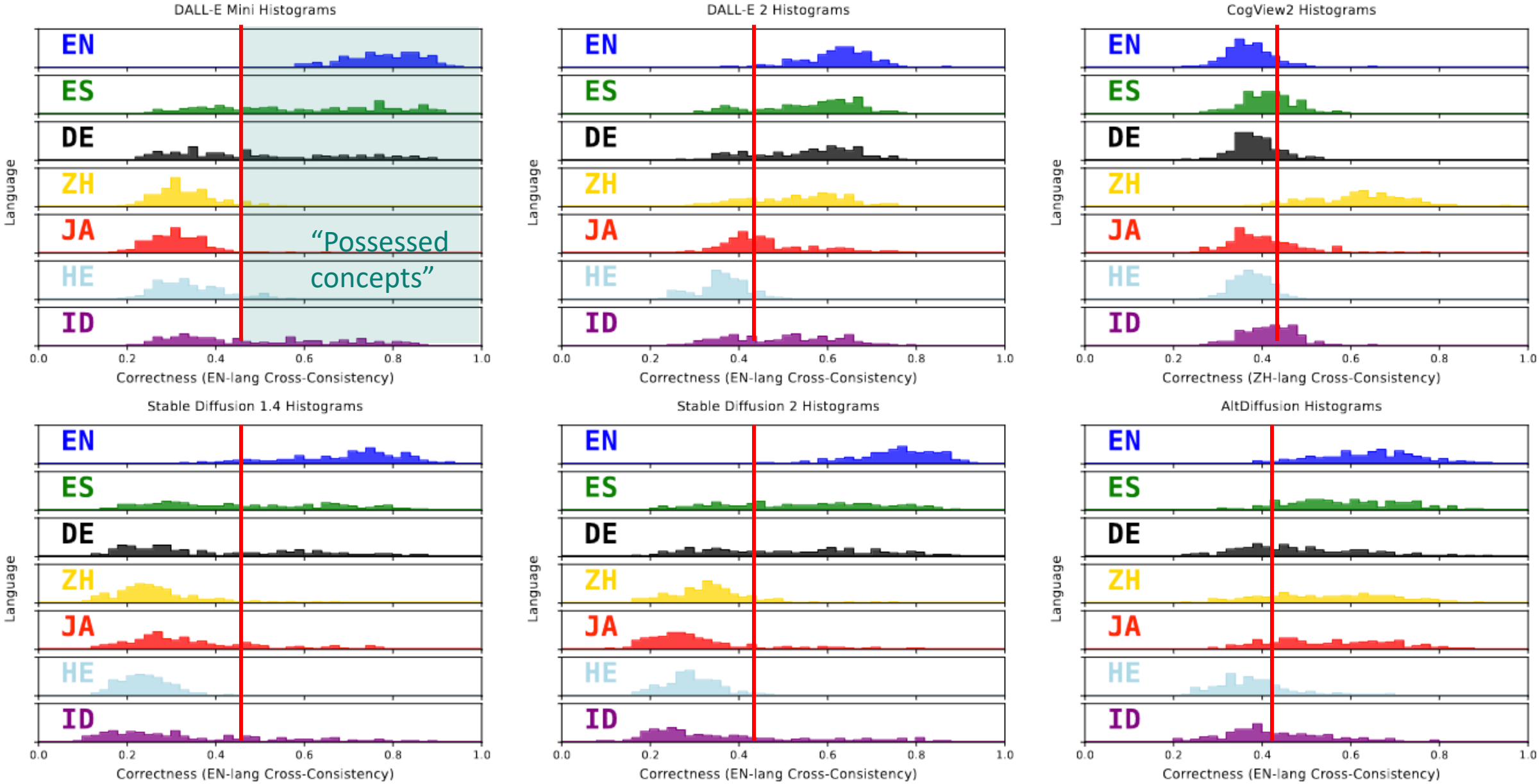
Can we use this to automatically find failure cases or verify conceptual possession?



(a) High-coverage concepts (b) Low-coverage concepts

- Easy! Sample from the correctness histograms!
- Fig above shows SD2 has (ES, Rabbit), (JA, snow), (ID, guitar); DALL-E 2 does NOT have (EN, Prince), (ZH, Ticket), (HE, eye)

“Metricate” the score with thresholds



Checking complex creativity with simple concepts?

- Now that we have a test of if a model "knows a concept" in some language
- Natural **within-model A/B tests between languages** where concept is/isn't known
- **We can test if our simple prompt-based knowledge test generalizes to complex settings**
- Given "a bird using a keyboard in the snow" in EN, ES, JA:
- **If a model does NOT possess a concept it will NOT be able to use it to make a creative prompt**
- Knowledge by our test is **necessary** but not **sufficient**

AltDiffusion

Concept	Xc	Wc	
bird	0.655	26	✔
keyboard	0.491	27	✔
snow	0.759	26	✔

Concept	Xc	Wc	
bird	0.646	26	✔
keyboard	0.489	26	✔
snow	0.704	25	✔

Concept	Xc	Wc	
bird	0.655	26	✔
keyboard	0.462	24	✗
snow	0.671	25	✔

Stable Diffusion 2

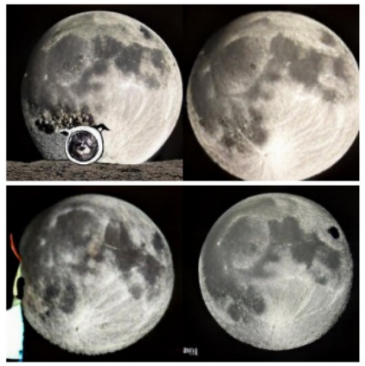
Concept	Xc	Wc	
bird	0.726	27	✔
keyboard	0.837	29	✔
snow	0.846	26	✔

Concept	Xc	Wc	
bird	0.697	26	✔
keyboard	0.789	29	✔
snow	0.818	26	✔

Concept	Xc	Wc	
bird	0.655	26	✔
keyboard	0.797	29	✔
snow	0.808	26	✔

Checking complex creativity with simple concepts?

- We find similar results for "a dog made of fire standing on the moon"
- This is the most exciting direction for future work
- The hope is that by **isolating model knowledge** as a factor we can test generalization on its own with these A/B conditions

	EN	ES	JA
DALL-E Mega			
AltDiffusion			

DALL-E Mega

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	✗

AltDiffusion

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	✔

Differential biases between languages

In increasingly sinister order…

Differential behavior: guess the language for *dog*

Different “typical” *dog* breeds: east vs west

EN

ES

JA

Differences for *hair* grows more stereotypical…

DE

ID

ES

Differences for *doctor* embed racial, gender biaseses*

EN

ZH

ID

*Grammatically gendered languages complicate this straightforward analysis of gender bias…

Racial bias and differential degrees of sexualization in images of women

DE

ES

JA

Additionally, orthogonal degrees of variation like pose, color vs black/white

AltDiffusion (the one multilingual model) has more uniform cross-lingual notions but less diversity

EN

ES

JA

“dog”

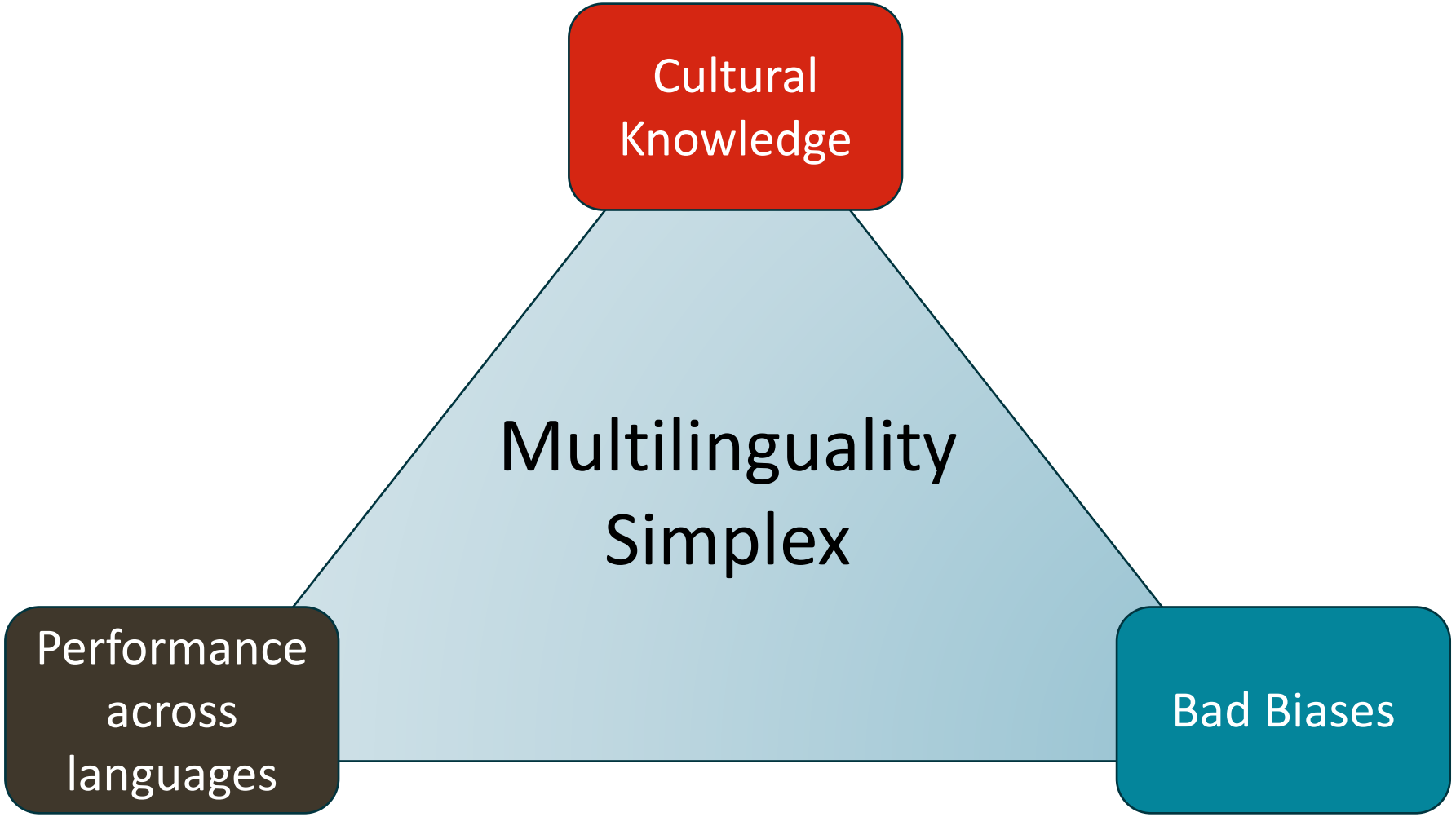
AltDiffusion (the one multilingual model) exhibits lower differential gender bias across languages, less diversity

"woman": note white is now the default, even for JA!

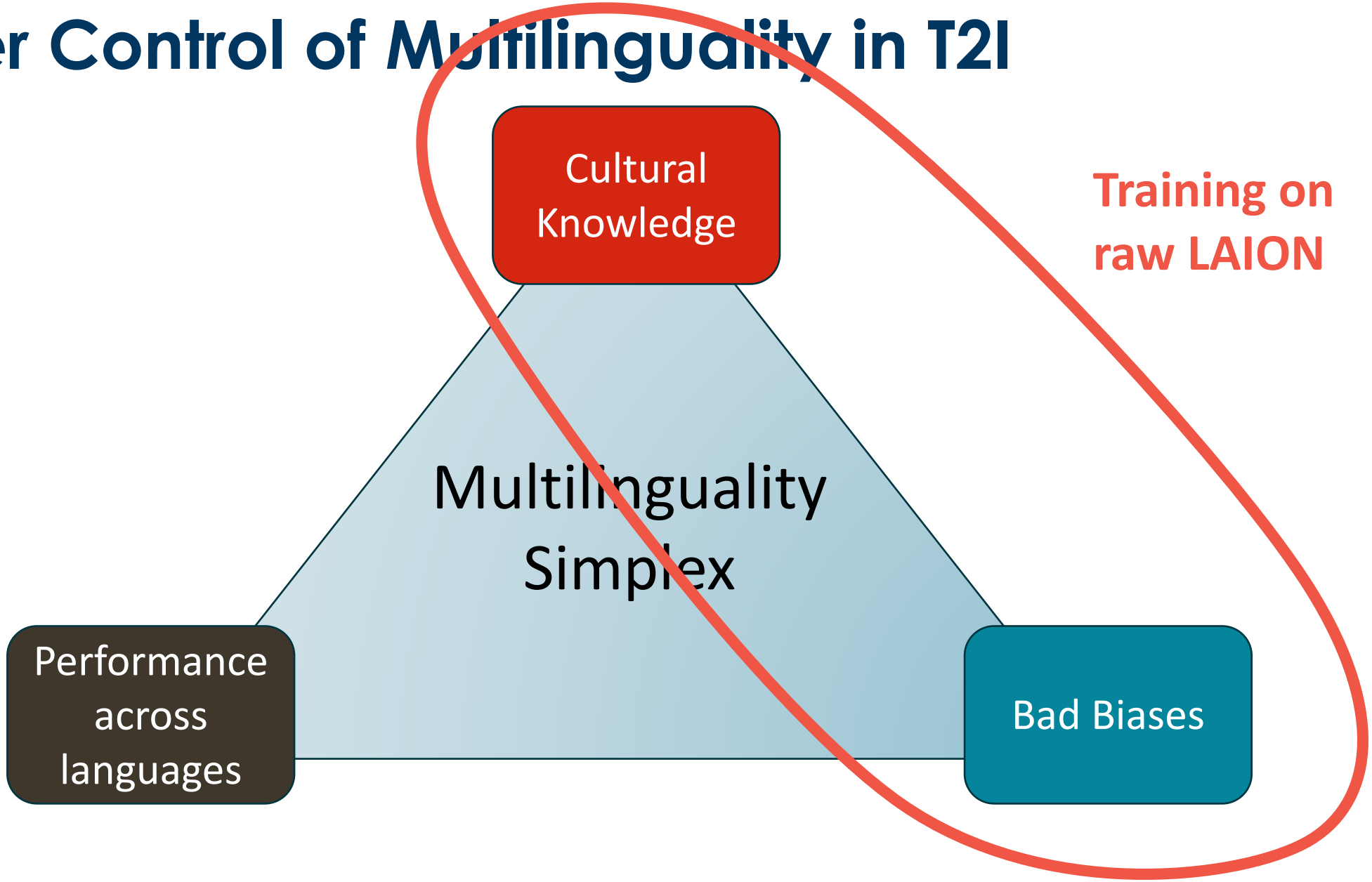
Better Control of Multilinguality in T2I?

- Goals are inherently in conflict
  - Bias mitigation
  - Cultural knowledge preservation
  - Cross-lingual performance
- Existing interventions (and non-interventions) compromise at least 1…
- How can we interpolate between the existing intervention methods?

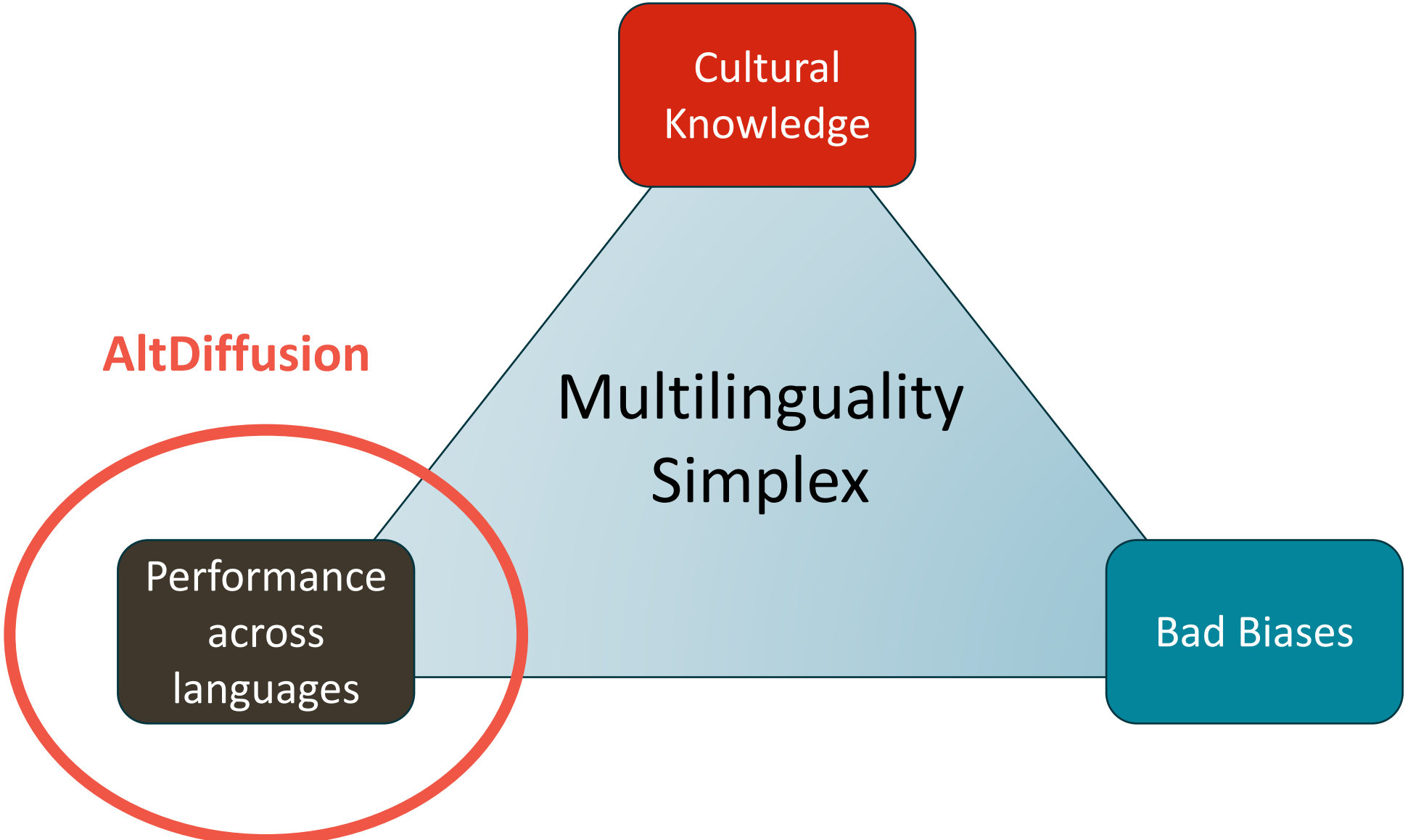
Better Control of Multilinguality in T2I



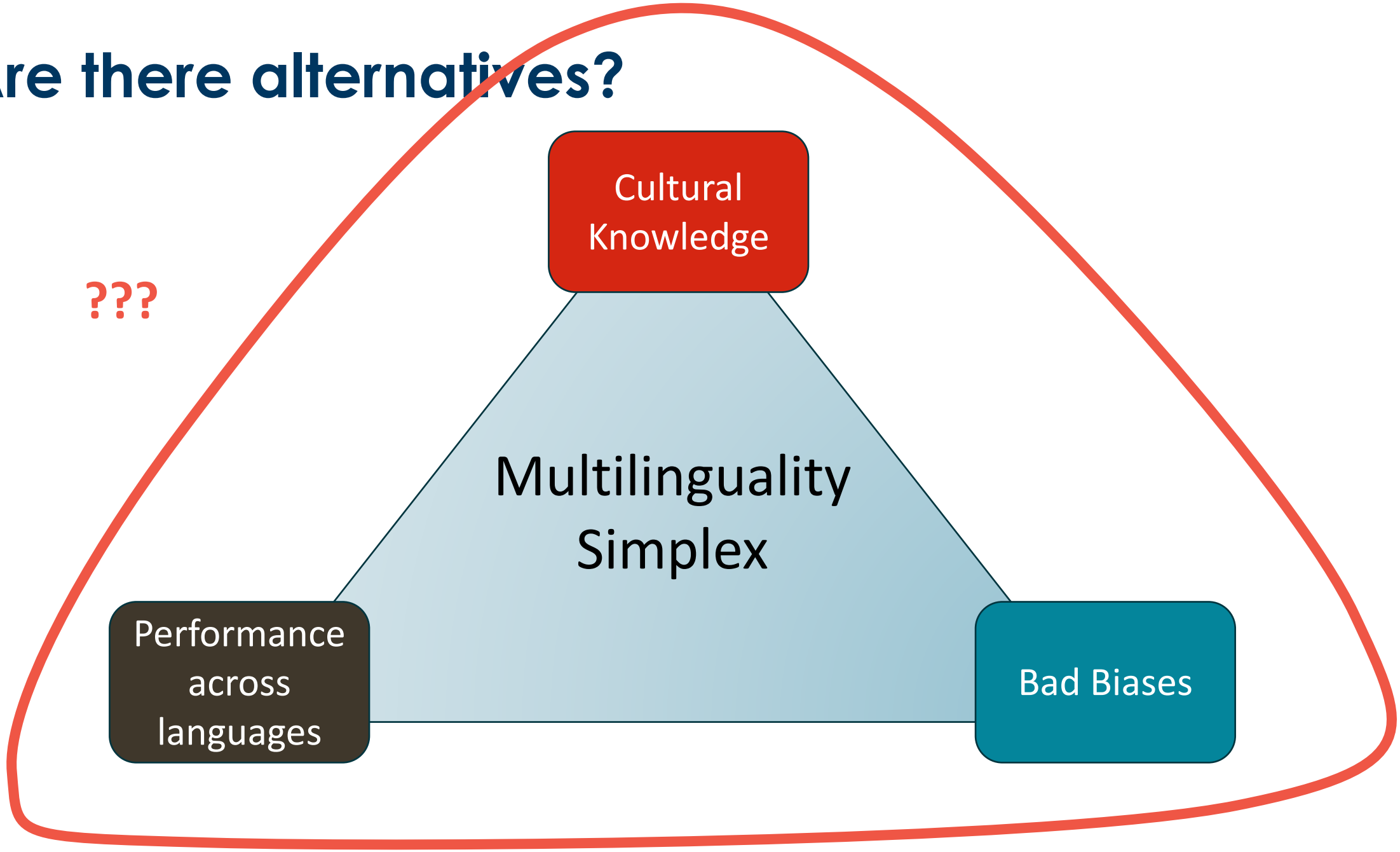
Better Control of Multilinguality in T2I



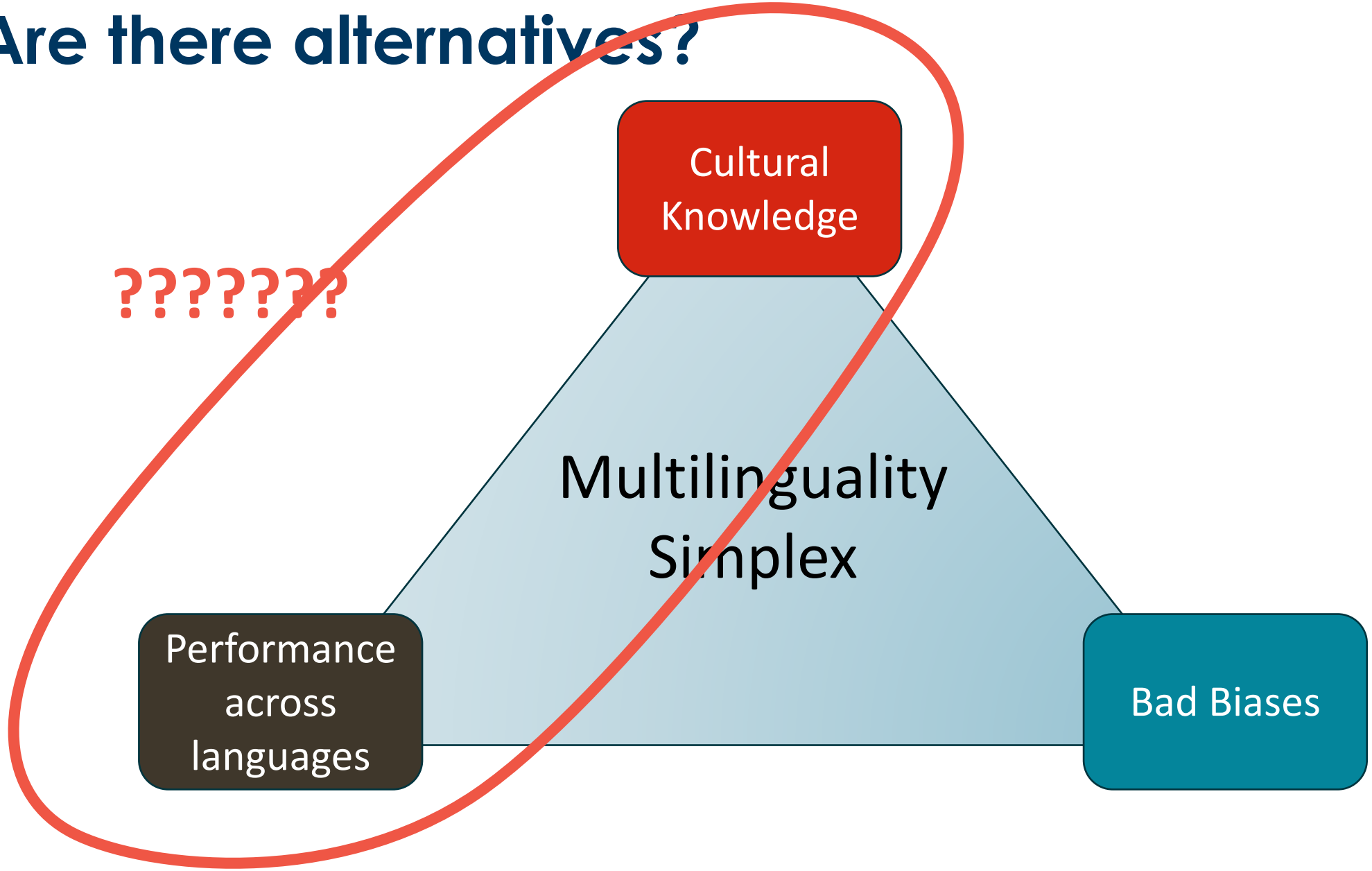
Better Control of Multilinguality in T2I



Are there alternatives?



Are there alternatives?

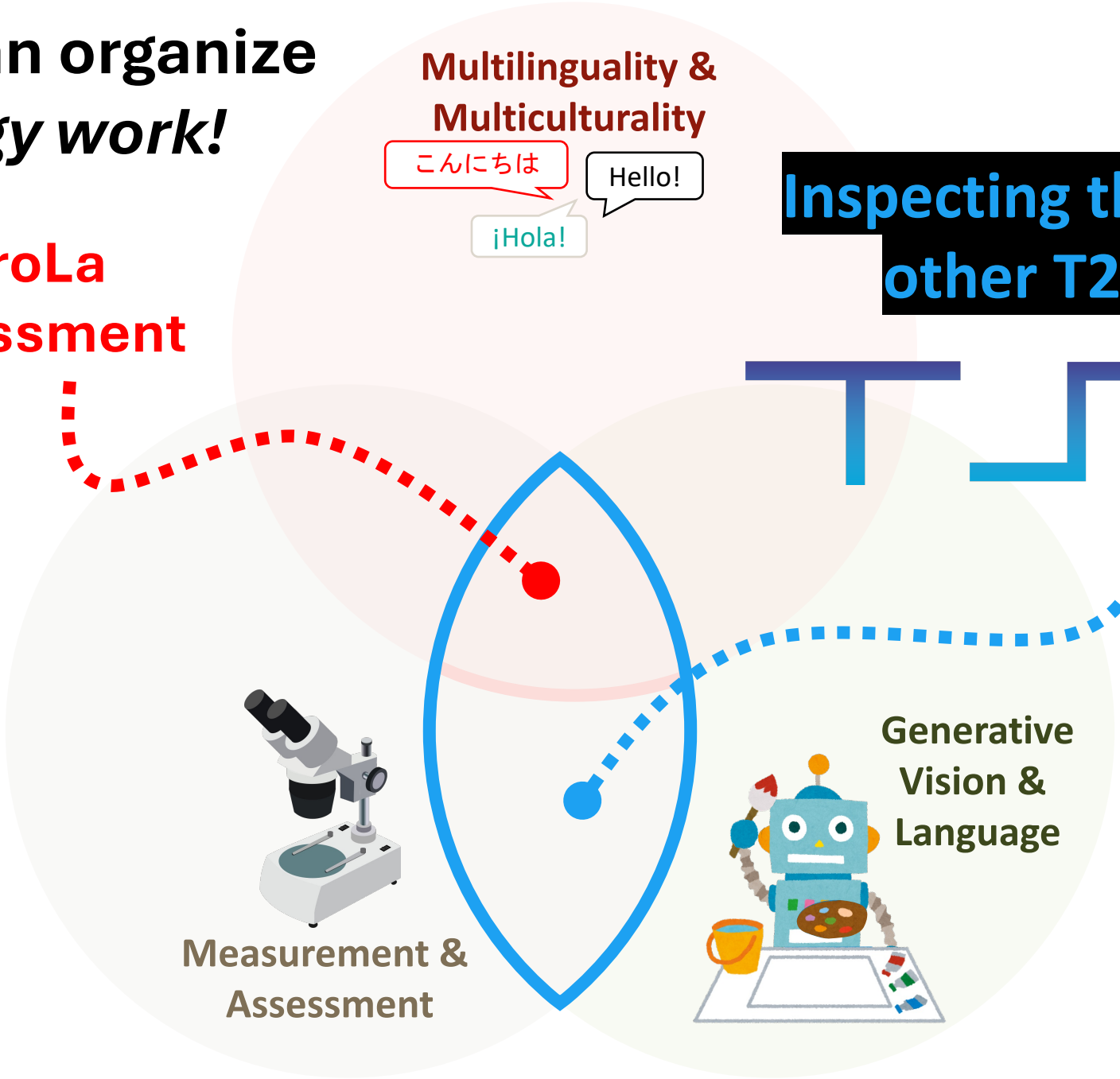


Research directions along “multilinguality simplex” can only be conceived because multilingual capability is *now measurable*

Ok, but why metrology as a field?

Because it can organize real *metrology work!*

CoCo-CroLa
Error Assessment

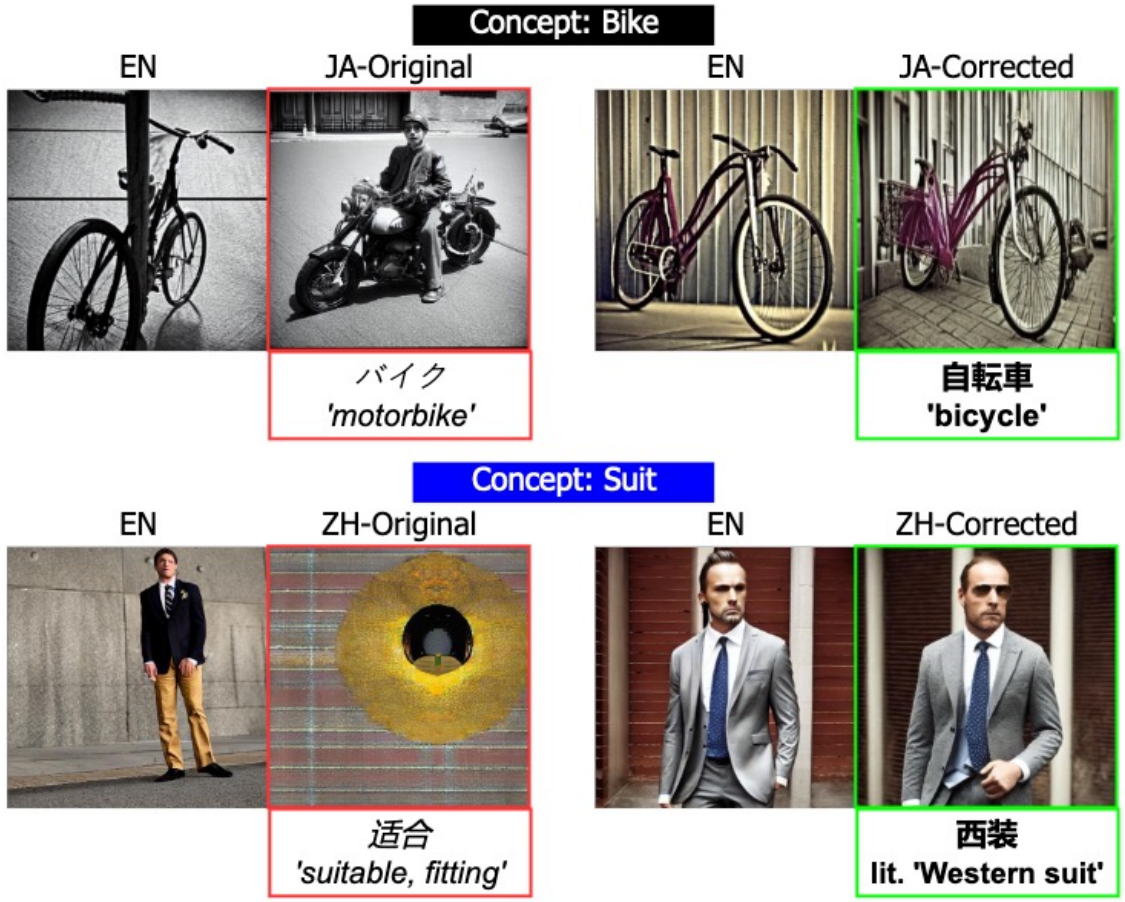


Inspecting the quality of other T2I metrics

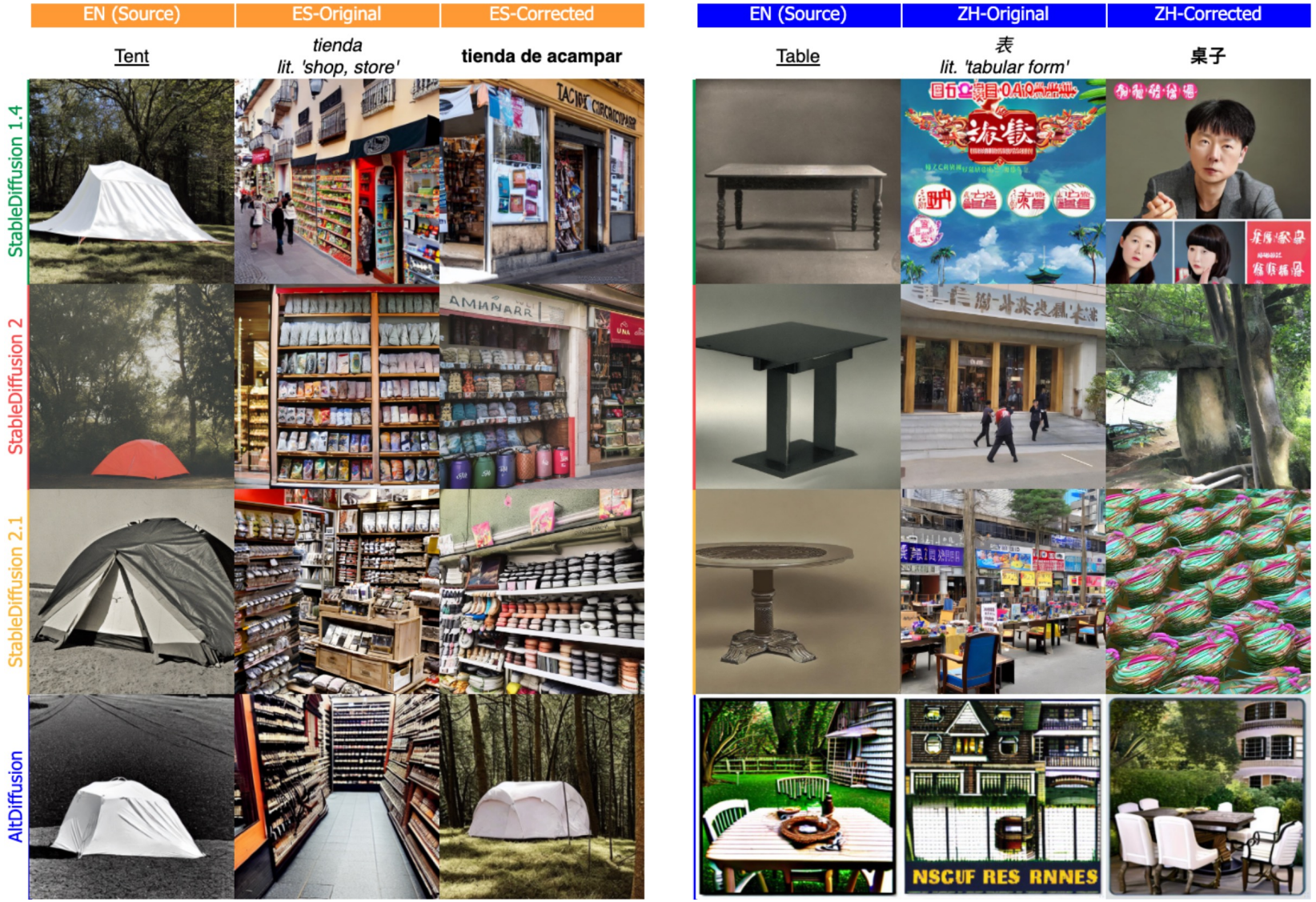
T2

Critiquing my own CoCo-CroLa benchmark

- When I presented this work at ASU, a student fluent in Chinese and Japanese checked my translations
- He found mistranslations and "suspicious translations" among those produced by CCCL v1.0
- Approx. 5-10% of translations depending on strictness
- Some of these errors did in fact cause false negatives to be assigned to models

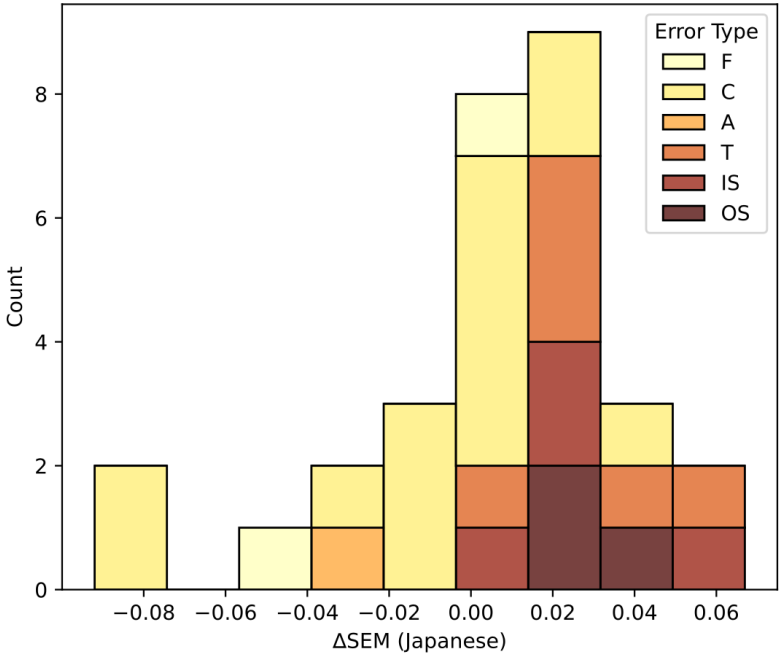


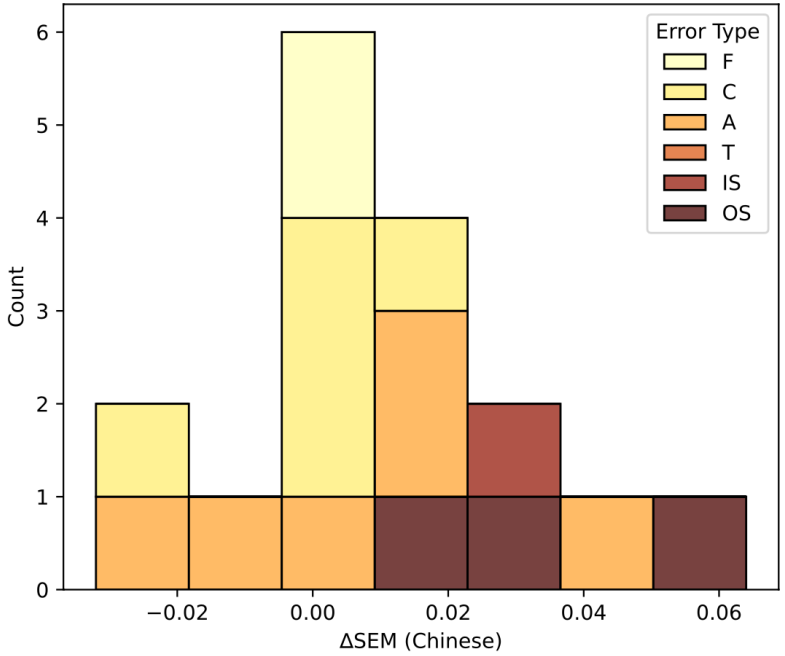
Errors found in CCCL

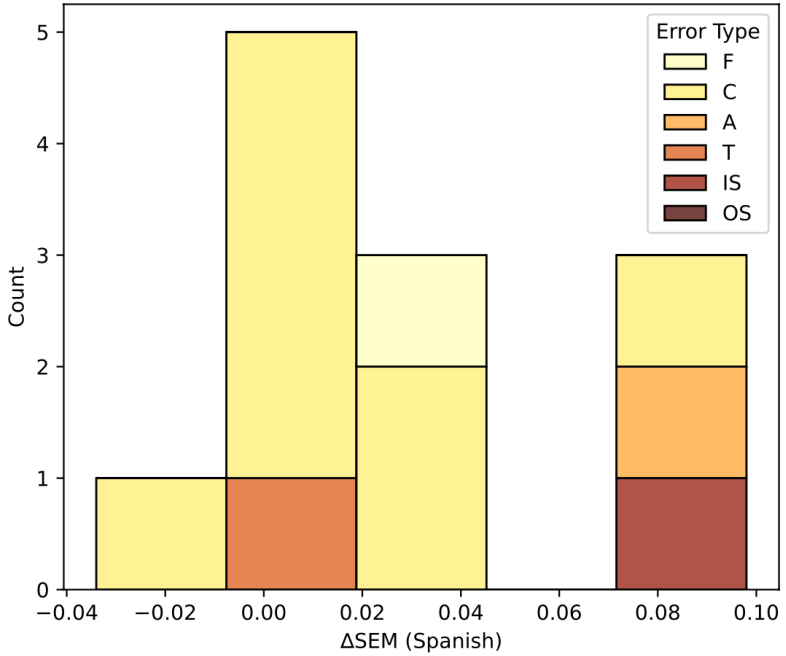


Errors found in CCCL

- We analyzed the errors by "type" (formality, commonality, sense error, etc) and by "severity": ΔSEM
- ΔSEM is the difference in distances between the translation before and after correction in a multilingual embedding space

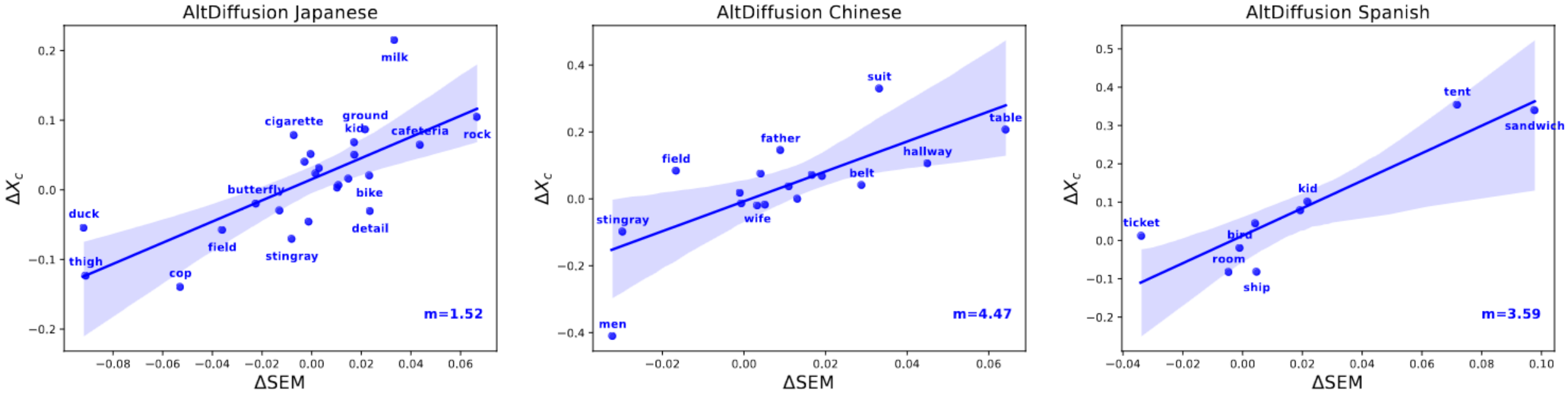






Critiquing my own CoCo-CroLa benchmark

- We did an error analysis, found:
  - ΔSEM correlated to more improvement in correctness score (more severe mistranslation -> more severe false negative)
  - **If the model actually "knows" the test language**



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

Using meta-analysis to improve metrics

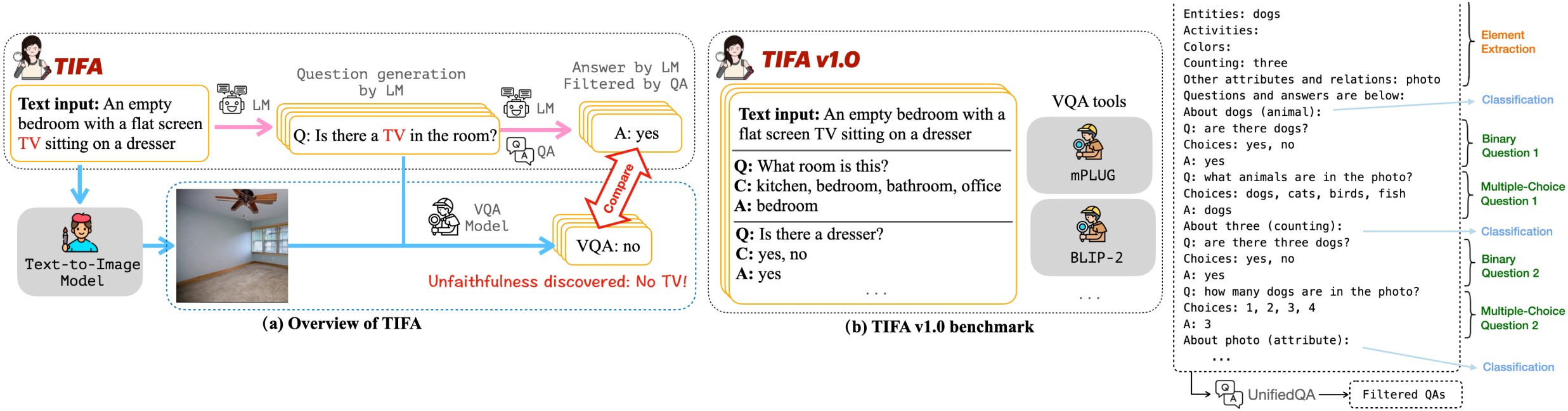
- These failures found in **CCCL** v1 generating process are being used for an update (diff translation proc.)
- Translating these simple sentences of "a picture of a dog" form is a novel MT domain!
- Using a different pipeline we are able to mitigate the errors
- **This is what model metrology work should look like:** diving deep and working to improve benchmark-generating processes

Attempts to build “verb assessments” found errs in other benchmarks

- As mentioned before, we wanted to measure verbal knowledge in T2I in addition to nouns
- We were worried that simple feature-space image comparisons (such as CLIP which we used) would be insufficient as they attend primarily to objects
- So we investigated an area where more sophisticated measures might exist: **prompt faithfulness metrics**

T2I prompt faithfulness

- The task of testing **does a generated image meet the semantic requirements of the prompt**
- Assessed either using multimodal embedding similarity (ie CLIP text-image comparison) or pipelines with VLMs/VQA
- Sophisticated T2I metrics use VLMs, could we adapt to verbs?



Example method: TIFA

- TIFA assesses the **prompt-faithfulness** of a generated image from a T2I model by taking a few steps:
  1. Expand the prompt into a list of requirements
     a dog riding a car -> "is there a dog", "is there a car"…
  2. Answer each expanded question separately using a vision question answering (VQA) setup with a VLM like GPT4, LLaVA, etc
  3. The correctness score of the image is the "success rate" of each question

Prompt-faithfulness example

Stable Diffusion v1.5 | Stable Diffusion v2.1

Text Input: A person sitting on a horse in air over gate in grass with people and trees in background.

GPT-3 generated + verified QAs (pre-generated in TIFA v1.0 benchmark)

Question: what is the animal?
Choices: cow, horse, sheep, dog
Answer: horse
VQA: Horse ✓ Horse ✓

Question: is there a gate?
Choices: yes, no
Answer: yes
VQA: No ✗ Yes ✓

Question: is the horse in air?
Choices: yes, no
Answer: yes
VQA: No ✗ Yes ✓

...

Accuracy on 14 questions

TIFA 71.4 100.0

✓ Fine-Grained ✓ Accurate ✓ Interpretable

Stable Diffusion v1.5 | Stable Diffusion v2.1

Text Input: Three cats and two dogs sitting on the grass.

GPT-3 generated + verified QAs (pre-generated in TIFA v1.0 benchmark)

Question: are there cats?
Choices: yes, no
Answer: yes
VQA: Yes ✓ Yes ✓

Question: are there dogs?
Choices: yes, no
Answer: yes
VQA: No ✗ Yes ✓

Question: how many cats are in the picture?
Choices: 1, 2, 3, 4
Answer: 3
VQA: 3 ✓ 4 ✗

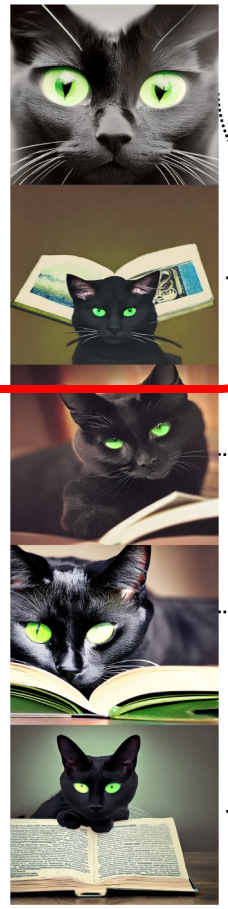
...

Accuracy on 8 questions

TIFA 62.5 62.5

✓ Fine-Grained ✓ Accurate ✓ Interpretable

When we tried adapting it to our setting, issues arose…



A photo of a black cat with green eyes reading a book.

TIFA Scores:

0.6666666667

0.6666666667

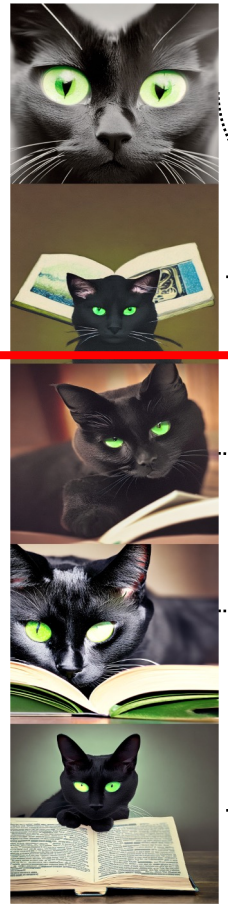
0.8888888889

1

1

One has a book, other doesn't, same score?!

When we tried adapting it to our setting, issues arose…



A photo of a black cat with green eyes reading a book.

TIFA Scores:

0.6666666667

0.6666666667

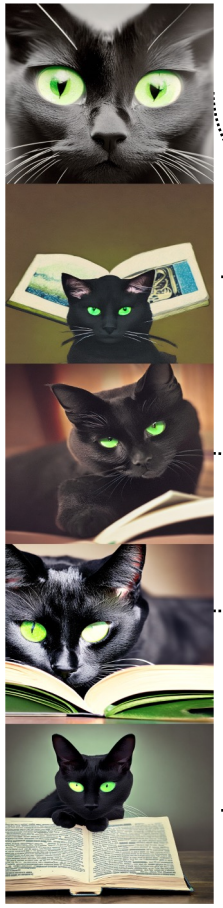
0.8888888889

1

1

All equally correct, one is lower!?

When we tried adapting it to our setting, issues arose…



A photo of a black cat with green eyes reading a book.

TIFA Scores:

0.6666666667

0.6666666667

0.8888888889

1

1

Errors are in same dynamic range as distinctions????

This finding presented another metrology opportunity…

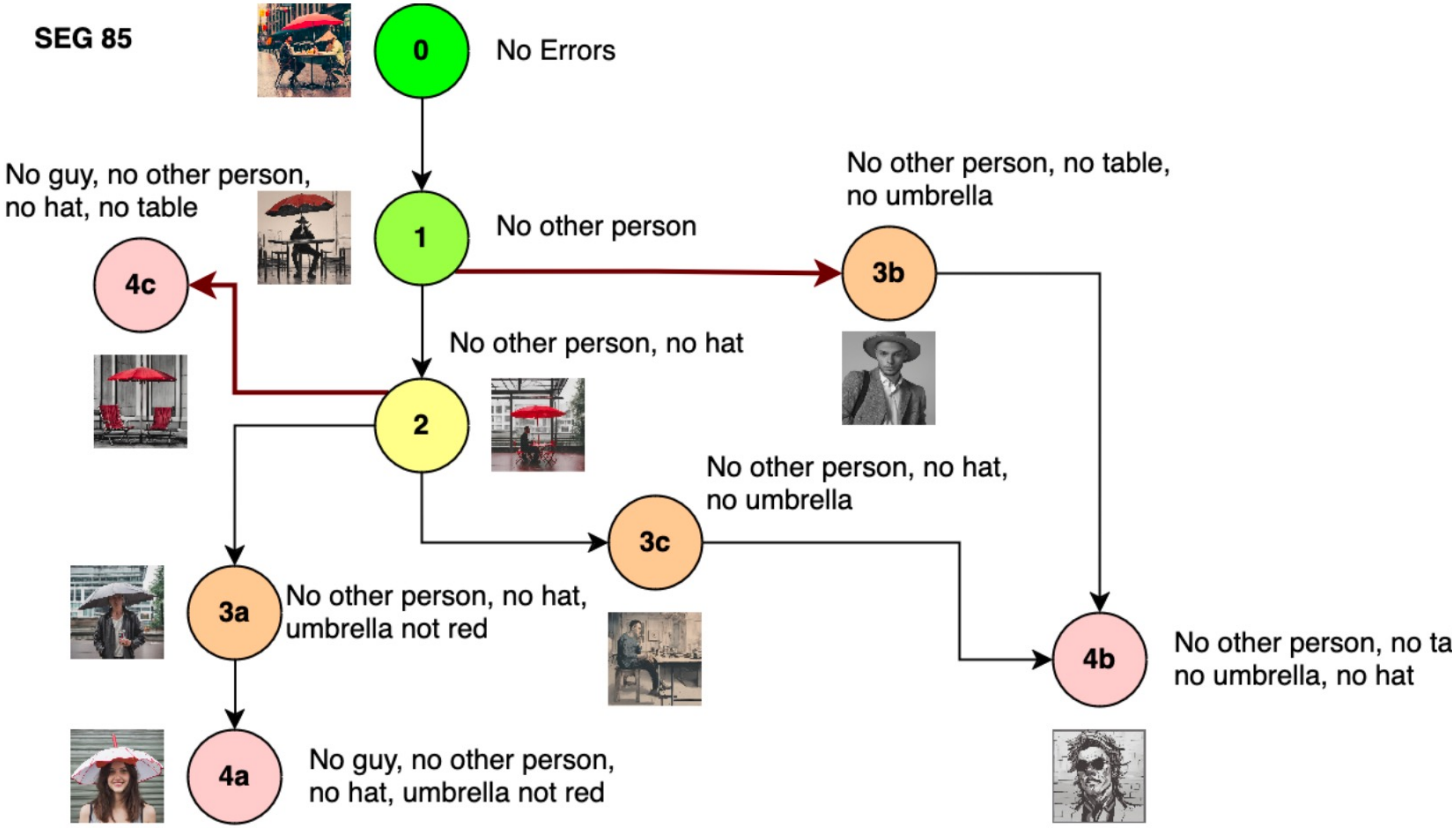
- Each new T2I faithfulness metric is being **evaluated separately**
- Ad-hoc eval sets put together by the same people proposing them, games the numbers a bit?
- They test against human preference correlation
- These eval sets **rarely test what we actually care about:**
- Ability to discriminate **similar images** wrt **same prompt**
- We designed our eval with capturing these characteristics in mind

Other methods evaluated

- *Correlation metrics*:
- **CLIPScore/ALIGNScore**: cosine similarity between image features and text features from prompt w/ CLIP or ALIGN
- *VQA metrics*
- **DSG** (Davidsonian Scene Graph): same as TIFA but using only yes/no questions over a graph to eval; all child nodes of a no are assessed as no
- *Captioning Metrics*:
- **LLMScore**: Use a captioning model like MPLUG to get a caption of the image, count errors between prompt and caption using LLM
- **VIEScore:** prompt a VLM to score the img wrt prompt

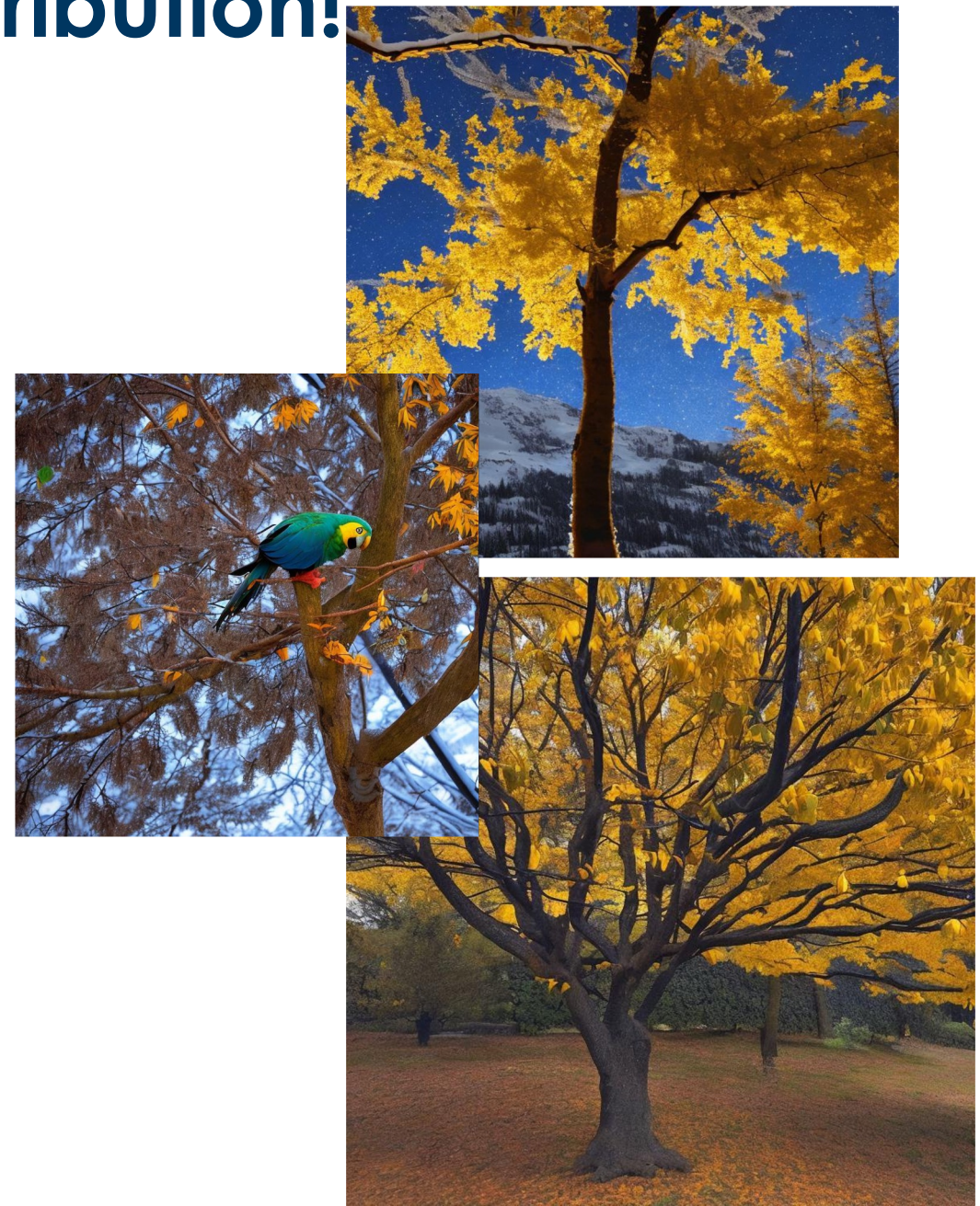
Capturing metric performance on similar images:

- We propose evaluating metrics on their ability to order and separate many images wrt the same prompt
- We organize these images on "**semantic error graphs**" to assess



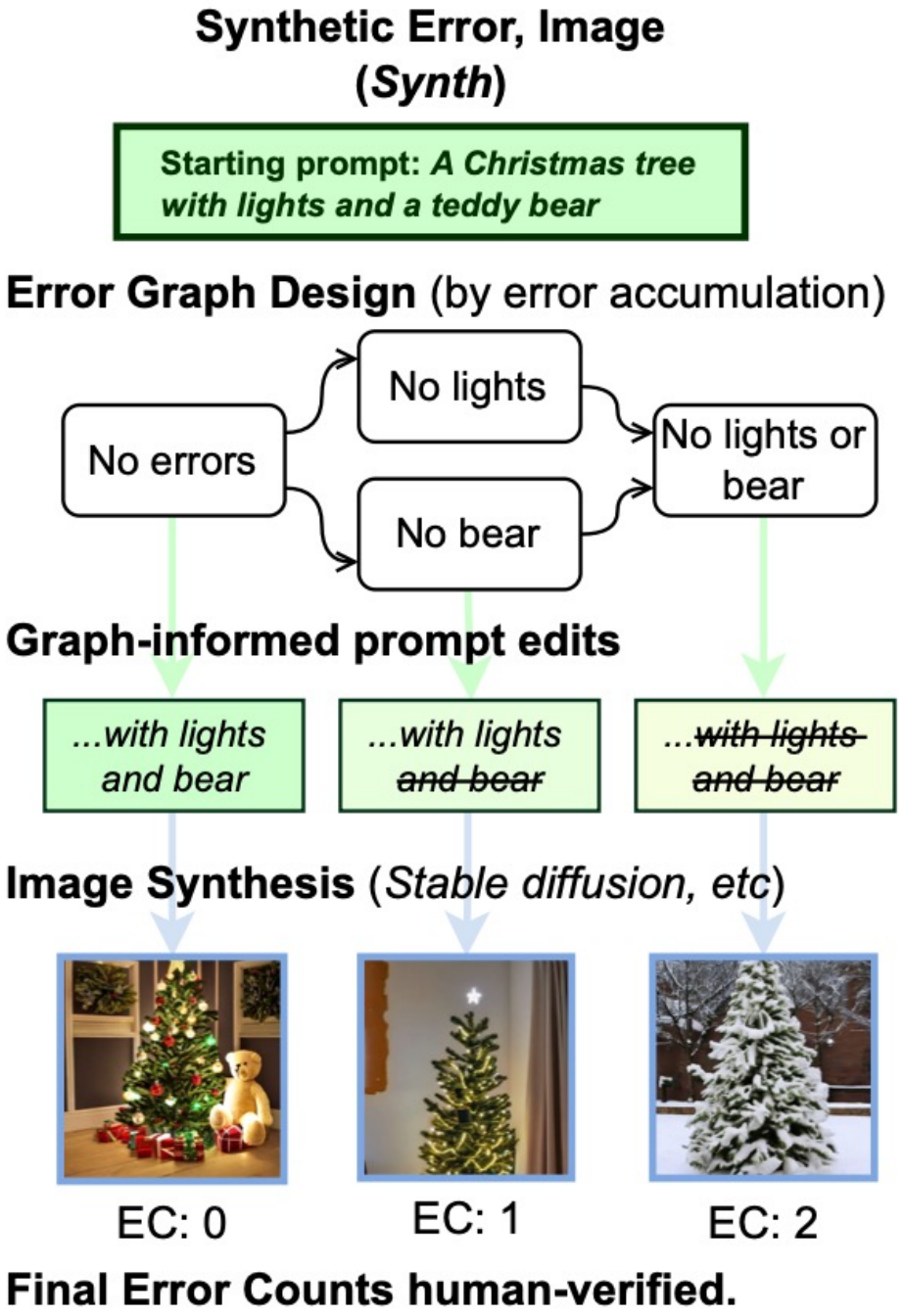
Construct validity of eval distribution!

- We consider different types of "error generating functions" to produce the SEGs
- **Naturally-occuring errors produced by a model** should be harder for a metric to identify than **synthetic errors** which we produced using prompt editing
- Ex: "a green bird in yellow trees against a blue sky"



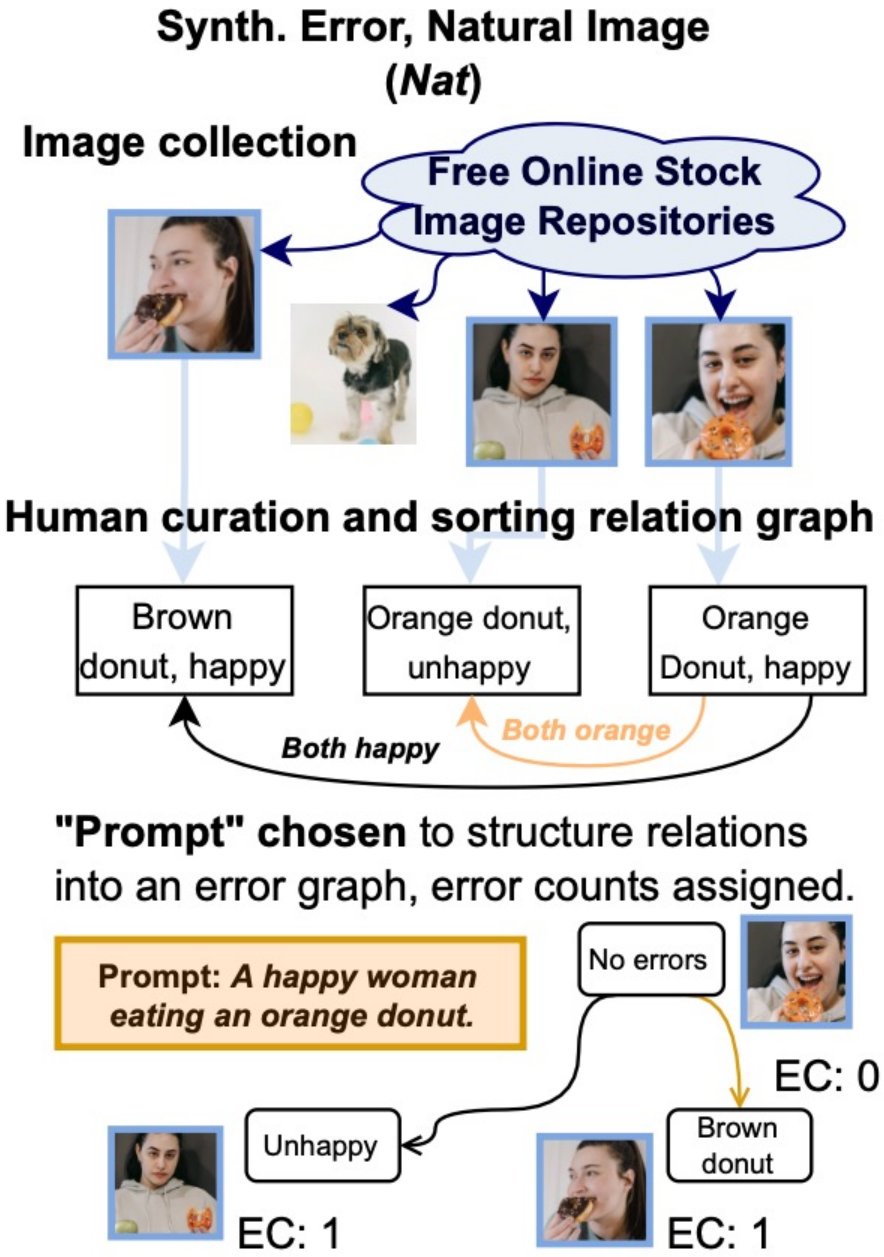
SEG method 1: Synth

- "Synthetic errors on synthetic images"
- We start with a COCO prompt
- Design the error graph by ablation of features
- Generate multiple images for each prompt to contain the errors
- Verify correctness



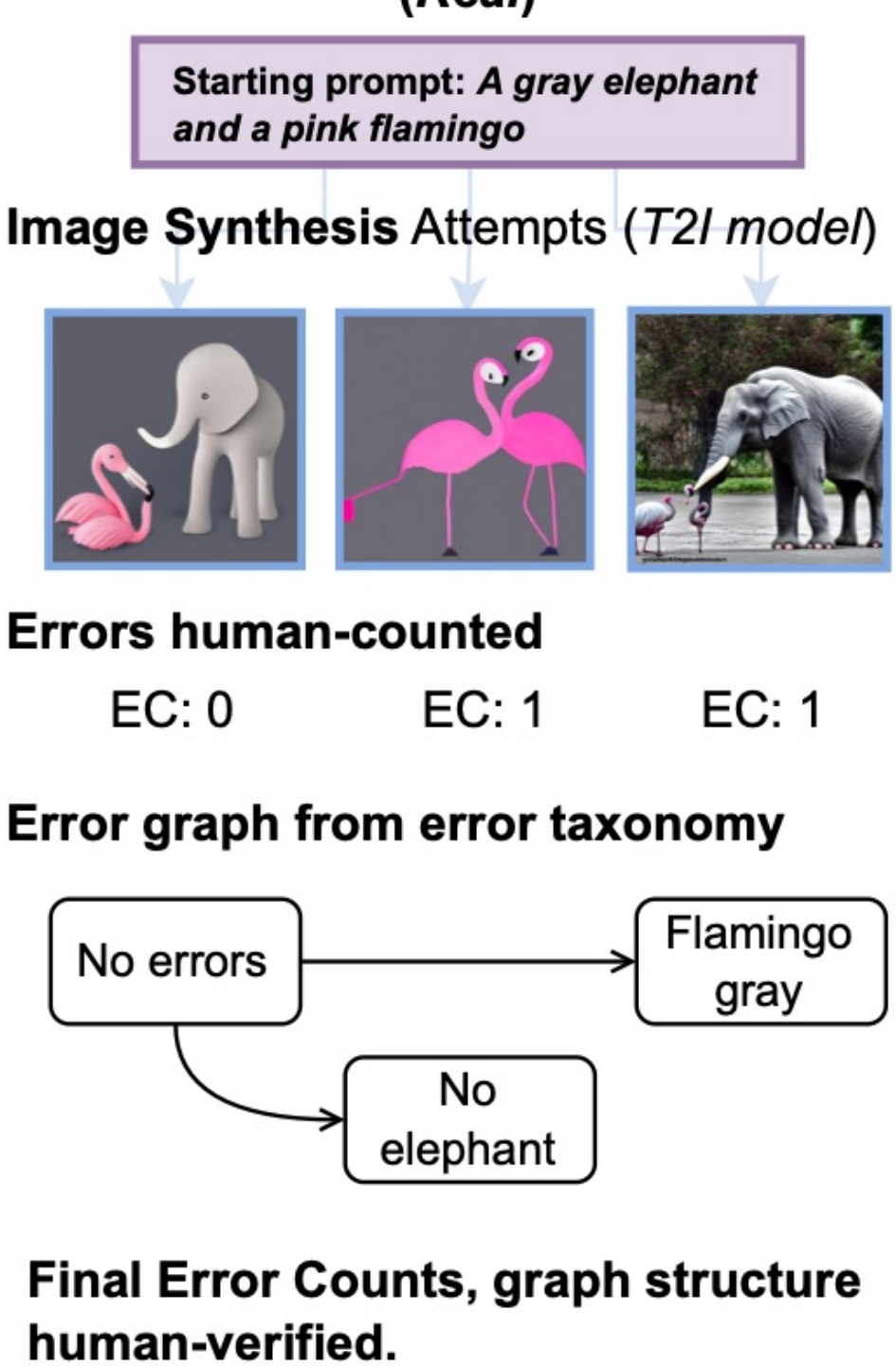
SEG method 2: Nat

- "Synthetic errors on natural images"
- Start with free stock photos
- We organize them by description and a graph of differences
- We select one image to be the "head node" and difference counts become error counts
- Write a "prompt" for the set

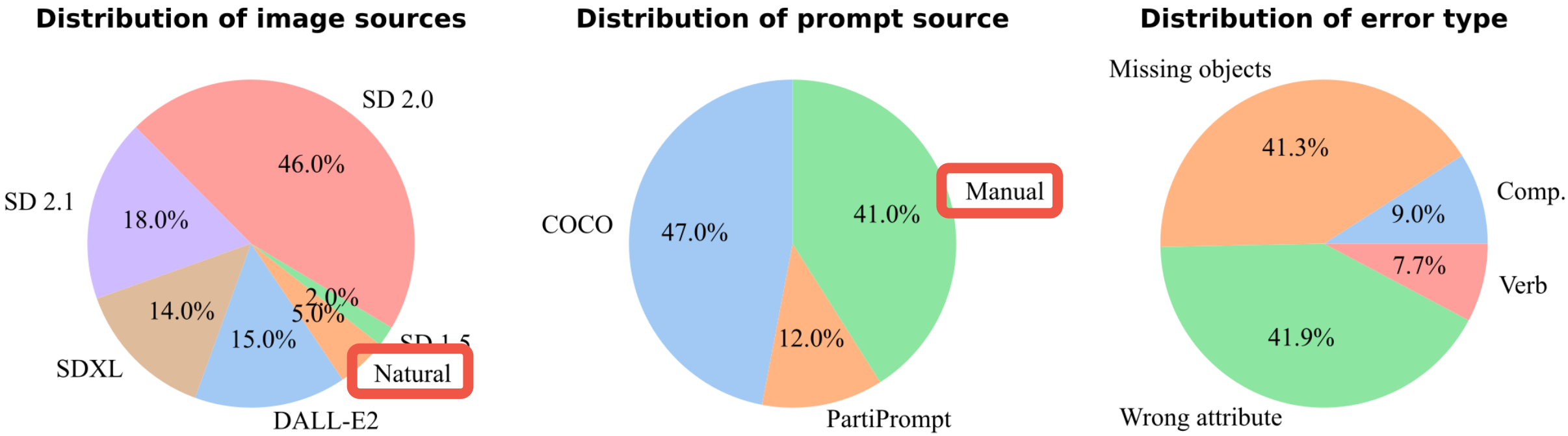


SEG method 3: Real

- "Real errors on synthetic images"
- Use a single prompt
- Generate images, which will include failures
- Count errors and organize the graph
- **This is by far the hardest partition**



We gather SEGs and their images from many sources



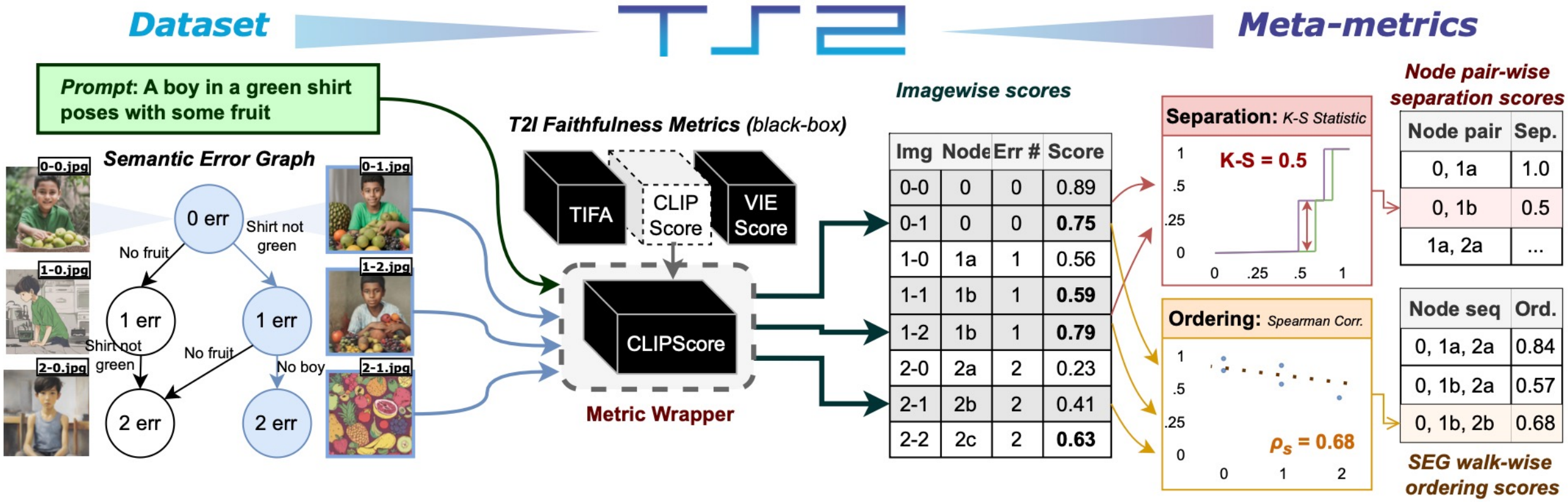
(a) Image Source (b) Prompt Source (c) Error node (SEG edge) type

In case natural images being ID wrt VLM, CLIP improves

To generate errors on artificial SEG graphs, to describe natural images

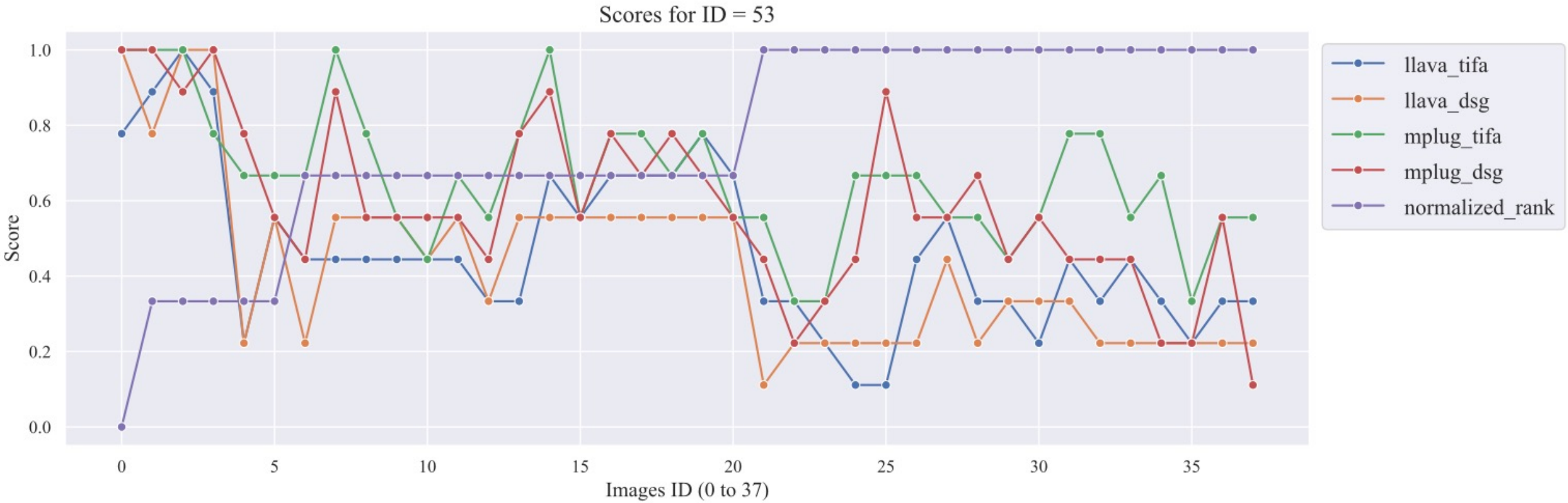
Leads to "T2IScoreScore" meta-metric (TS2)

- We slot in a metric like a black box, assess its scores for every image in every SEG, and check if the individual nodes are **separated** and **ordered** by a metric



Ordering Meta-Metric: Spearman complexity over all walks

- Over a graph walk of ascending error count, the score assigned by a good metric should monotonically decrease
- Goal: negative rank-order correlation



Separation Meta-metric: Kolmogorov-Smirnov

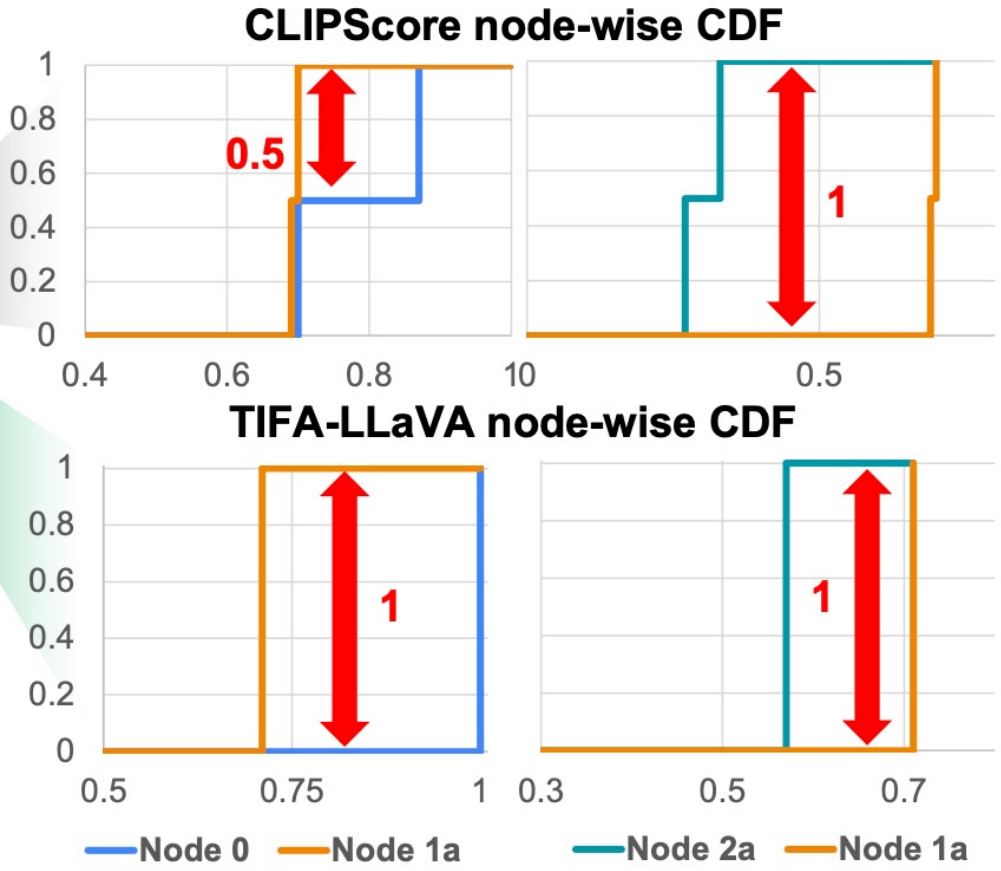
- Tests if two distributions are "different" (in this case, scores assigned to images on two adjacent nodes in SEG)

(a)

Node	0	0	1b	1b	2b	2b
CLIPScore↑	0.87	0.7	0.7	0.69	0.33	0.27
TIFA-LLaVA↑	1	1	0.71	0.71	0.57	0.57
LLMScore↓	0	1	6	1	14	5

(b)

Node	0	1a	1a	1a	1a	1a
CLIPScore↑	0.65	0.54	0.5	0.49	0.57	0.52
TIFA-LLaVA↑	0.4	0.4	0.2	0.4	0.4	0.4
LLMScore↓	6	4	4	4	4	5

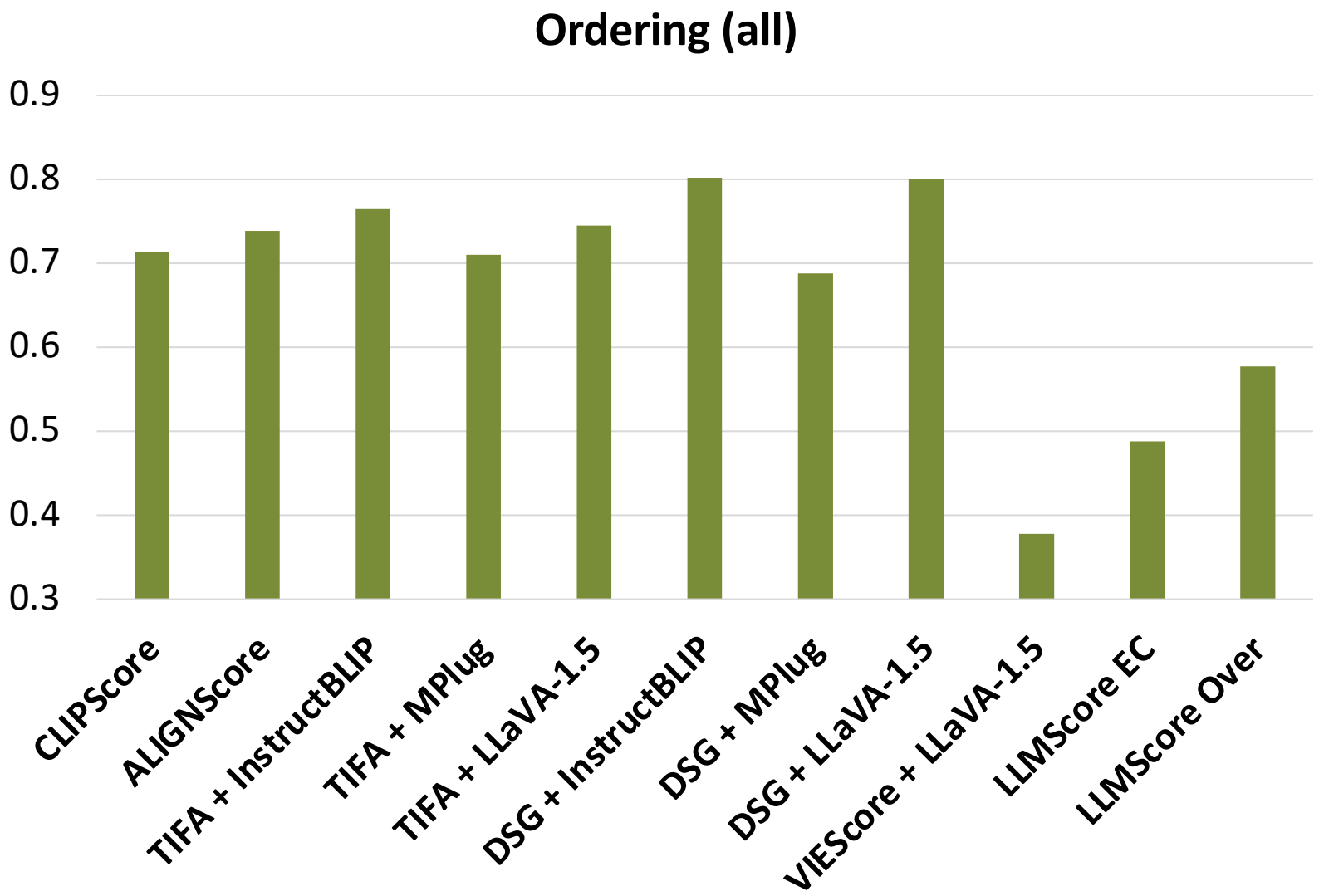


Results

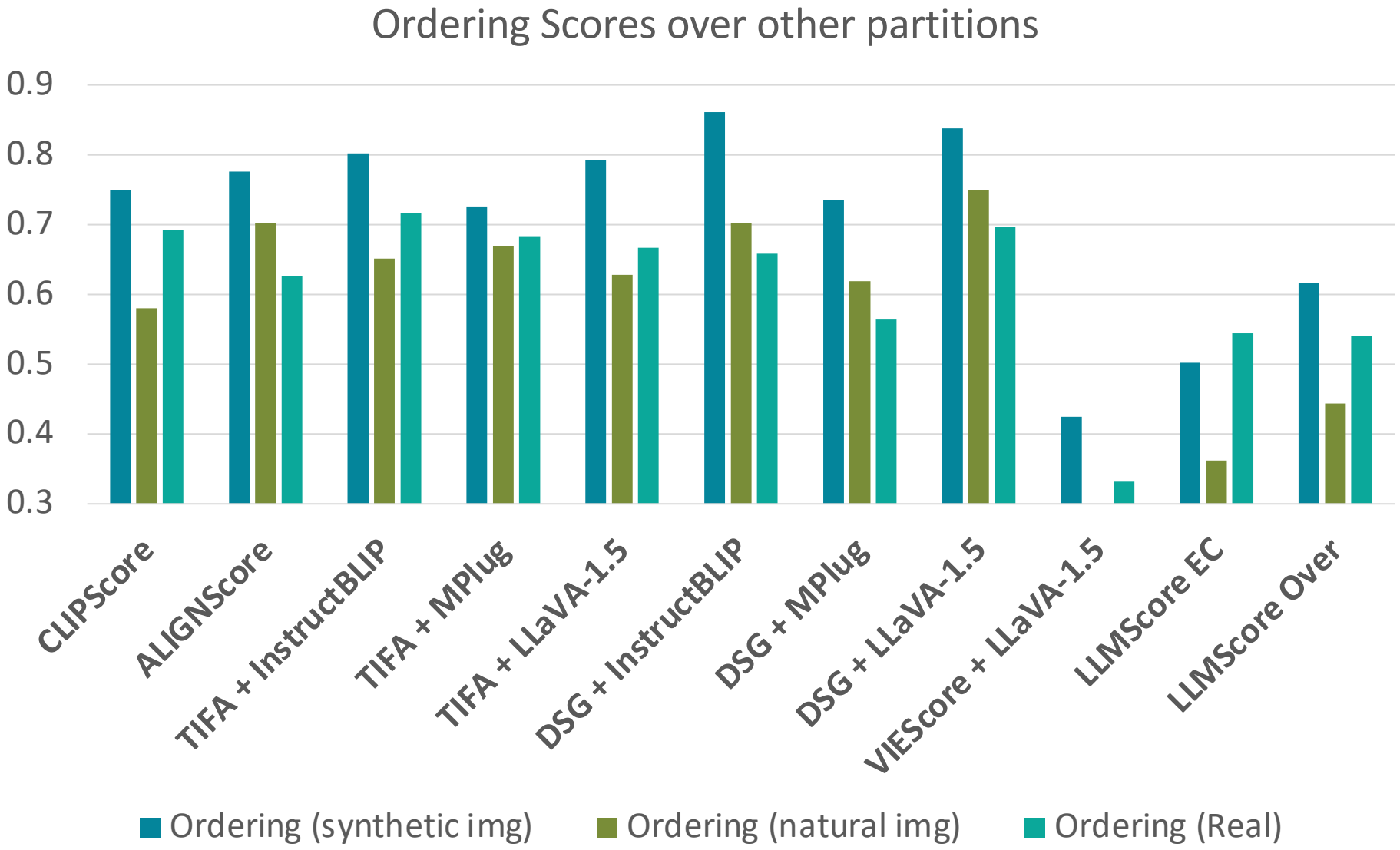
- **Ordering**: the VQA-based metrics only **slightly** better than correlation metrics, despite claims of significant superiority
- **Separation**: correlation significantly outperforms

		CLIPScore	ALIGNScore	mPLUG	LLaVA 1.5	LLaVA 1.5 (alt)	InstructBLIP	BLIP1	Fuyu	mPLUG	LLaVA 1.5	LLaVA 1.5 (alt)	InstructBLIP	BLIP1	Fuyu	LLMScore EC	LLMScore Over	VIEScore
		Emb-based		TIFA						DSG						Caption-based		
Ord.	Avg	71.4	73.9	71.0	74.5	74.4	76.5	73.8	38.7	70.4	76.2	75.0	**79.0**	76.6	29.5	48.8	57.7	37.8
	Synth	75.0	77.6	72.6	79.2	79.2	80.2	78.8	44.5	74.6	80.1	81.6	**85.1**	81.6	35.4	50.2	61.6	42.5
	Nat	58.0	*70.2*	66.9	62.8	64.0	65.1	62.2	23.5	65.3	65.9	68.8	*70.7*	**71.6**	20.5	36.2	44.4	22.4
	Real	*69.3*	62.6	*68.2*	66.7	64.5	71.6	64.0	29.7	58.4	**70.0**	54.2	62.0	61.2	14.2	54.4	54.1	33.2
Sep.	Avg	*90.7*	**92.8**	80.6	84.3	81.9	85.0	81.8	67.2	78.4	83.1	80.3	84.2	80.8	63.6	73.6	73.5	51.8
	Synth	90.5	**94.1**	80.6	87.5	85.2	86.7	84.1	67.3	80.9	85.7	83.9	87.8	84.9	65.8	71.1	72.8	53.7
	Nat	*91.5*	**92.6**	84.2	83.4	75.6	82.8	77.9	75.7	71.2	80.9	76.7	81.8	73.3	68.6	80.5	76.7	44.5
	Real	**90.3**	87.9	77.4	72.7	74.4	80.5	76.4	59.3	75.1	74.5	69.4	71.9	70.8	50.3	77.3	73.6	50.7

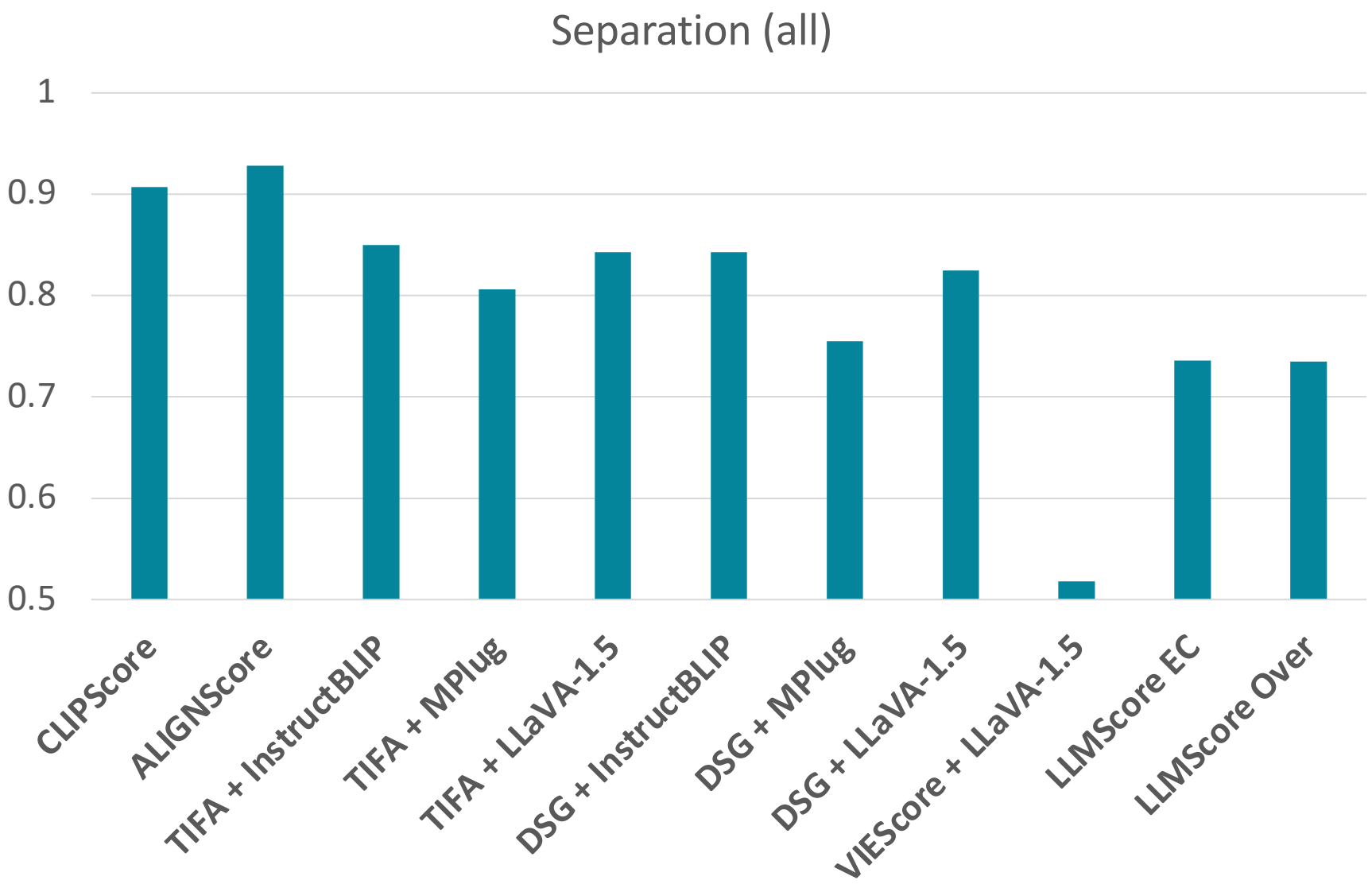
Findings


Ordering (all)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
CLIPScore
ALIGNScore
TIFA + InstructBLIP
TIFA + MPlug
TIFA + LLaVA-1.5
DSG + InstructBLIP
DSG + MPlug
DSG + LLaVA-1.5
VIEScore + LLaVA-1.5
LLMScore EC
LLMScore Over

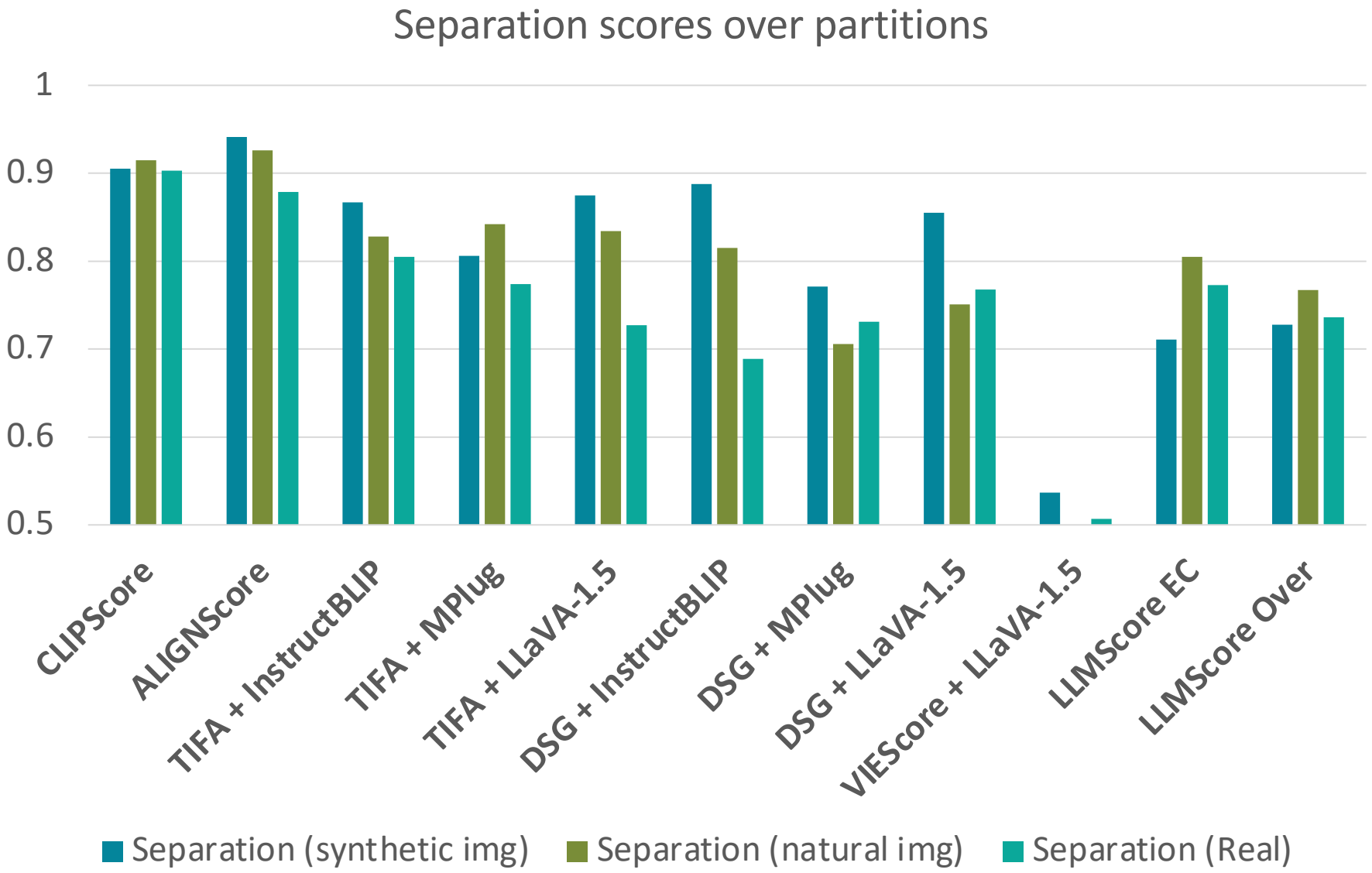
Findings


Ordering Scores over other partitions
0.9
0.8
0.7
0.6
0.5
0.4
0.3
CLIPScore
ALIGNScore
TIFA + InstructBLIP
TIFA + MPlug
TIFA + LLaVA-1.5
DSG + InstructBLIP
DSG + MPlug
DSG + LLaVA-1.5
VIEScore + LLaVA-1.5
LLMScore EC
LLMScore Over
Ordering (synthetic img)
Ordering (natural img)
Ordering (Real)

Findings


Separation (all)
1
0.9
0.8
0.7
0.6
0.5
CLIPScore
ALIGNScore
TIFA + InstructBLIP
TIFA + MPlug
TIFA + LLaVA-1.5
DSG + InstructBLIP
DSG + MPlug
DSG + LLaVA-1.5
VIEScore + LLaVA-1.5
LLMScore EC
LLMScore Over

Findings


Separation scores over partitions
1
0.9
0.8
0.7
0.6
0.5
CLIPScore
ALIGNScore
TIFA + InstructBLIP
TIFA + MPlug
TIFA + LLaVA-1.5
DSG + InstructBLIP
DSG + MPlug
DSG + LLaVA-1.5
VIEScore + LLaVA-1.5
LLMScore EC
LLMScore Over
Separation (synthetic img)
Separation (natural img)
Separation (Real)

Failure causes

- **Turns out VLMs VQAs hallucinate in the same ways that T2I models do!**
- **Makes sense, T2I built on VQA-like models**

SEG 143

Prompt: A yellow school bus and a red stop sign and a blue car.

Node 0

- Yellow school bus ✅
- Blue car ✅
- Red stop sign ✅

Target Score = 1

- Q1: Is there a yellow school bus?
  VQA: *Yes there is a yellow school bus in the image. (Correct, **should be yes**)*
- Q2: Is there a red stop sign?
  VQA: *No (Incorrect, **should be yes**)*
- Q3: Is there a blue car?
  VQA: *No (Incorrect, **should be yes**)*

Correct DSG Score	DSG LLaVA	CLIPScore
1 (3/3)	0.33	0.67

Node 1a

- Yellow school bus ❌
- Blue car ✅
- Red stop sign ✅

Target Score = 0.67

- Q1: Is there a yellow school bus?
  VQA: *Yes there is a yellow school bus in the image. (Incorrect, **should be no**)*
- Q2: Is there a red stop sign?
  VQA: *No (Incorrect, **should be yes**)*
- Q3: Is there a blue car?
  VQA: *No (Incorrect, **should be yes**)*

Correct DSG Score	DSG LLaVA	CLIPScore
0.67 (2/3)	0.33	0.56

Takeaways

Deep metrology work is worth the effort!

- **New measures enable new directions:**
- CoCo-CroLa multilinguality simplex is only assessable because we can measure multilingual capability!
- TS2 measures of metric quality enables objective assessment, finds previously unknown weaknesses
- **Problems with measures create new technical requirements:**
- Better hallucination-proofed VQAs, strategies required to improve T2I metrics
- New translation approaches needed to mitigate CCCL errs
- **Better measures for better science:**
- CCCL for bottom-up approach to reasoning assmt in T2I?

こんにちは

Hello!

¡Hola!

CoCoCroLa

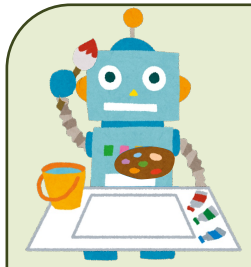
Sharon Levy (->Rutgers)

Yiran Luo (ASU)

Yezhou Yang (ASU)

My Boss

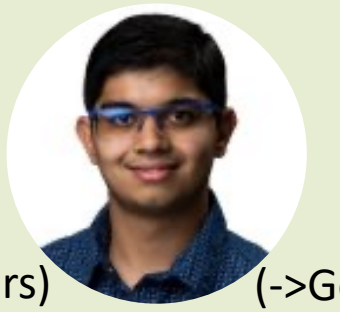
William Wang



TS2

Mahsa Khoshnoodi (->SBU)

Fatima Jahara (->Rutgers)



Aditya Sharma (->Google)

Yujie Lu (UCSB)

Model Metrology

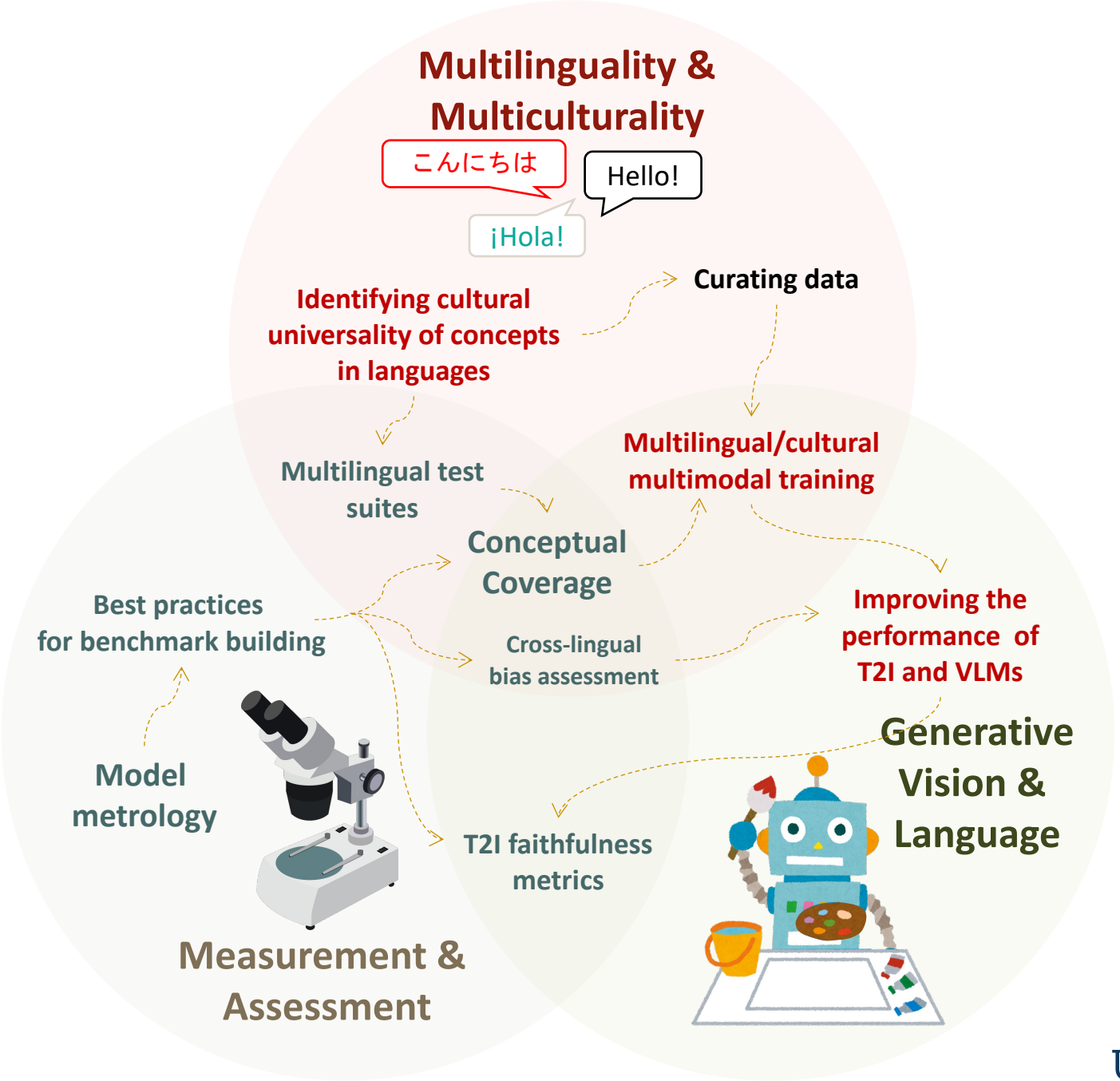
Peter West (UW)

Ari Holtzman (->U Chicago)

Naomi Saphra (Harvard)

Literally me

Fin.



@m2saxon

UC SANTA BARBARA