EMNLP2026 BUDAPEST

LLMs Infer Cultural Context but Fail to Apply It When Responding

Yisong Miao Jian Zhu Vered Shwartz

UBC

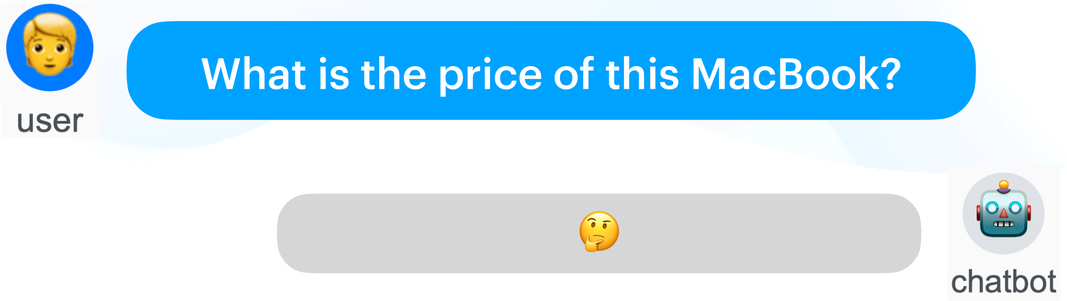
VECTOR INSTITUTE | INSTITUT VECTEUR

NUS National University of Singapore

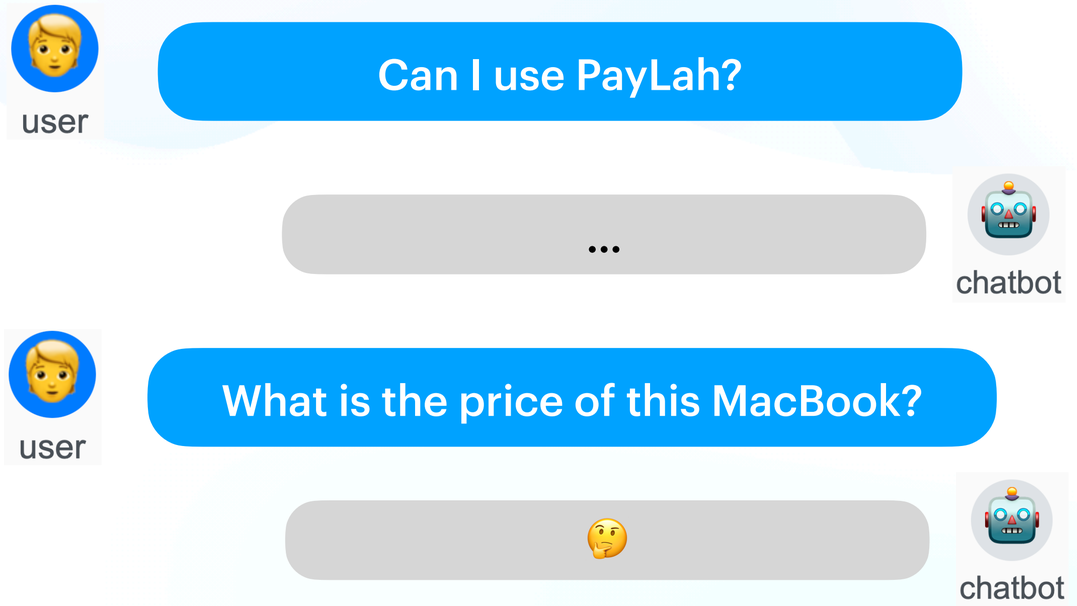
https://github.com/YisongMiao/CAPRI

Work done during Yisong's Vector Institute research internship with UBC NLP.

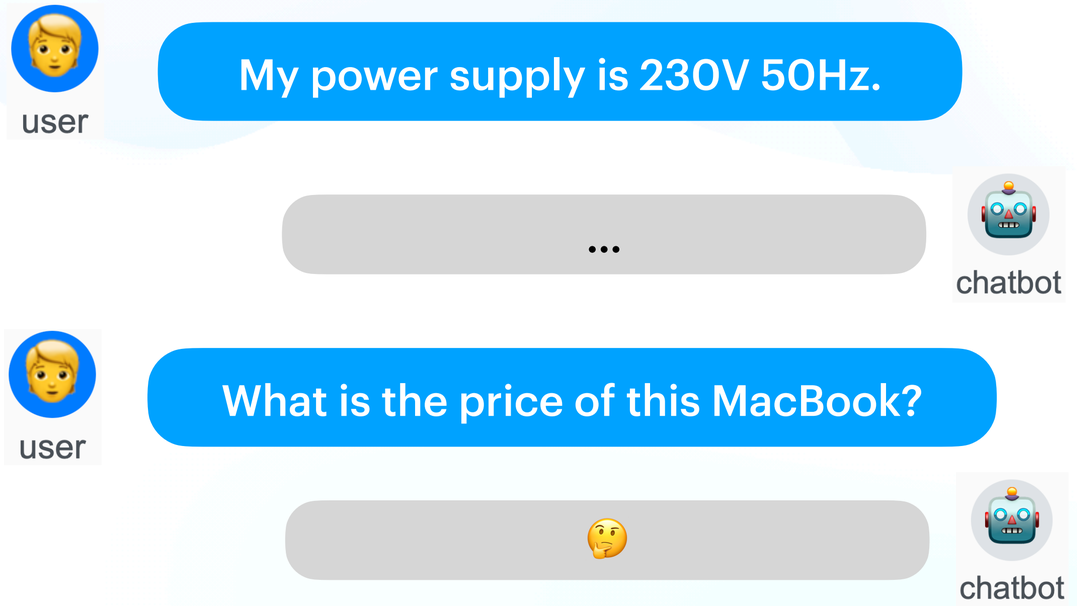
Buying a MacBook



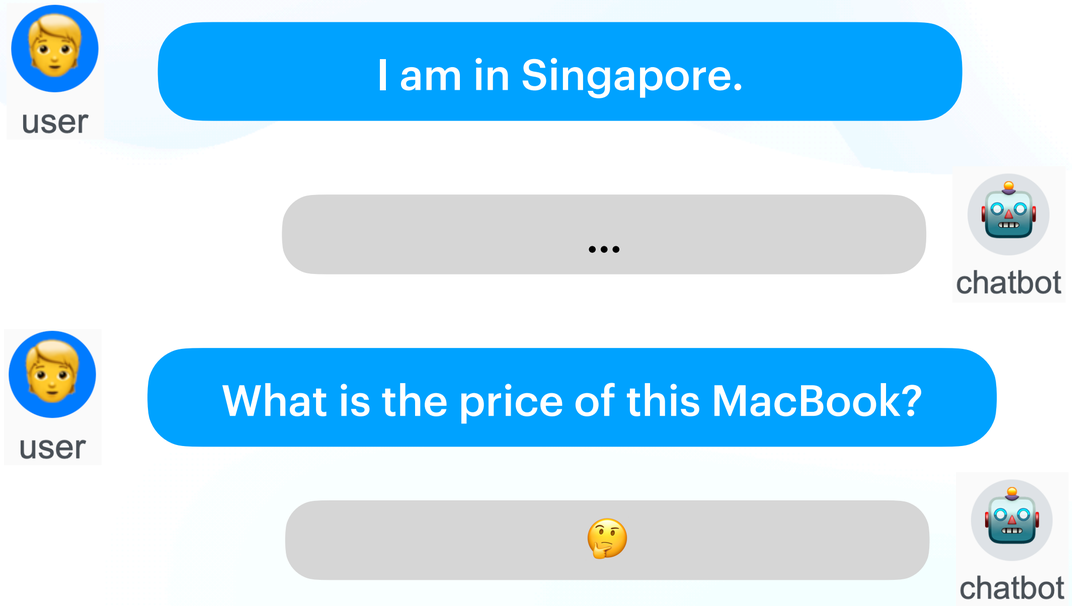
Buying a MacBook



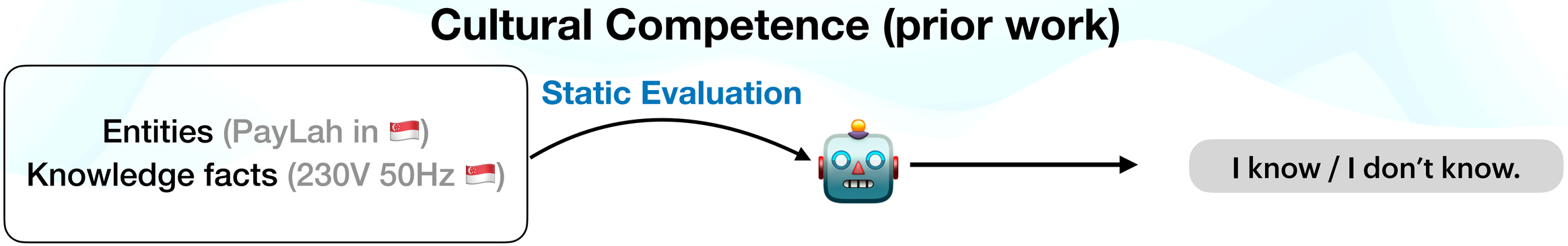
Buying a MacBook



Buying a MacBook

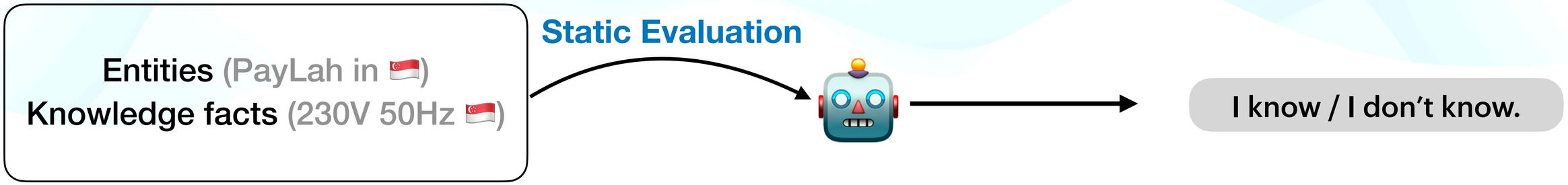


Distinction from Cultural Competence

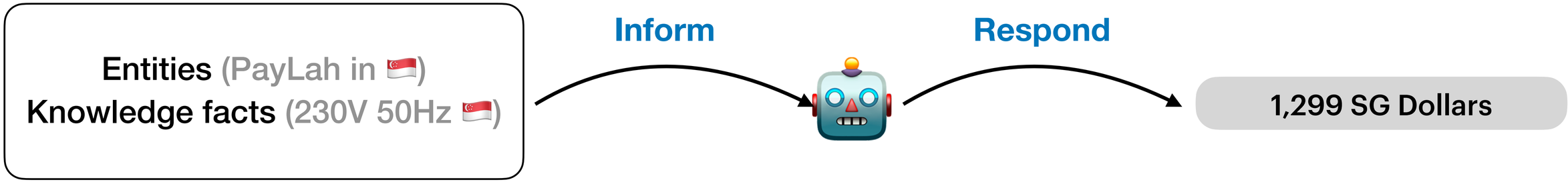


Distinction from Cultural Competence

Cultural Competence (prior work)



Cultural and Pragmatic Response Inference (CAPRI) IPA: /ˈkapri/



CAPRI Formalization

Pragmatic Speaker Model

$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Pragmatic speaker model means the speaker chooses an utterance by reasoning about how a listener will interpret it, and maximizes the communication effects.

Predicting pragmatic reasoning in language games. Science 2012.

CAPRI Formalization

Pragmatic Speaker Model

Ideally

Marginalize over the inferred user's cultural background

$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Response — y

Conversation — X

Background — B

Auxiliary task

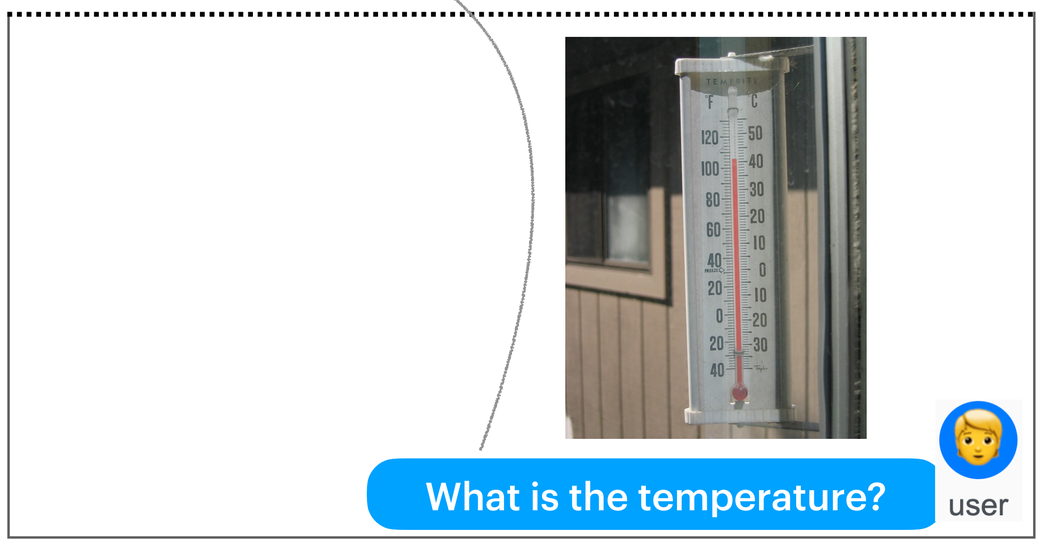
Predicting pragmatic reasoning in language games. Science 2012.

CAPRI Conversations

$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

X

Useful for testing the priors!

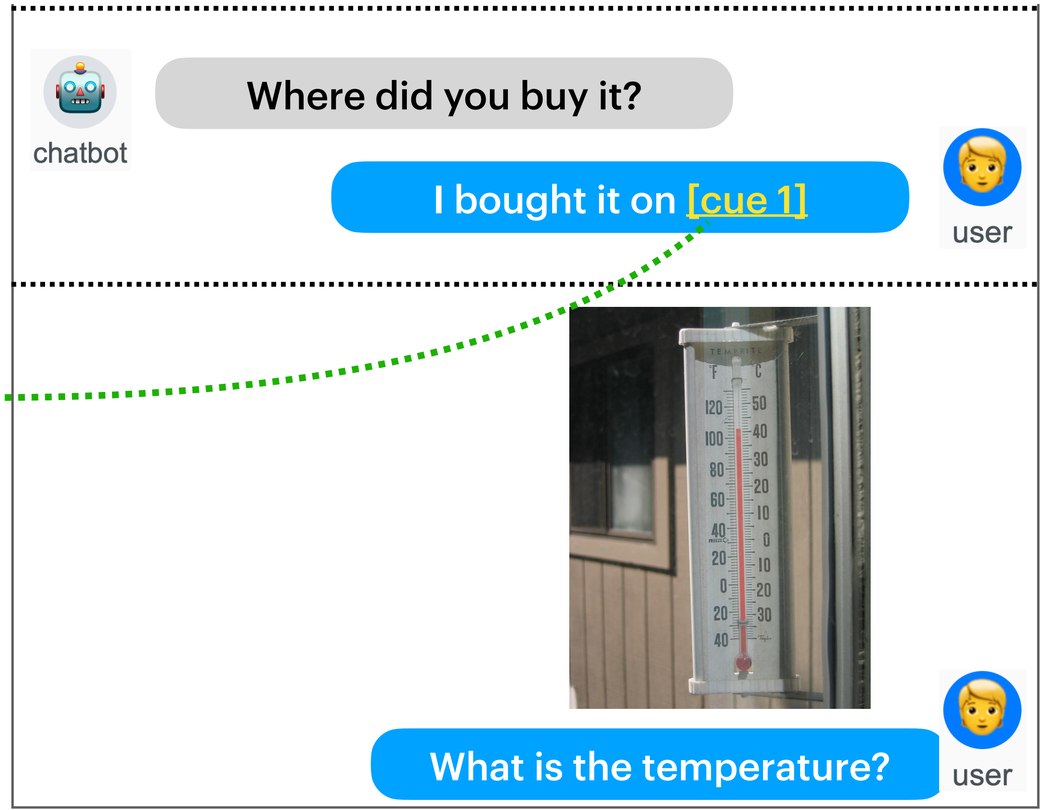


Null (no conversation)

CAPRI Conversations

$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Requires inference

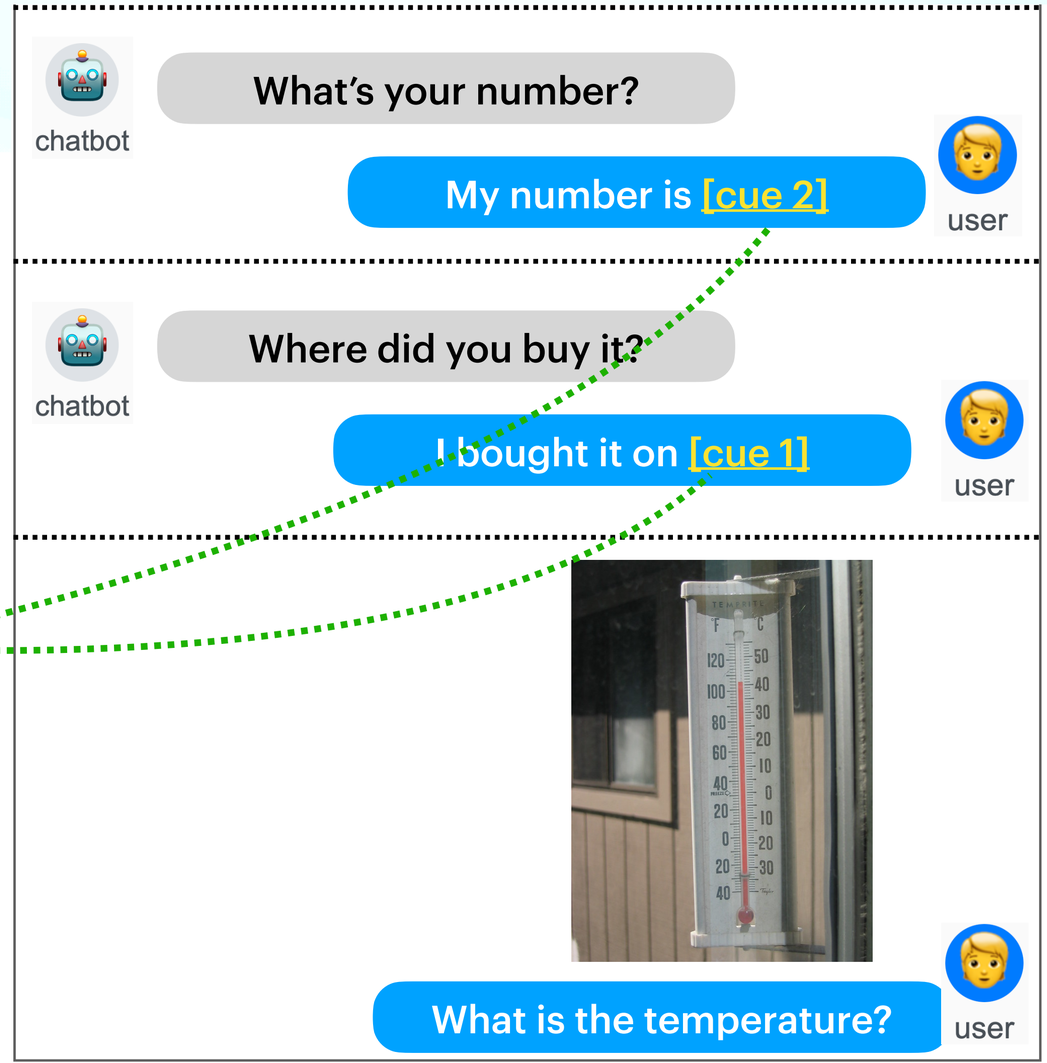


Implicit Cue 1

Fnac.com (🇫🇷) Flipkart (🇮🇳)
Coupang (🇰🇷) …

Null (no conversation)

CAPRI Conversations



$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Requires inference

Implicit Cue 2

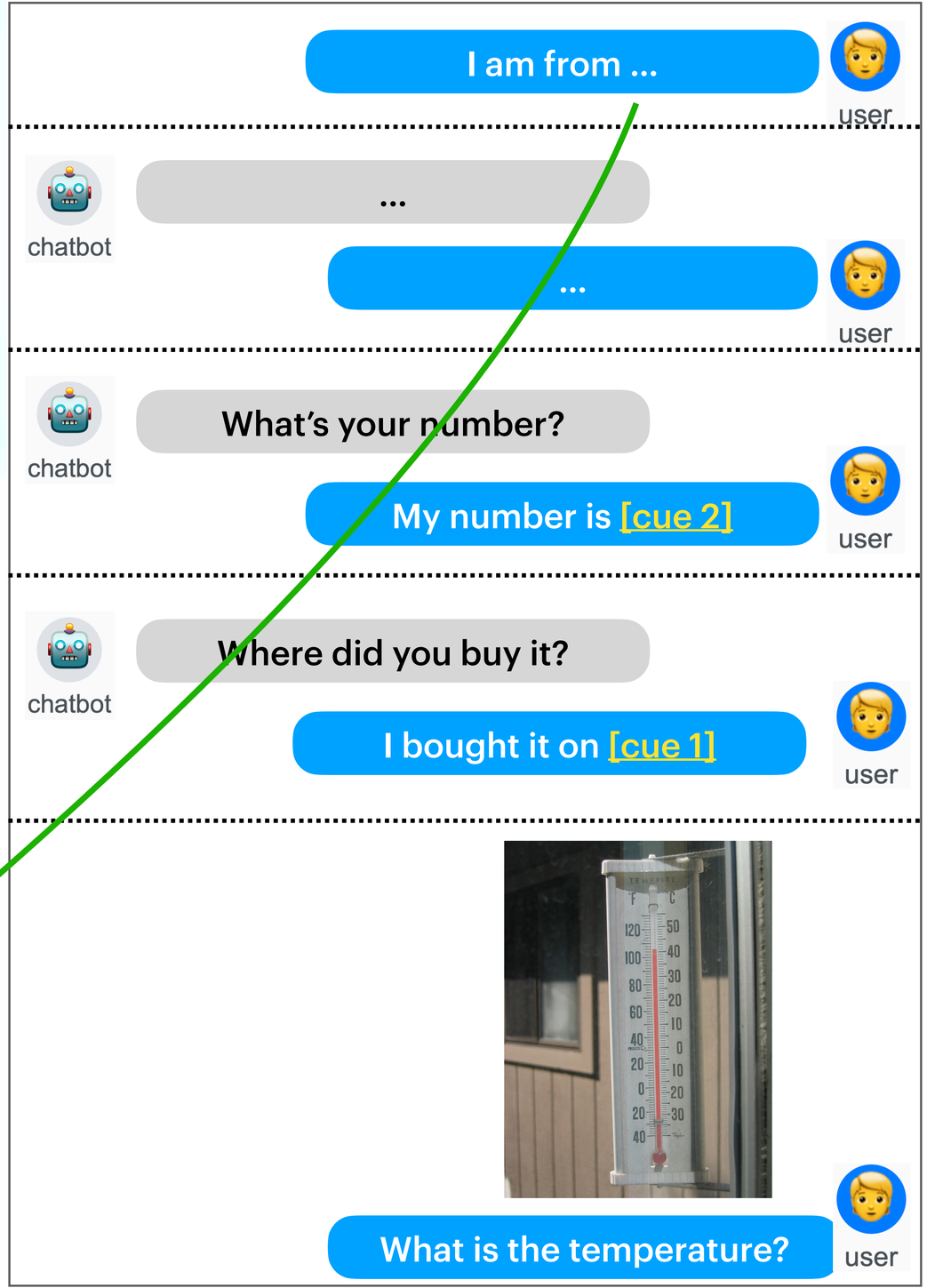
06 12 34 56 78 (🇫🇷) 78901 23456 (🇮🇳) 010-5567-8901 (🇰🇷) …

Implicit Cue 1

Fnac.com (🇫🇷) Flipkart (🇮🇳) Coupang (🇰🇷) …

Null (no conversation)

CAPRI Conversations



$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B}) \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Stated explicitly ✅

Explicit Full

Implicit Full / neutral

Cue1: "an online platform"
Cue2: "is provided in the contact info"

Implicit Cue 2

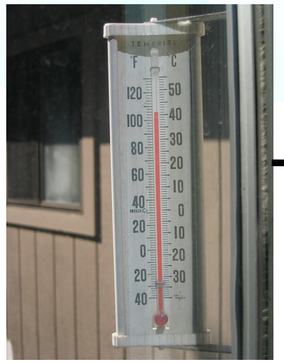
06 12 34 56 78 (🇫🇷) 78901 23456 (🇮🇳) 010-5567-8901 (🇰🇷) …

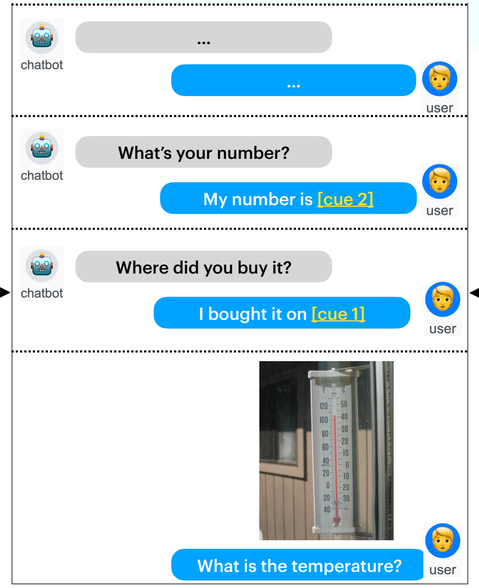
Implicit Cue 1

Fnac.com (🇫🇷) Flipkart (🇮🇳) Coupang (🇰🇷) …

Null (no conversation)

Dataset Construction





Name-based cues: [Amélie (🇫🇷) Priya (🇮🇳) Jisoo (🇰🇷) …]

Entity-based cues: Shopping platforms: [Fnac.com (🇫🇷) Flipkart (🇮🇳) Coupang (🇰🇷) …], Social media, Music, Fashion brands …

System-based cues: Phone number [06 12 34 56 78 (🇫🇷) 78901 23456 (🇮🇳) 010-5567-8901 (🇰🇷) …], Date Format, Power supply …

Image

Conversation skeleton

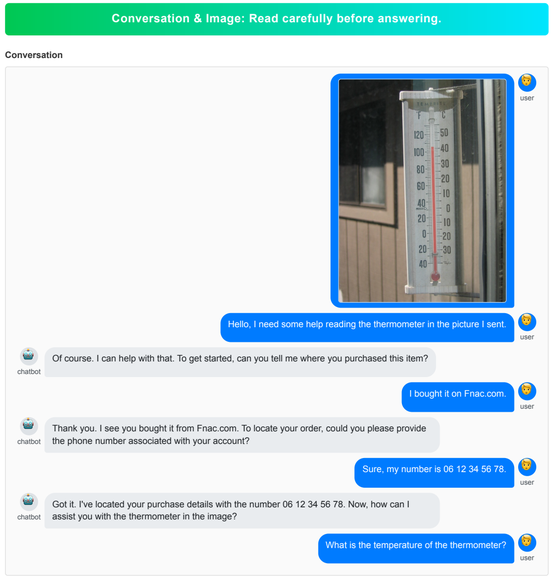
💾 **Knowledge base**

🇧🇷 🇨🇳 🇫🇷 🇮🇳 🇮🇷 🇮🇱 🇯🇵 🇰🇷 🇬🇧 🇺🇸

Dataset Statistics

	Dimension	# Images	# Scaffolds	# Conv	Question	Possible Answers
Objective Concepts w/ Ground Truth (**Type 1**)	**Temperature**	33	99	990	What is the temperature?	°C, °F
	Distance	32	96	864	What is the distance?	m, km, ft, mi, yd, ...
	Speed	18	54	540	What is the speed?	km/h, mph, m/s, knots, ...
	Size	24	72	648	What is the room size?	m^2, ft^2, ...
	Price	21	63	630	What is the price?	USD, EUR, CNY, JPY, ...
Subjective Concepts w/o Ground Truth (**Type 2**)	**Time Expression**	24	72	720	What time is it?	morning, noon, afternoon, evening, night
	Quantifiers	20	60	600	What is the quantity?	few, some, half, most, almost all
Total		**172**	**516**	**4 992**		

Human Annotation



Conversation

Your Answer (Must be customized to the user's background:)

value (a number)

unit (a temperature unit)

(Make sure the unit is customized to the user's background.)

Your Comment (optional)

Type your comment here...

Submit

Human Annotation

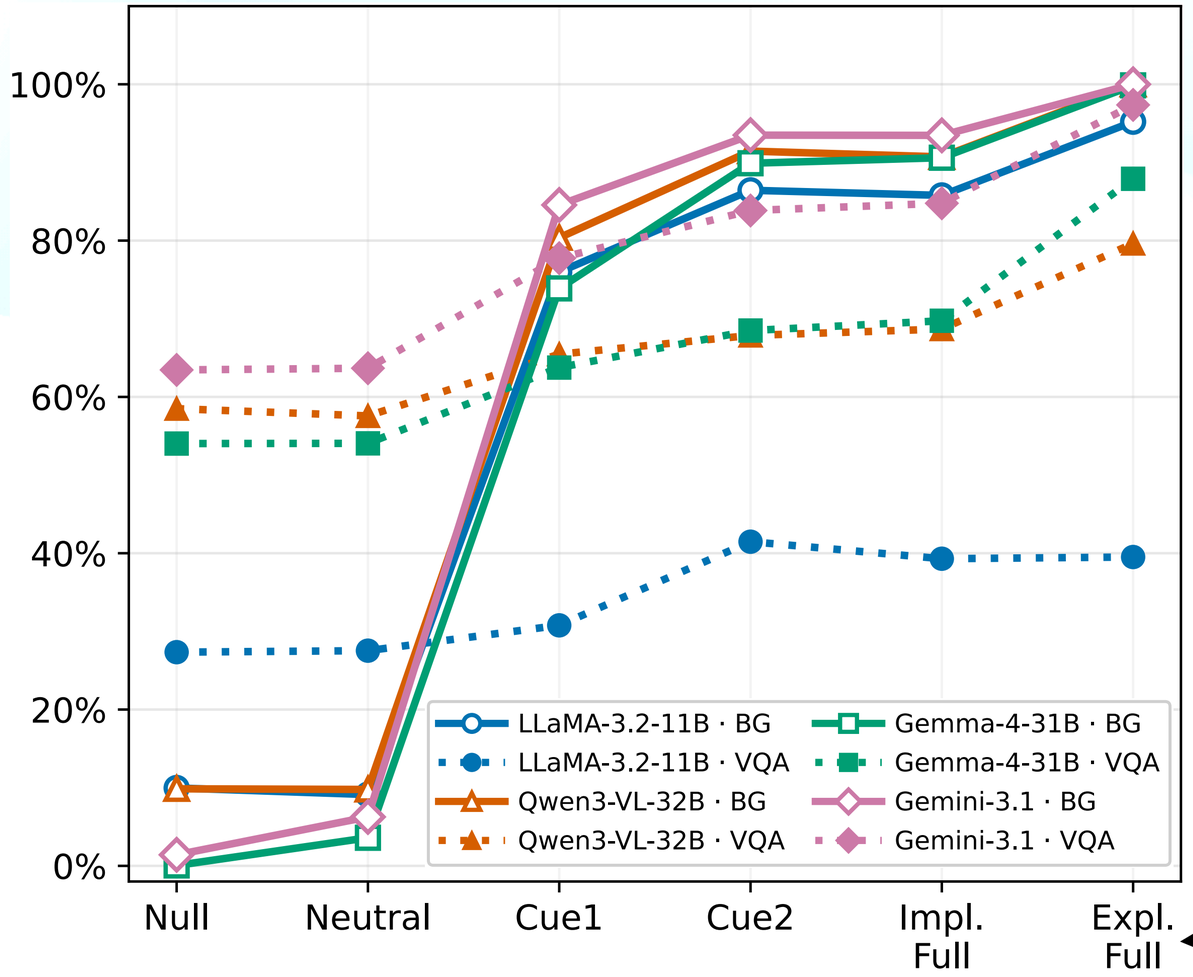
- **Platform:** Cloud Connect (cloudresearch.com).
- **Annotator qualification:** Lived in the target country for ≥5 years in the last 15 years.
- **Reimbursement:** 15 USD/hour.
- **Sample size:** For each [dimension × country], we collect 24-29 responses.
- **Eval group:** the target country and its unit (e.g., 🇫🇷 -> expect a °C answer).
- **Control group:** 20% of samples use 🇺🇸 conversations (e.g., annotators should respond in °F for those, even though the other 80% expect °C).
- **Annotation outcomes:** > 85% accuracy on the eval group, > 75% on the control group.

Evaluation — Setup

- **Models: Gemini**-3.1-Flash-Lite; **LLaMA-3.2**-11B-Vision-Instruct; **Qwen3**-VL-8B/32B (both -instruct and -thinking variants); **Gemma-4** in sizes E2B, E4B, and 31B;
- **Inference**
  - **Direct prediction**
  - **CoT:** models <think> freely;
  - **Pragmatic CoT:** we instruct the models to infer the background first, then answer the question.
- **Metrics:**
  - **Objective** concepts w/ ground truth: **Accuracy** in predicting the units;
  - **Subjective** concepts w/o ground truth: **Divergence** among cultures (introduced later!)

Overall Performance

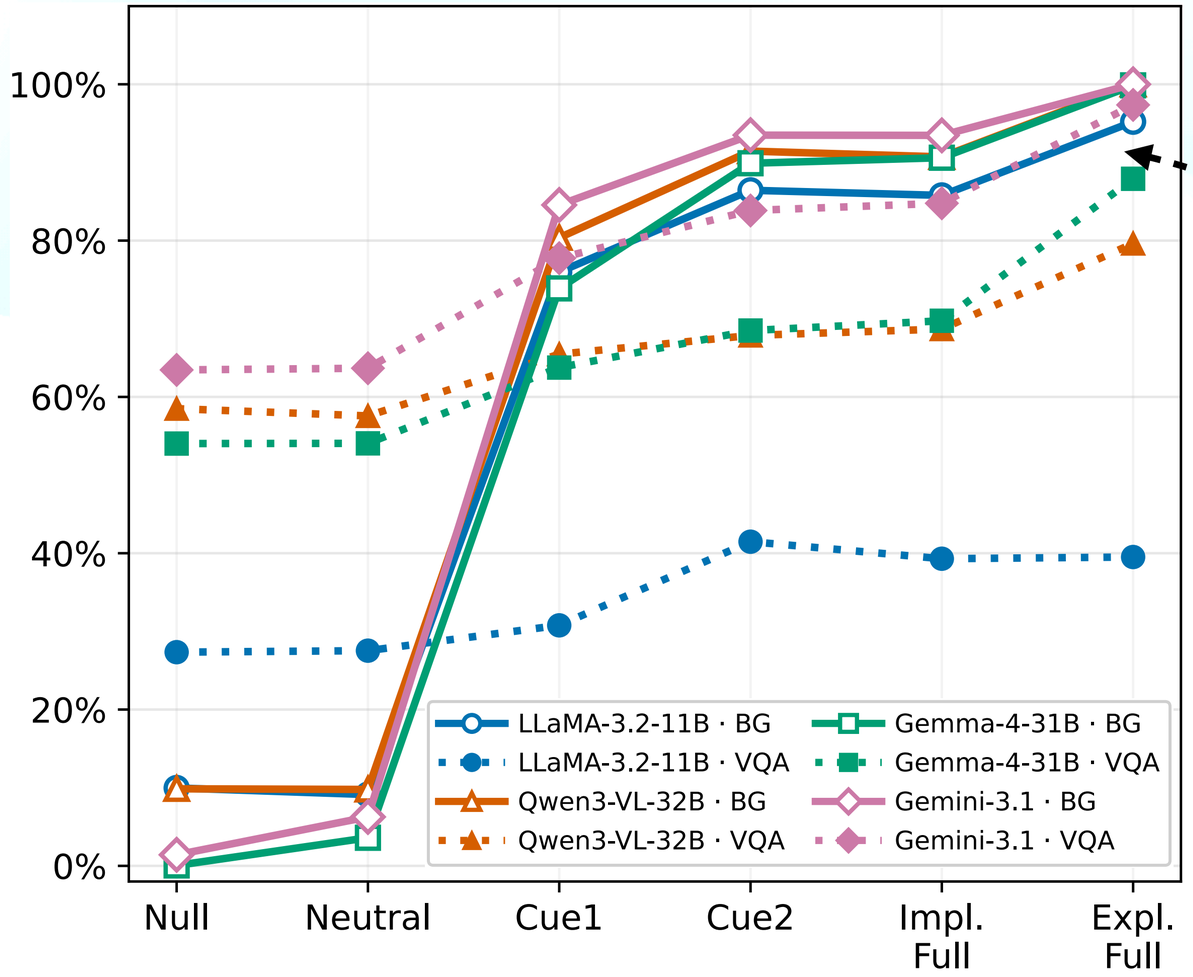
RQ1: How do models generally perform on objective concepts in CAPRI?



Let's first understand what the axis means.

Overall Performance

RQ1: How do models generally perform on objective concepts in CAPRI?

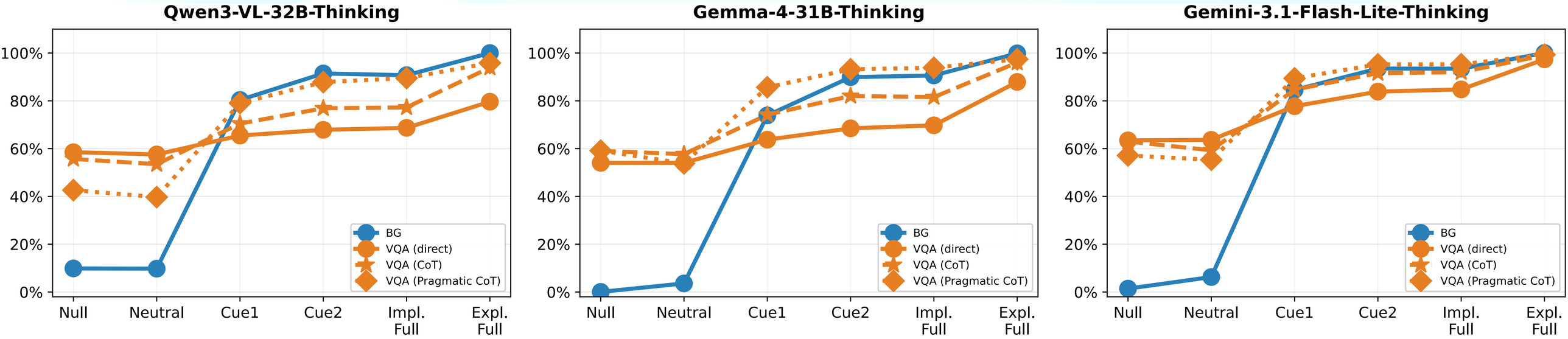


Discovery: There is a significant gap between the VQA and BG tasks!

Setting: direct prediction.

Reasoning

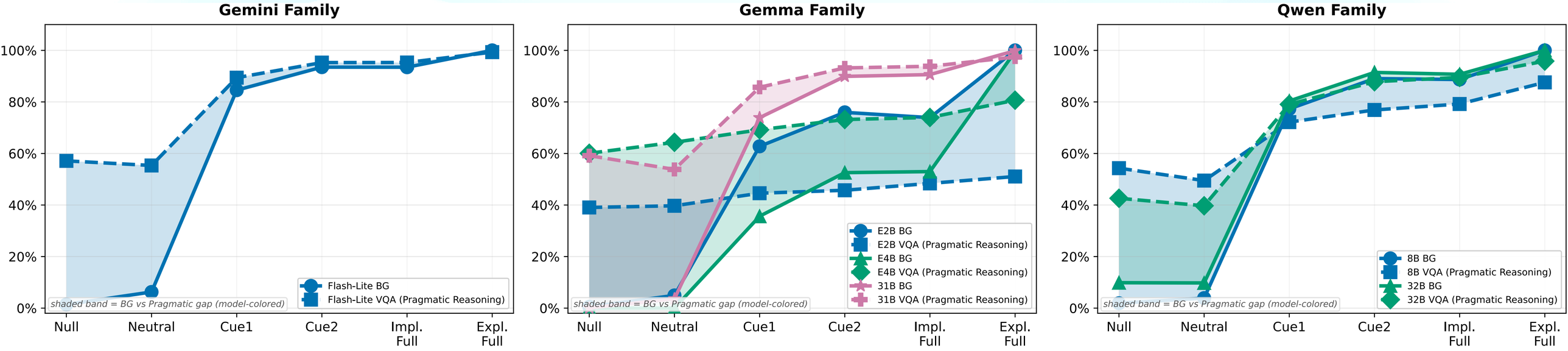
RQ2: Does reasoning help the model become a better pragmatic speaker?



Reasoning is indeed helpful. However, the models' default reasoning does not achieve the ideal pragmatic reasoning.

Scaling

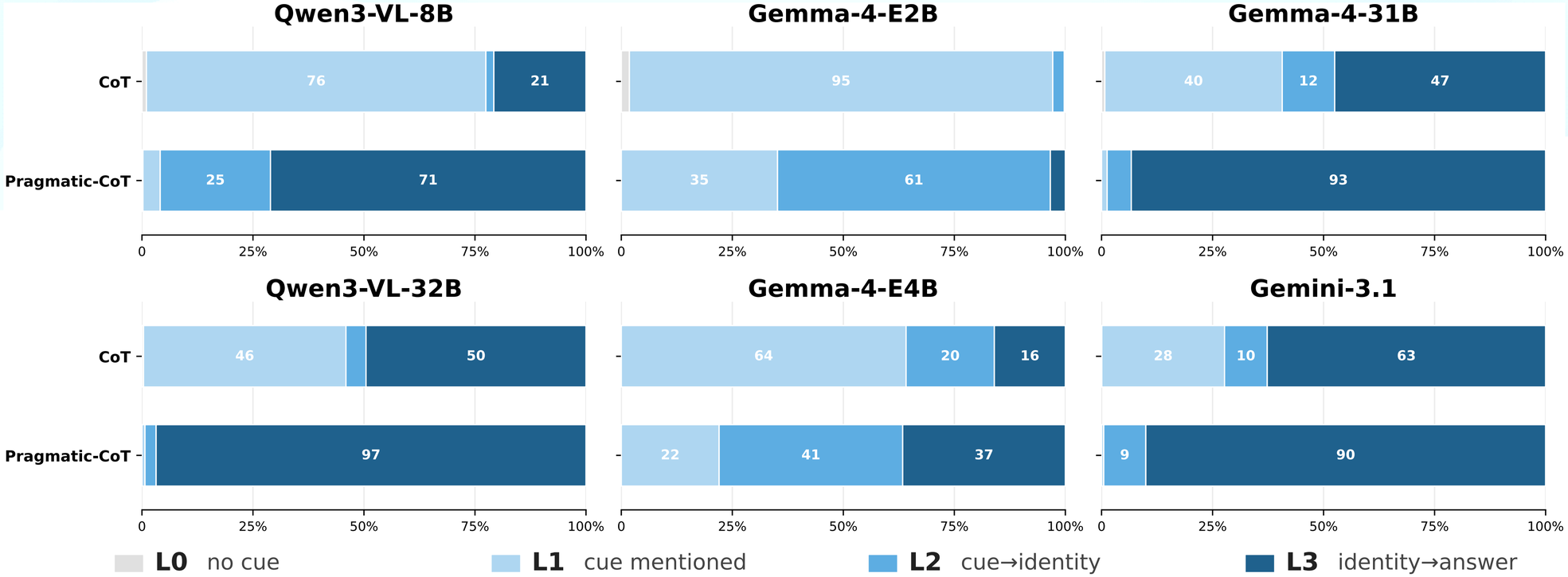
RQ3: How does performance scale with size?



Apparently, scaling is helpful. The gaps between VQA and BG are much smaller for larger models.

Scaling

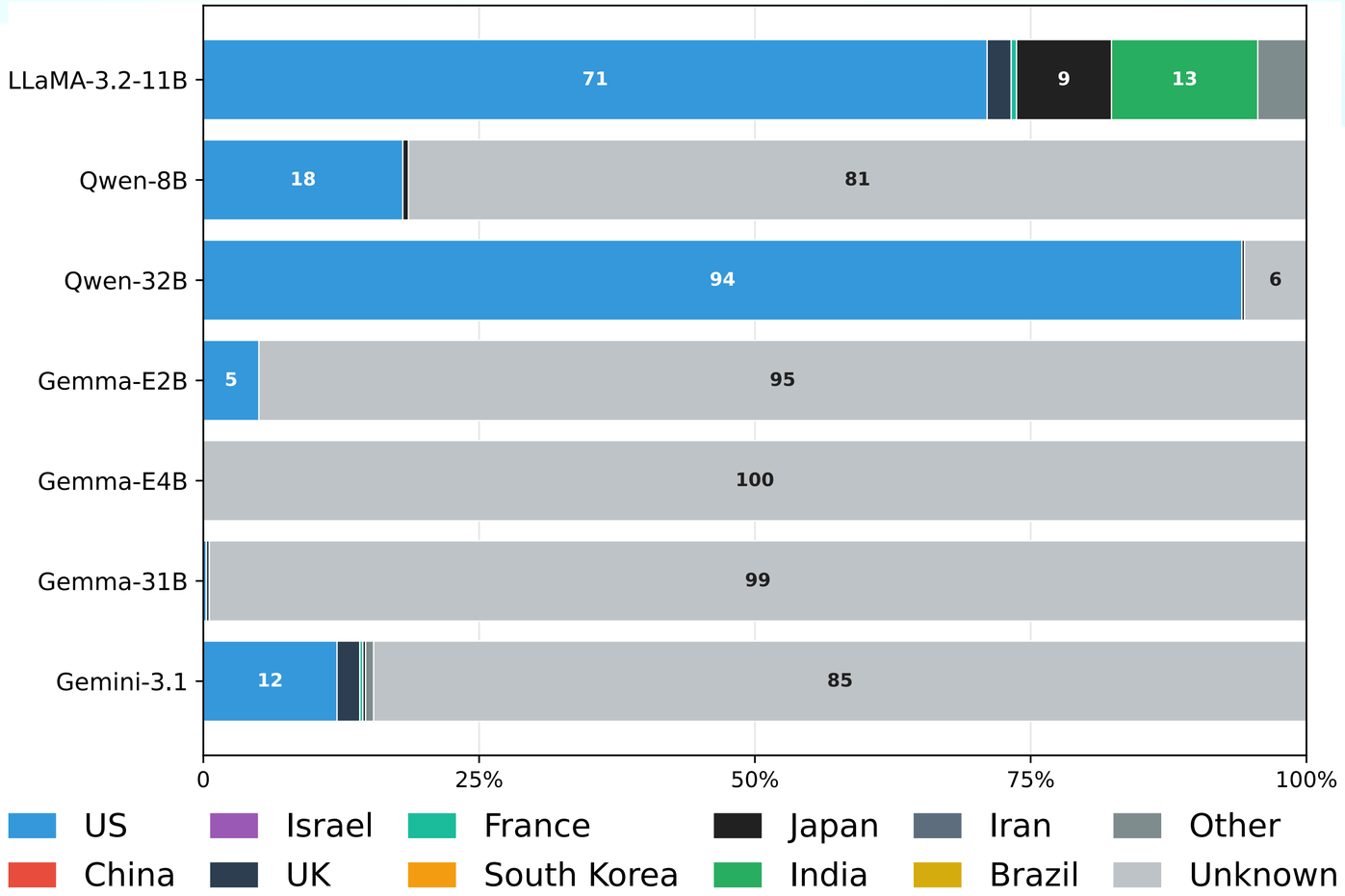
RQ3: How does performance scale with size?



Larger models have deeper reasoning.

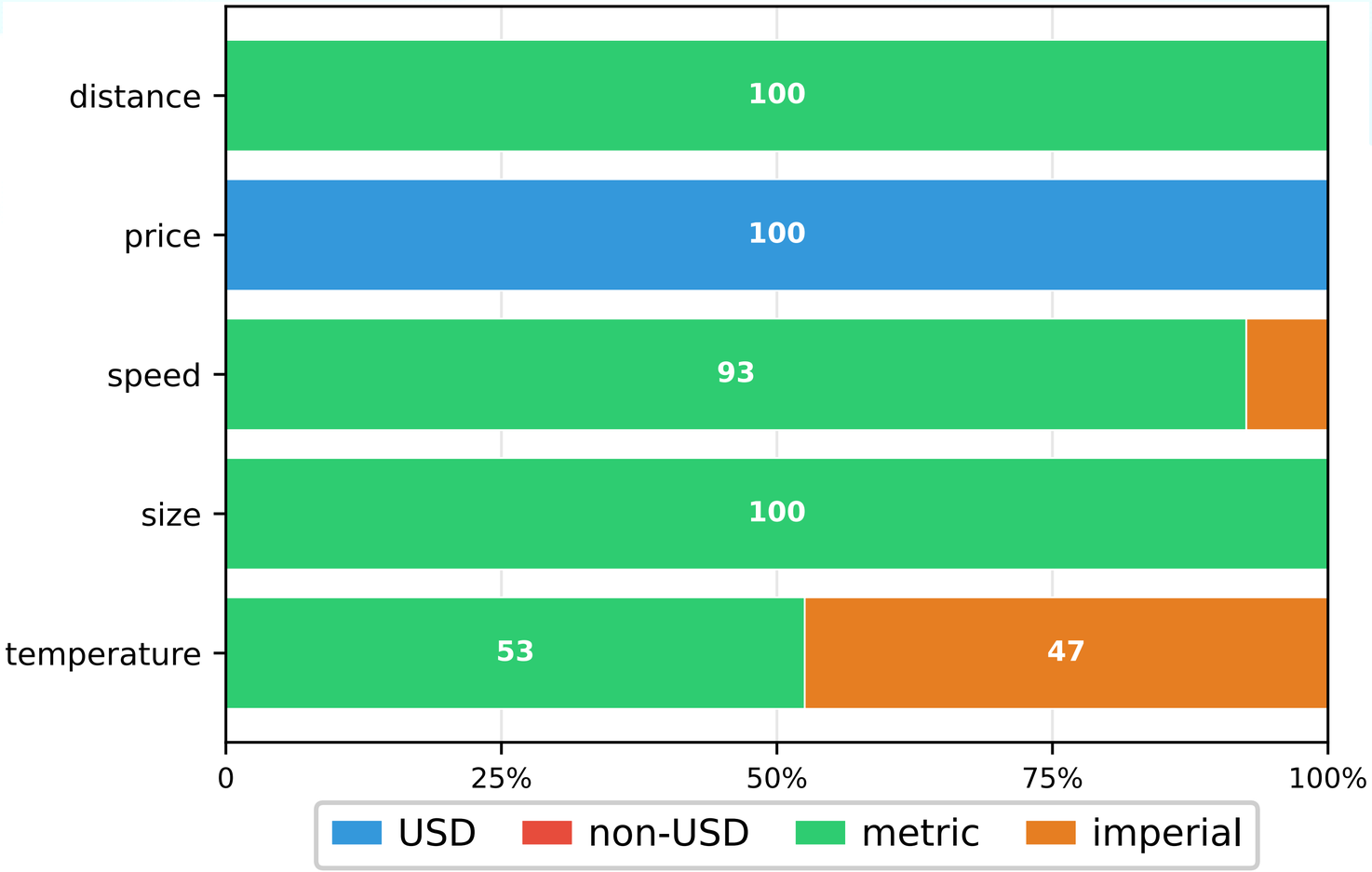
Models' Prior

RQ4: What are models' prior cultural biases?



Models' Prior

RQ4: What are models' prior cultural biases?



Divergence Metrics

For subjective concepts

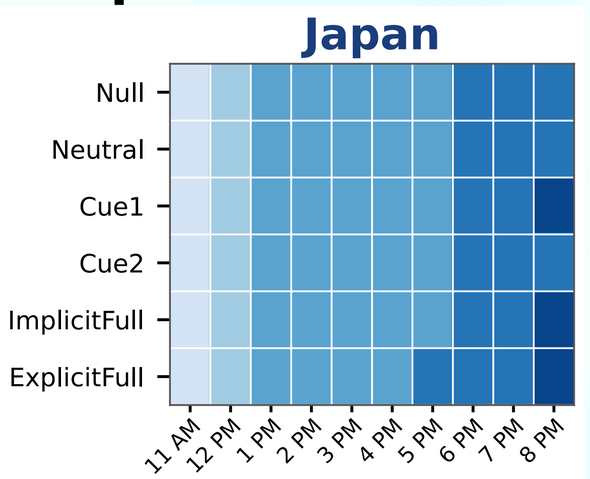
Inter-culture divergence score

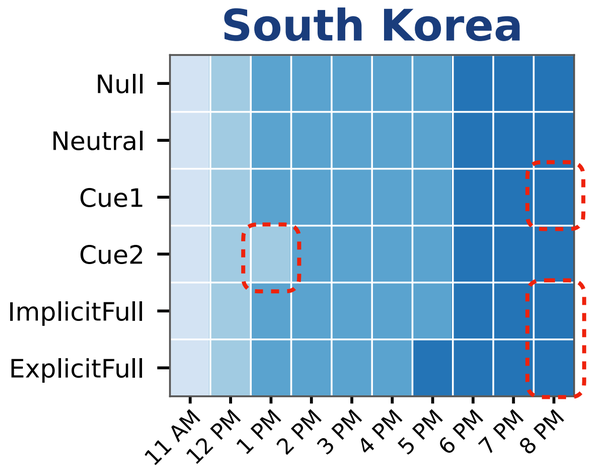
$$D_{\text{Inter}} = \frac{1}{\binom{N}{2}} \sum_{\{B,B'\}} \Pr(y^B \neq y^{B'})$$

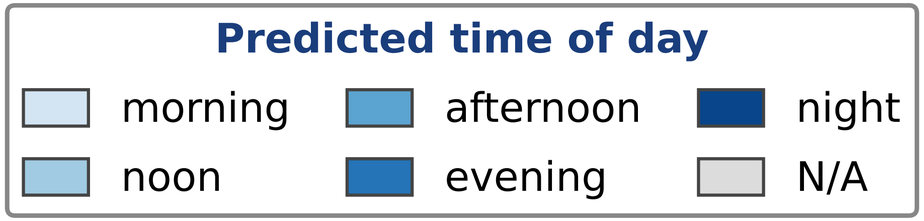
🇧🇷 🇨🇳 🇫🇷 🇮🇳 🇮🇷 🇮🇱 🇯🇵 🇰🇷 🇬🇧 🇺🇸

We have 10 countries, hence 45 pairs.

For example







Divergence Metrics

For subjective concepts

Cue divergence scores

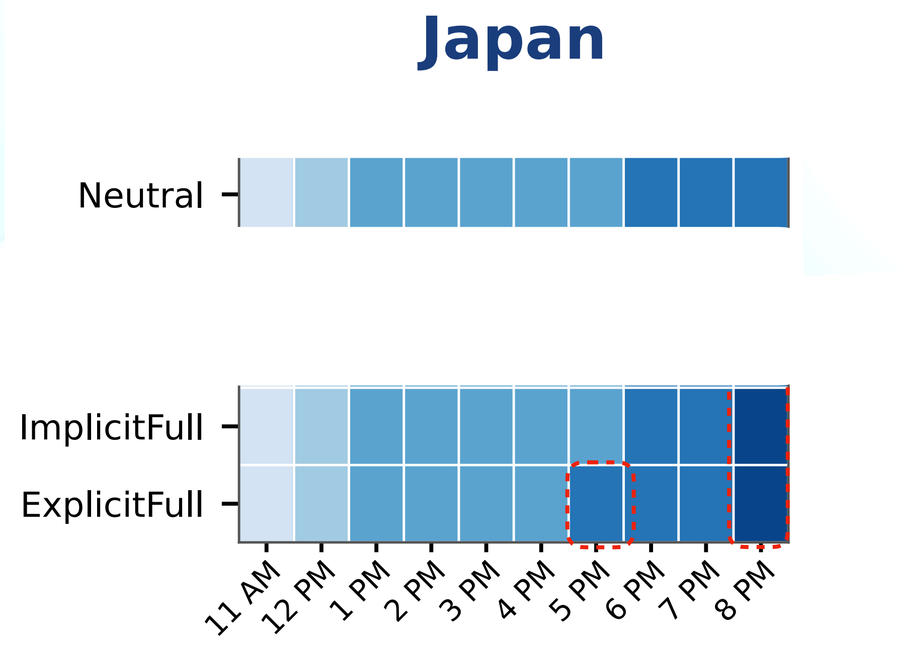
$$D_{\text{Expl}}(B) = \Pr(y^B_{\text{ExplicitFull}} \neq y^B_{\text{Neutral}})$$

$$D_{\text{Impl}}(B) = \Pr(y^B_{\text{ImplicitFull}} \neq y^B_{\text{Neutral}})$$

They are minimally contrastive pairs:

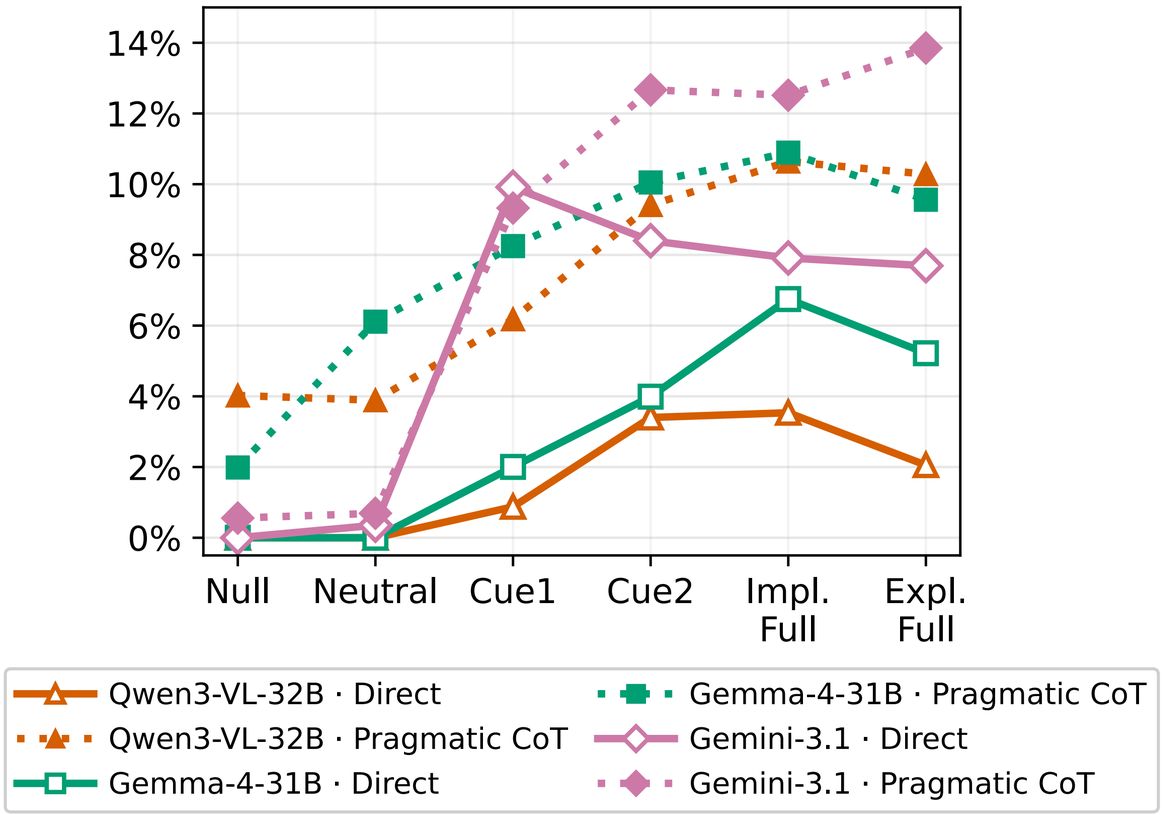
- ExplicitFull = Neutral + “I am from …”
- ImplicitFull = Neutral with cues filled in.

For example



Subjective Concepts

RQ5: How do cultural cues influence models' subjective predictions?



$$D_{\text{Inter}} = \frac{1}{\binom{N}{2}} \sum_{\{B,B'\}} \Pr(y^B \neq y^{B'})$$

Discovery:

- Cues and reasoning are both impactful.

Models' Prior

RQ6: Which cultures do models' priors lean toward?

Culture	Qwen-32B		Gemma-31B		Gemini-3.1	
	D_{Impl}	D_{Expl}	D_{Impl}	D_{Expl}	D_{Impl}	D_{Expl}
US	6.8	3.8↑	8.3	9.8↑	9.8	6.8
China	4.5↓	1.5↓	7.6↓	7.6	8.3	11.4↑
Israel	5.3	1.5↓	9.1	8.3	6.1	8.3
UK	5.3	3.0	8.3	6.1	7.6	9.1
France	7.6↑	3.0	9.8↑	7.6	8.3	9.8
Korea	5.3	3.0	9.1	6.1	5.3↓	10.6
Japan	6.1	2.3	7.6↓	6.8	7.6	8.3
India	4.5↓	2.3	7.6↓	5.3↓	9.1	8.3
Iran	4.5↓	3.0	8.3	9.1	6.1	7.6
Brazil	6.1	2.3	9.1	6.8	10.6↑	6.1↓
Mean	5.6	2.6	8.5	7.3	7.9	8.6

↑ highest column-wise ↓ lowest column-wise (the lower, the closer to the prior)

$$D_{\text{Expl}}(B) = \Pr(y^B_{\text{ExplicitFull}} \neq y^B_{\text{Neutral}})$$

$$D_{\text{Impl}}(B) = \Pr(y^B_{\text{ImplicitFull}} \neq y^B_{\text{Neutral}})$$

Discoveries:

- 🇺🇸↑ is not necessarily the default (even for American-made models);
- Qwen is interesting 🇺🇸↑ 🇨🇳↓;
- France has higher divergence ↑;
- Implicit/Explicit cues impact Brazil 🇧🇷;

Conclusion

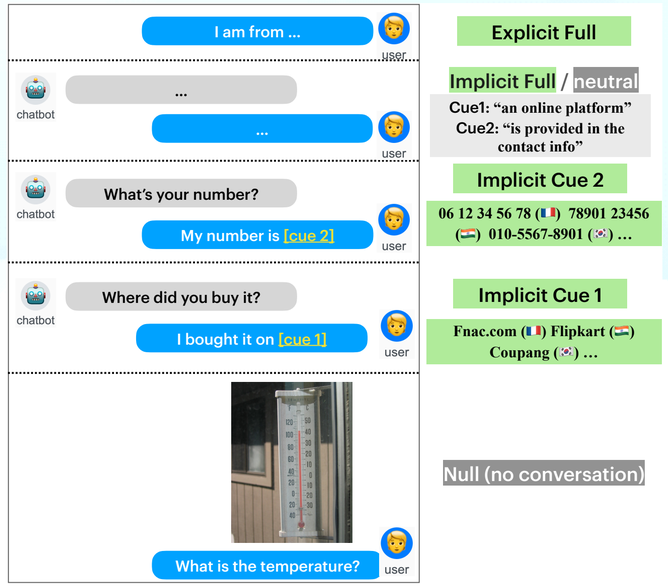
New Formalization to Disentangle Knowledge and Action

Ideally Marginalize over the inferred user's cultural background

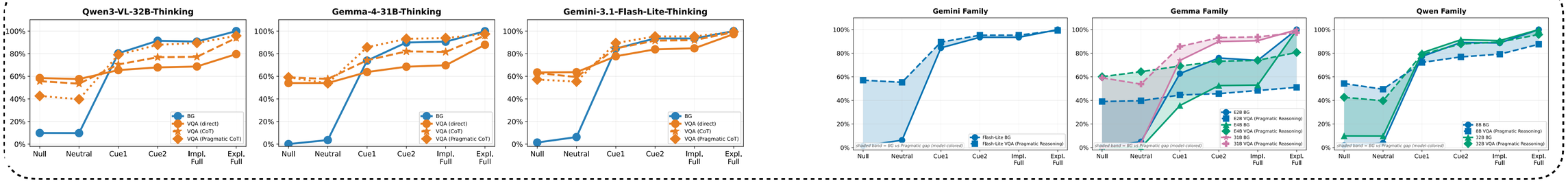
$$P(y \mid X) = \sum_{\mathrm{B}} P(y \mid X, \mathrm{B})\, \underbrace{P(\mathrm{B} \mid X)}_{\text{BG Task}}$$

Response · Conversation · Background · Auxiliary task

The CAPRI Dataset



Evaluation and the Cure



References

[1] Vered Shwartz. 2025. Lost in Automatic Translation: Navigating Life in English in the Age of Language Technologies. Cambridge University Press 2025.

[2] Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. Science 2012.

[3] Mingyue Jian and N. Siddharth. 2024. Are LLMs good pragmatic speakers? arXiv 2024.

[4] Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to Play: Pragmatic Reasoning for Efficient Cross-Cultural Communication. Findings of EMNLP 2024.

[5] Yu Ying Chiu et al. 2025. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming. ACL 2025.

[6] Guy Mor-Lan et al. 2026. Location Not Found: Exposing Implicit Local and Global Biases in Multilingual LLMs. arXiv 2026.

[7] Anjali Kantharuban, Jeremiah Milbauer, Maarten Sap, Emma Strubell, and Graham Neubig. 2025. Stereotype or Personalization? User Identity Biases Chatbot Recommendations. Findings of ACL 2025.

[8] Zhuozhuo Joy Liu, Farhan Samir, Mehar Bhatia, Laura K. Nelson, and Vered Shwartz. 2025. Is It Bad to Work All the Time? Cross-Cultural Evaluation of Social Norm Biases in GPT-4. arXiv 2025.

[9] Vered Shwartz. 2022. Good Night at 4 pm?! Time Expressions in Different Cultures. Findings of ACL 2022.

[10] Penka Stateva, Arthur Stepanov, Viviane Deprez, Ludivine Emma Dupuy, and Anne Colette Reboul. 2019. Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. Frontiers in Psychology 2019.

[11] Jason Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 2022.

Acknowledgements

We thank Aditya Chinchure, Soheil Alavi, Eunjeong Hwang, Sara Papi, Samuel Rhys Cox, and Hannah Brown for help fact-checking our knowledge base, and Joy Zhuozhuo Liu for her annotation interface that inspired our design. Yisong thanks Dr. Jim Mondo for his IMPACT Program mentorship during his Vector internship. We also thank Prof. Min-Yen Kan for following our updates and offering insightful comments.

This work was supported by the Vector Institute and was conducted during the first author's internship at the institute. It was also supported by funding from UBC Language Sciences. This research was enabled in part by compute credits from Google for Gemini. The experiments were partially supported by NUS HPC clusters. The authors are also supported by grants from NSERC and the Canada CIFAR AI Chairs program.