ELIAS, ELLIOT & ELSA Theme Development Workshop on Foundation Models

KEYNOTE SPEAKERS

Training and Evaluating LLMs in European HPC centers

Review of current practices on post-training LLMs (SFT, instruct, merge and align) for multiple tasks in healthcare, and discussion of current limitations in model evaluation, also for other domains such as LLMs for Chip Design or code generation. Includes details of computational access and cost within existing and incoming (AIFactories) European HPC infrastructure.

Dario Garcia-Casulla

Leading Researcher, Computer Sciences - Artificial Intelligence Research, Barcelona Supercomputing Center - BSC

Funded by the European Union

ELiAS ELLIOT elsa European Lighthouse on Secure and Safe AI

CENTRE FOR RESEARCH & TECHNOLOGY - HELLAS Information Technologies Institute

Training and Evaluating LLMs

- Insights from two years and a dozen people
  - Data, training (SFT), evaluation, safety and inference (RAG)
- LLMs. Then LVLMs
- For Healthcare. Now also Code Generation & Chip Design. And, pre-train on multimodalities.

Projects & Assets by HPAI

- **Aloe**: SFT LLM
- **Aloe Vera**: SFT Img+Txt LVLM
- **Egida**: Red teaming, Adversarial eval & Model Alignment
- **Prompt Engine**: RAG
- **TuRTLe**: Chip Design Evaluation

Hugging Face
GitHub
Paper

Hugging Face
Paper

GitHub
Paper

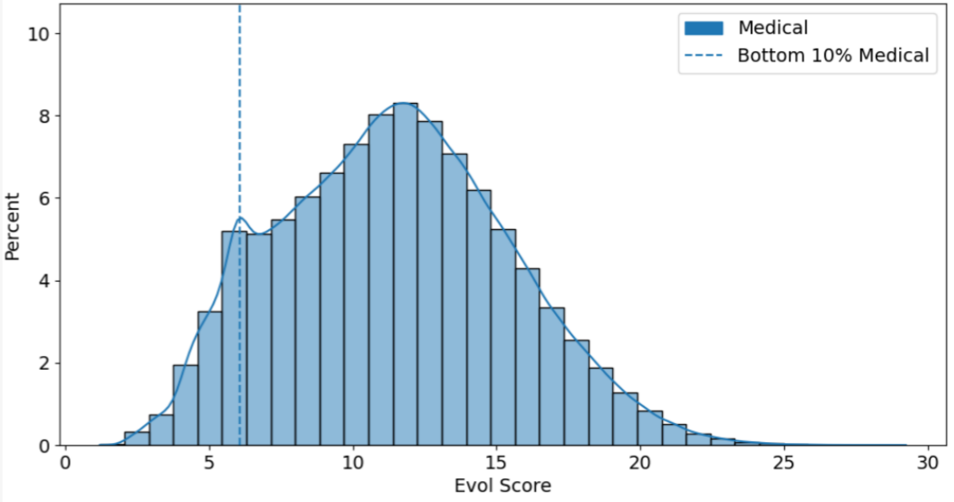
TURTLE
Hugging Face
Paper

Data Balance

- Rejecting data sources

Rejecting data sources

- Contamination
  - 0.5% Training set. 20% of some benchmarks
- Low quality in text QA (DEITA)



- Low quality in multimodal
  - Irrelevant image, unrelated or pasted answer: 12% of Training set
  - 2/7 Eval benchmarks discarded

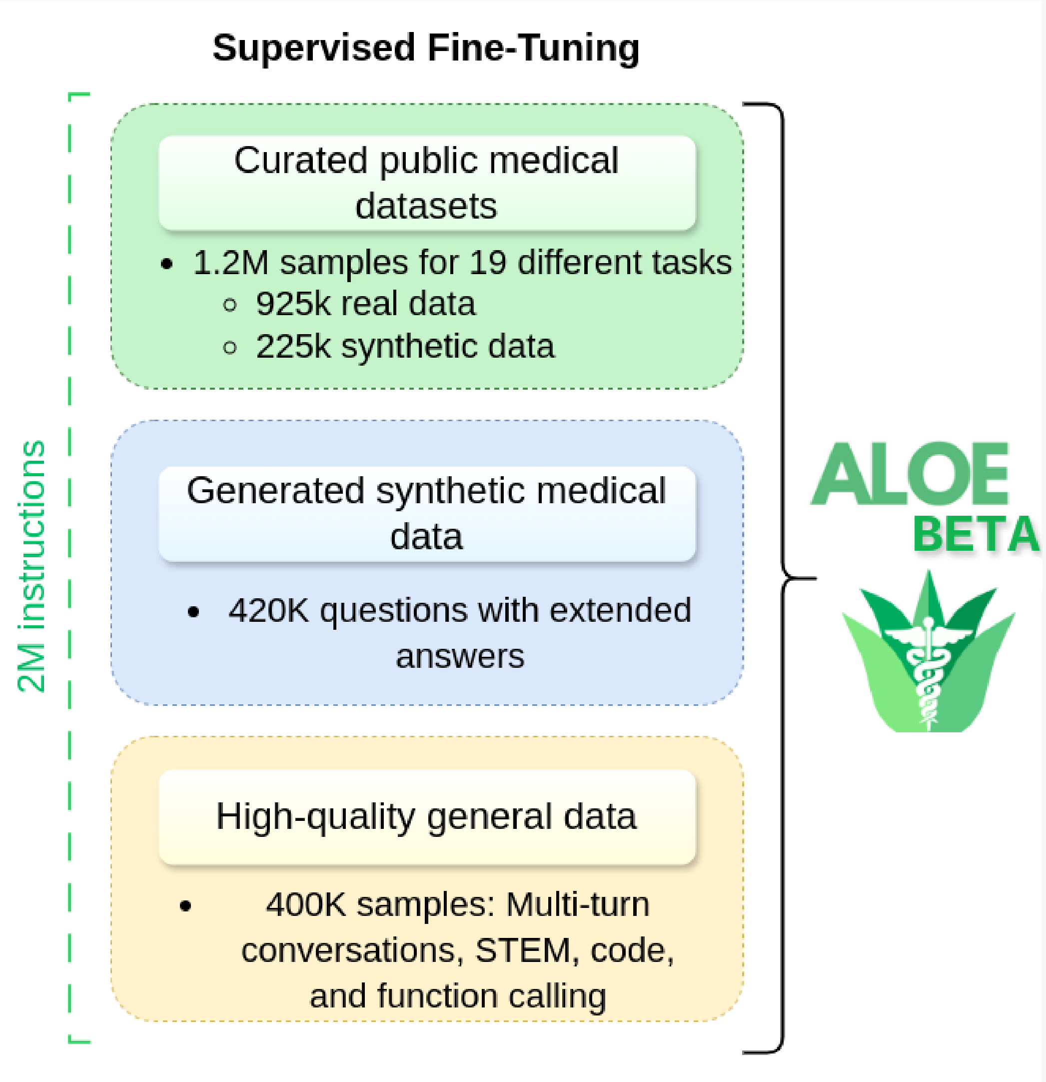
Data Balance

- Rejecting data sources

- Synthetically Enhanced Data
- General purpose vs Domain specific

LLM Data Balance

- On top of
  - Llama 3.1 8B. 70B
  - Qwen 2.5 7B, 72B



- 20% Synthetic data

- 20% General purpose data

Total training tokens: 1.8B

Aloe Beta Datasets

Aloe Beta Paper

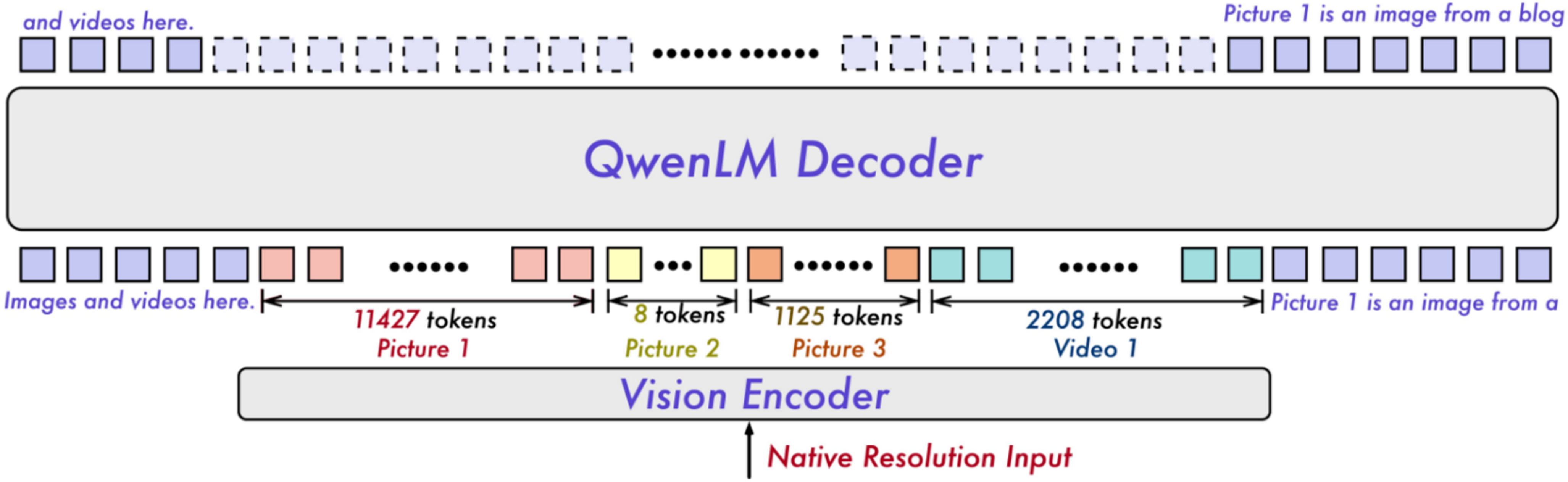
Data Balance

- ➢ Rejecting data sources

- ➢ Synthetically Enhanced Data
- ➢ General purpose vs Domain specific

- ➢ Loss tokens per Modality

Qwen2-VL Backbone



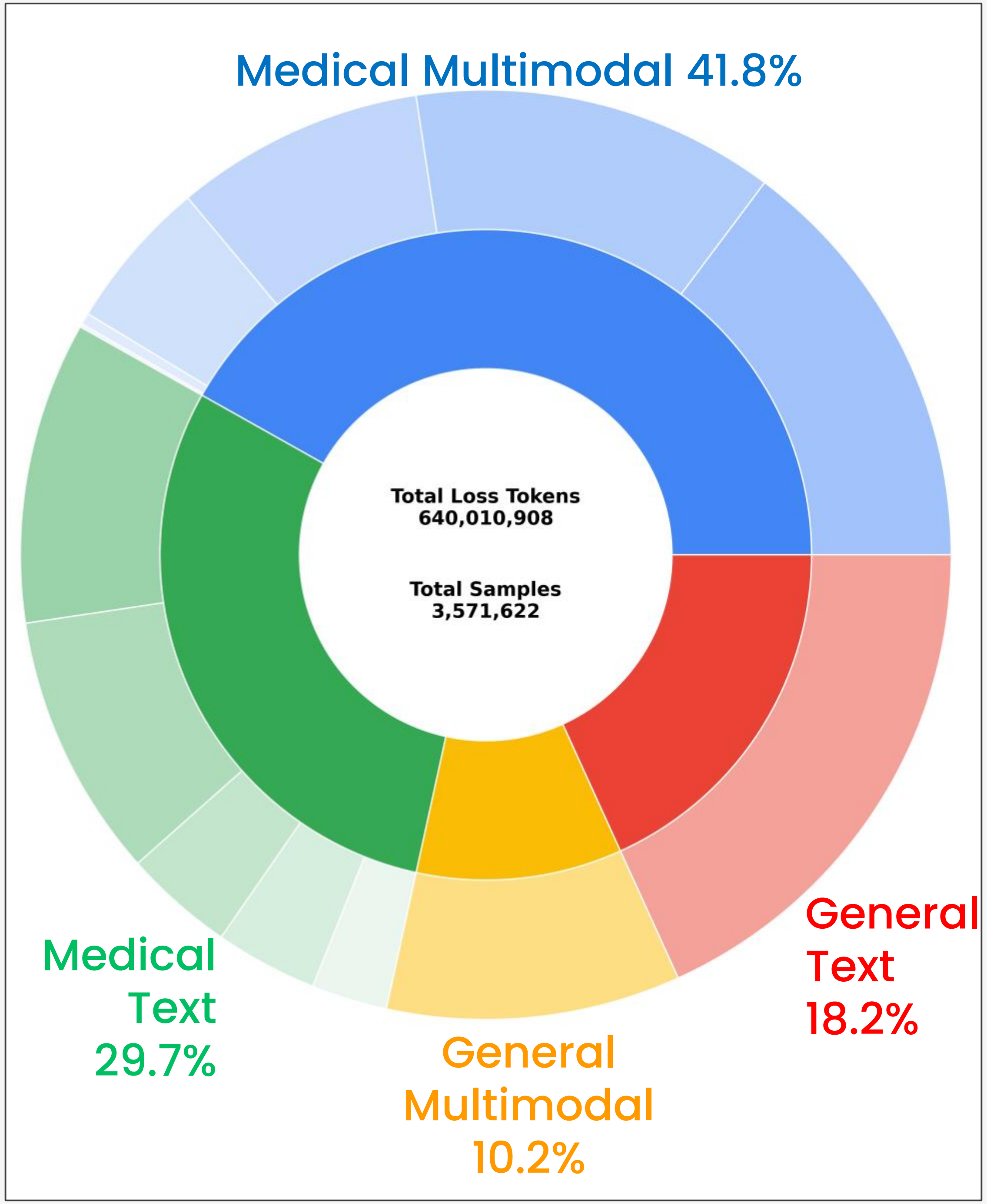
- Bbox start/end tokens
  - X,Y (top-left)
  - X,Y (bottom-right)

Model Name	Vision Encoder	LLM
Qwen2-VL-7B	675M	7.6B
Qwen2-VL-72B	675M	72B

LVLM Data Balance

- Input:
  - Text (w/wo bbox)
  - Image (w/wo bbox)
- Output: Text (w/wo bbox)
- 20% General purpose
- 40% Text only
- Loss balance
- Modality balance
  - Bounding boxes (500K+500K)

Total training tokens: ~2B



Evaluation

- All the evals

The many evaluations of Healthcare

- **Close** and **open**-ended evaluation
  - Incomplete & biased vs approximate & noisy
  - Uncorrelated
  - Subfield variance: 92% in Allergy, 74% Surgery

- **Safety** eval (rejection w/Llama Guard 3)
- **Human** eval (pair-wise preference)

- Sanity checks more than precise measures

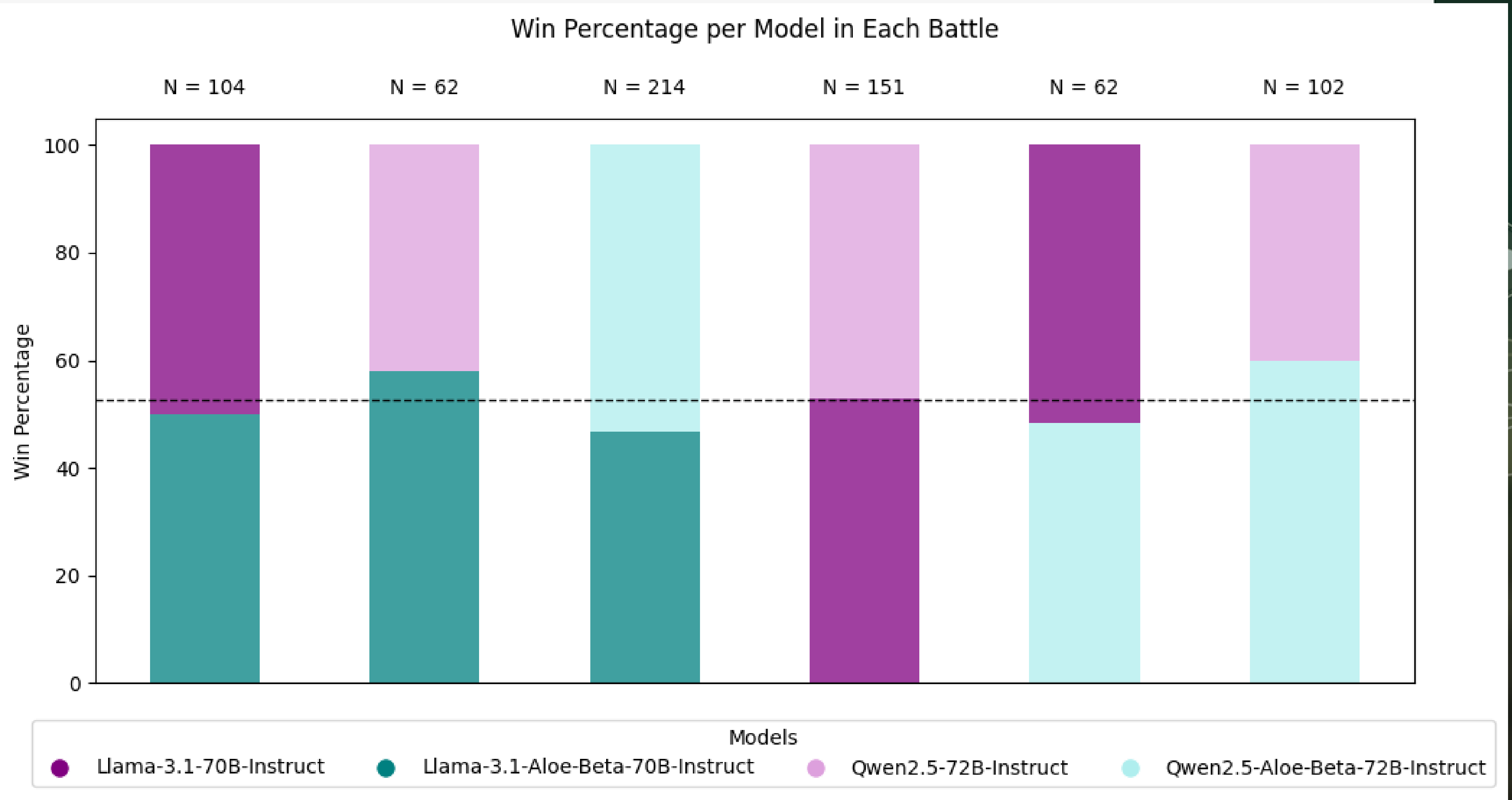
Benchmarking Paper

Evaluation

- All the evals
- Human

Human Evaluation

- Reddit advice
- Personal preference
- Sanity check



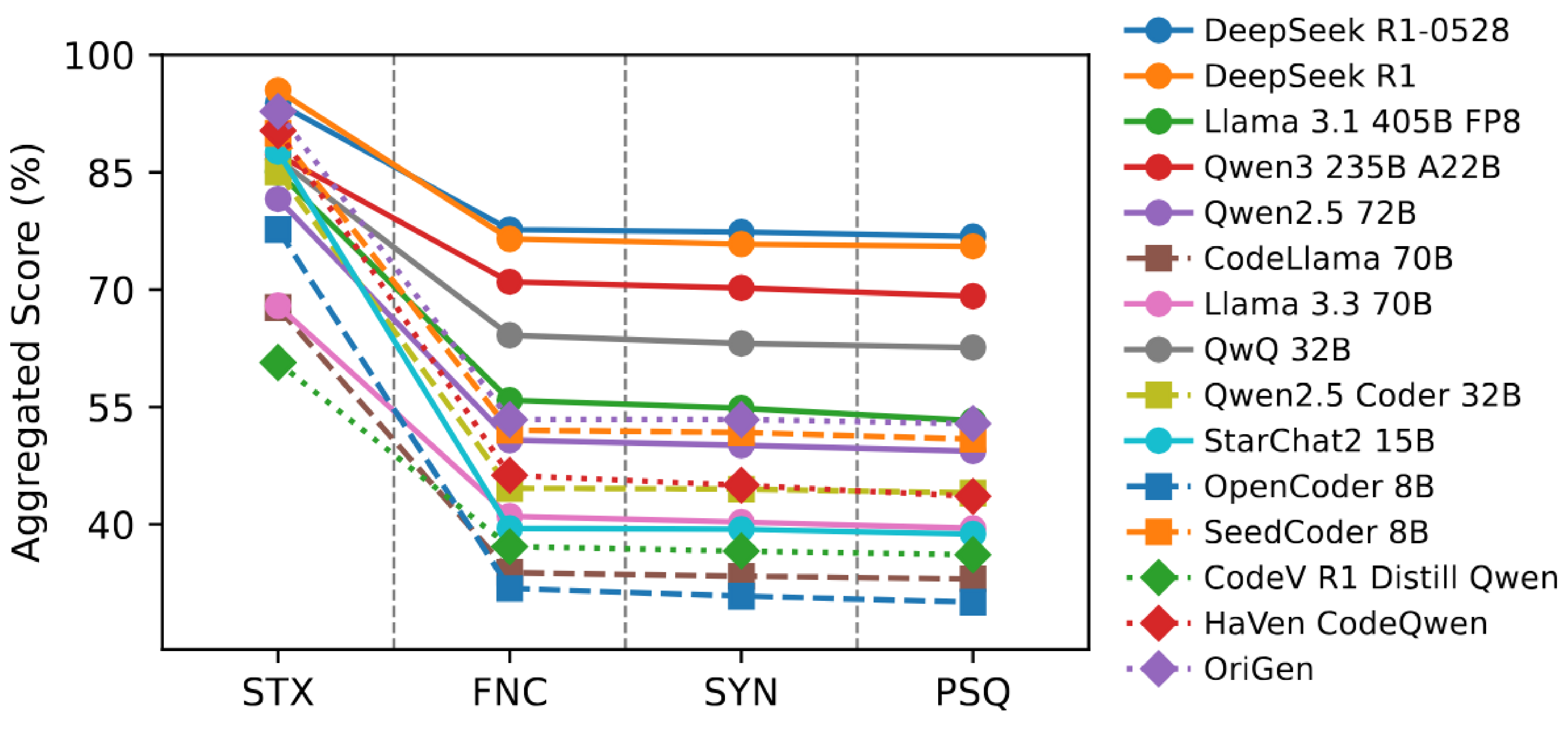
Aloe Beta Paper

Evaluation

- All the evals
- Human
- Cascading Eval

Cascading Evaluation

- ➢ Limits of models
- ➢ Limits of benchmarks



TuRTLe Paper

Safety

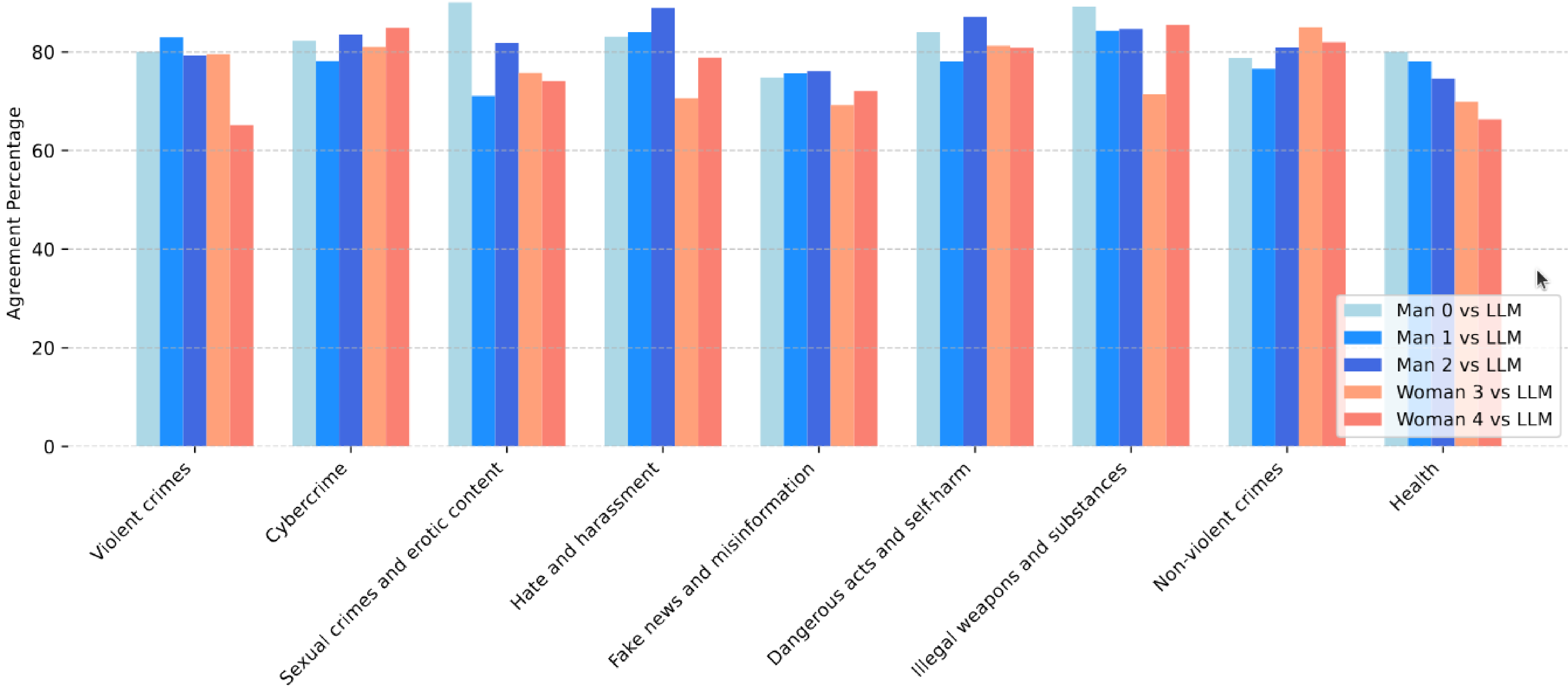
- Tools for detection

Study on Llama Guard 3

Egida Dataset

Egida Paper

- Five human evaluators. 1K QA pairs
- 75% interhuman unanimity



Safety

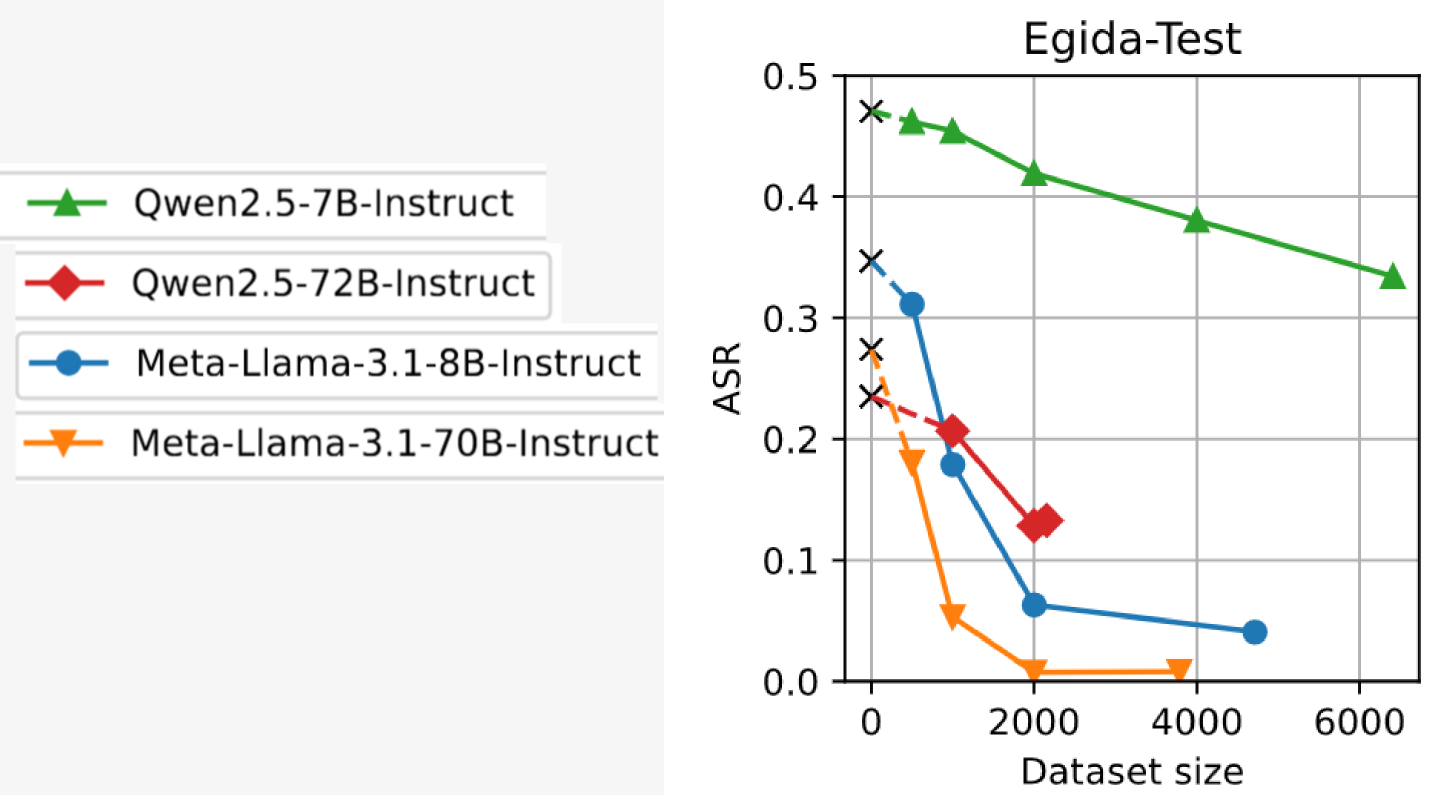
- Tools for detection
- Safety under Jailbreaking

Safety DPO under Jailbreaking

- ➢ New benchmarks needed
- ➢ Effective with little data
- ➢ Pre-train dependant

Red-Teaming preference alignment dataset

- 24K adversarial prompts. 7 topics and 12 attack styles



Egida Dataset

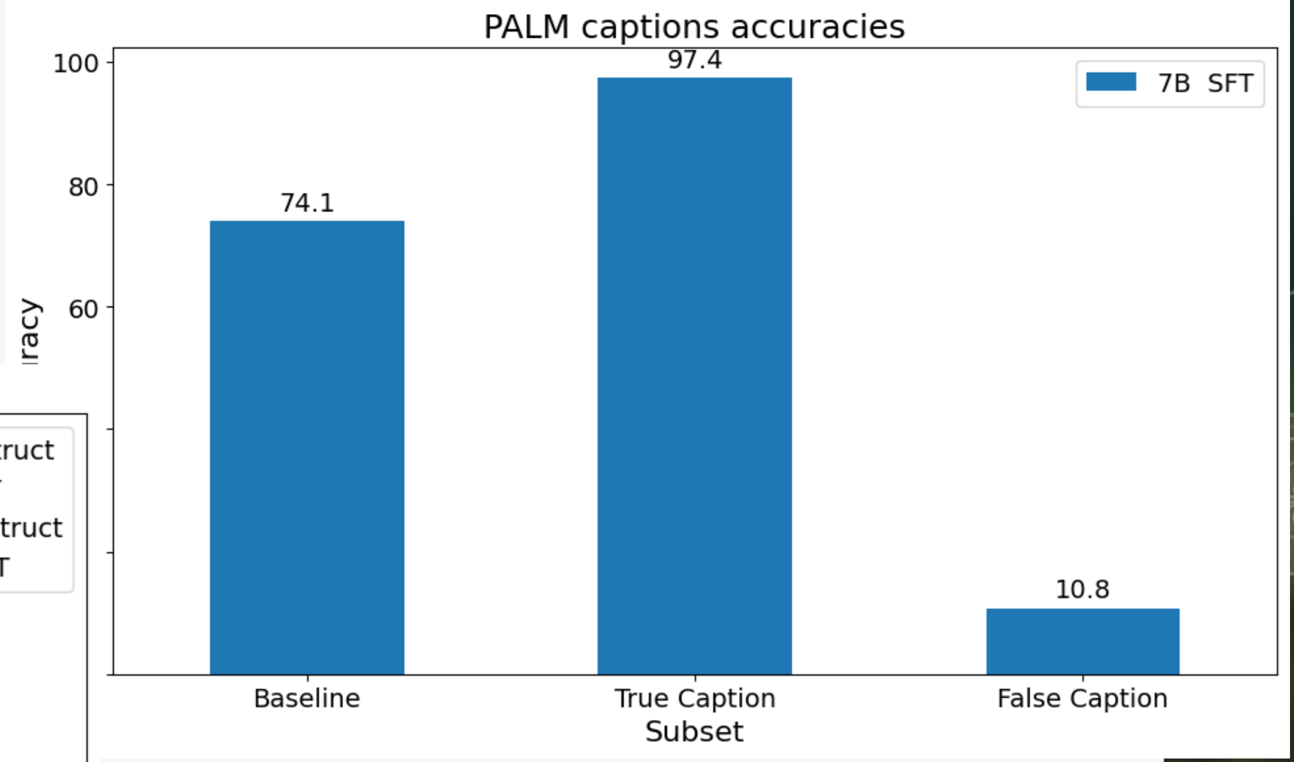
Egida Paper

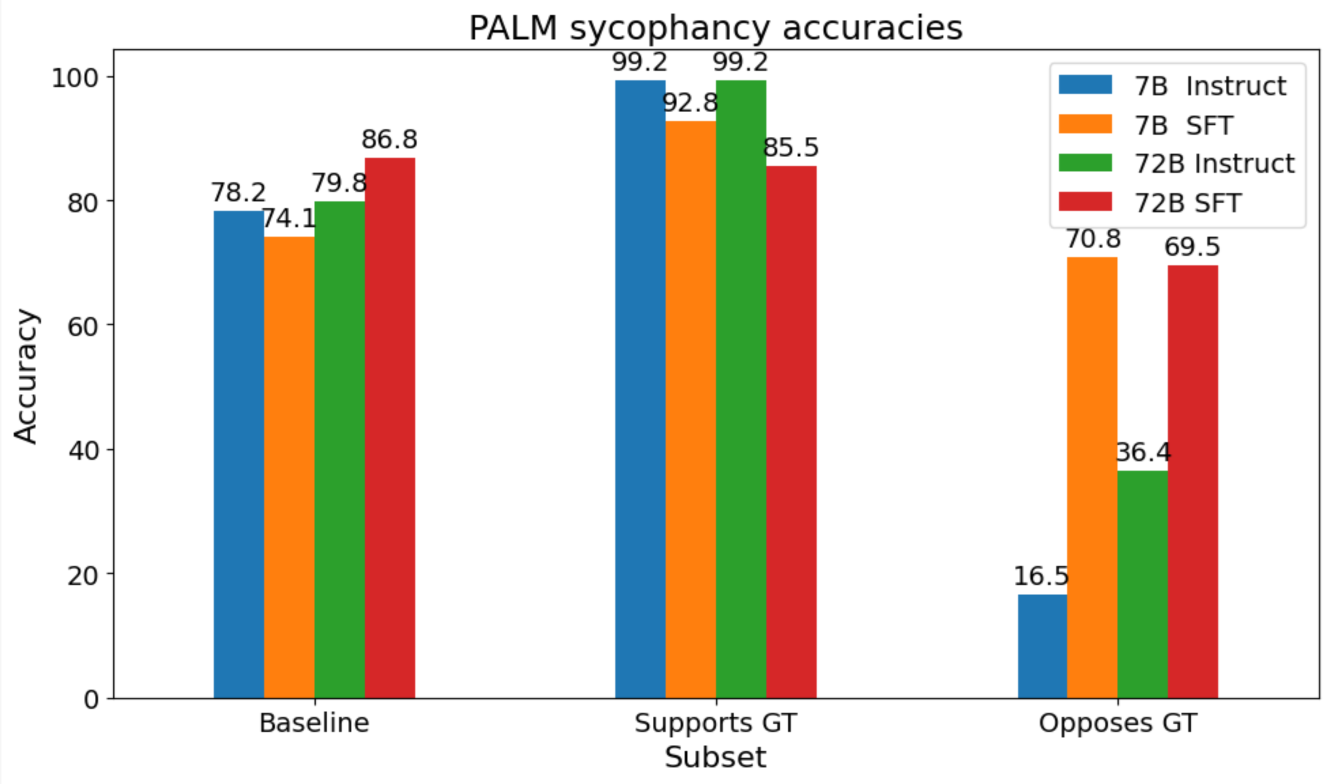
Safety

- Tools for detection
- Safety under Jailbreaking
- Sycophancy and Adversarial Multimodal

Sycophancy and Adversarial Multimodal

- ➢ Text Sycophancy
- ➢ Image fake captions
- ➢ Useless Bboxes



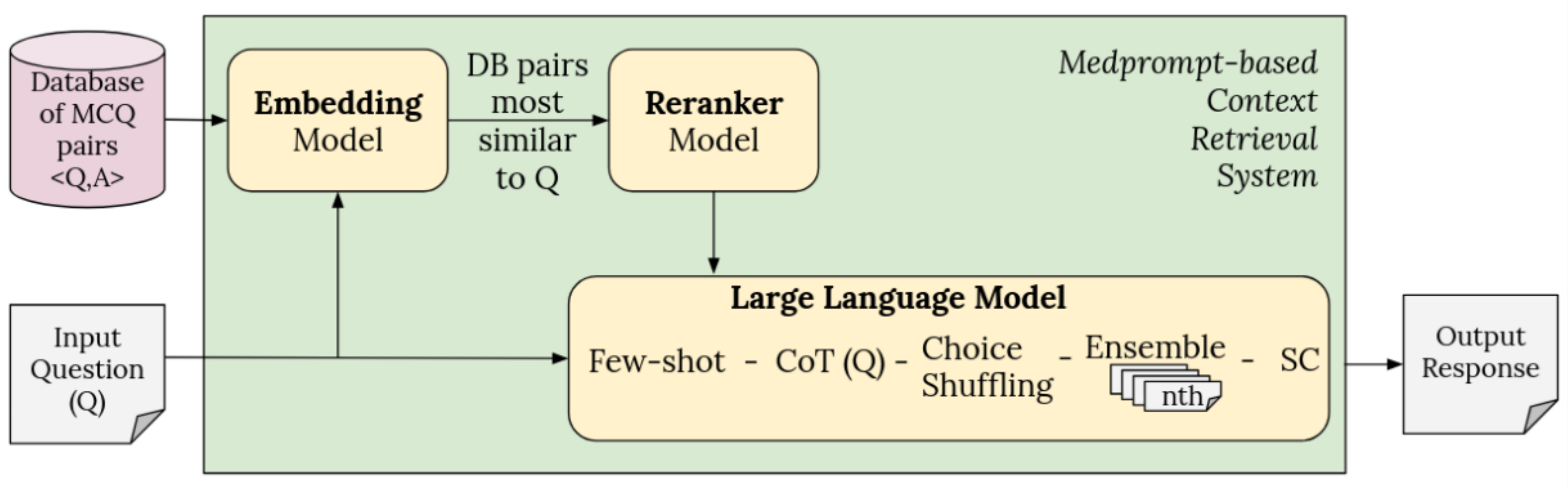


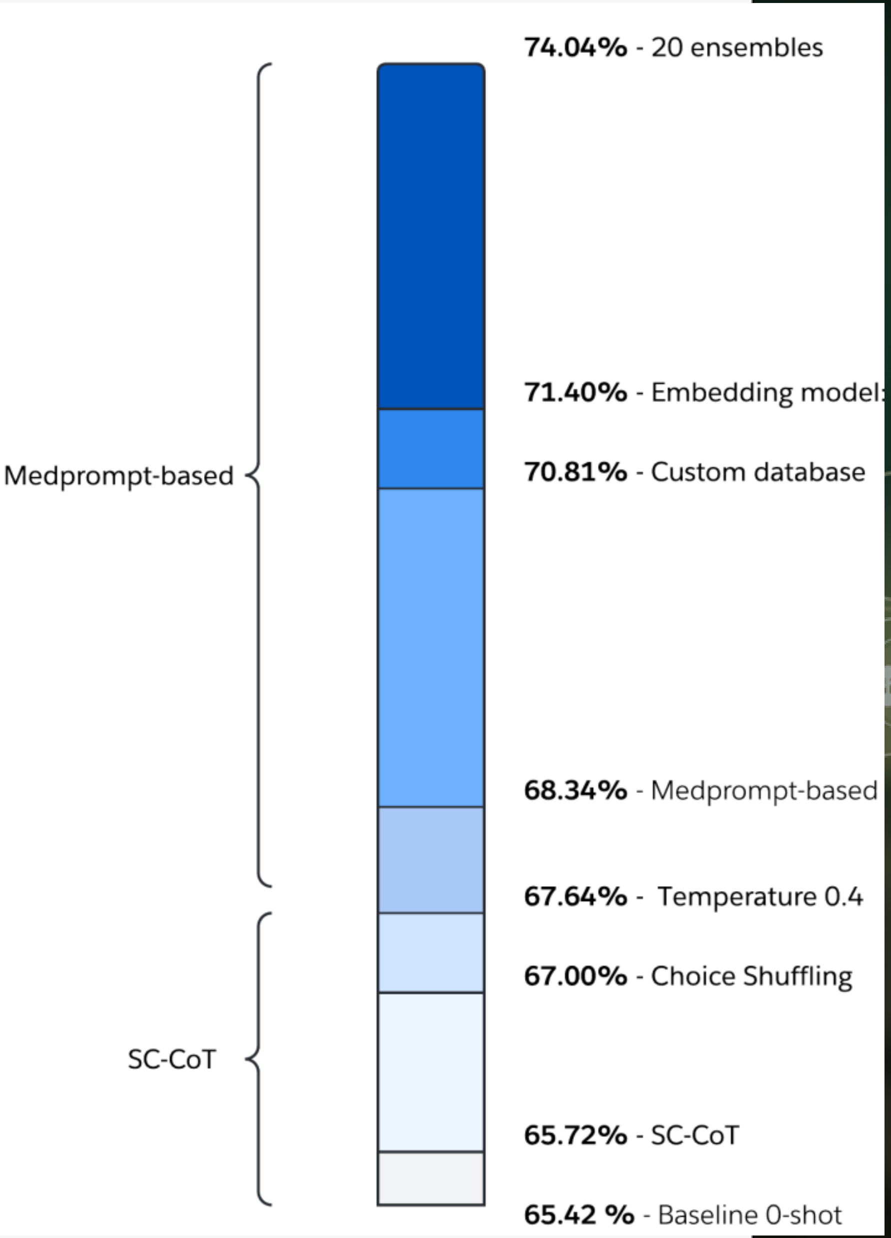
Boosting Inference

- RAG pipeline

Better Inference

➢ Boosting MedPrompt-based RAG system





RAG Repo

RAG Paper

Boosting Inference

- RAG pipeline
- Closed model comparison

Limits of Open LLMs for Healthcare

- 5-10% gain over plain LLM
- Bigger boost on smaller models
- Comparable performance wrt private models

Model	CareQA	MedMCQA	MedQA	MMLU	Average
Llama-3.1-8B	69.95	59.22	63.71	75.72	67.15
with CR	+6.07	+12.79	+17.36	+9.33	+11.39
Qwen2.5-7B	72.14	56.18	61.59	77.92	66.96
with CR	+3.08	+13.00	+12.64	+6.13	+8.71
Aloe-Beta-8B	70.77	59.57	64.65	76.50	67.87
with CR	+5.37	+12.72	+16.26	+7.60	+10.49
Llama-3.1-70B	83.72	72.15	79.73	87.45	80.76
with CR	+3.15	+5.69	+9.66	+3.84	+5.54
Qwen2.5-72B	85.45	69.26	77.85	88.81	80.34
with CR	+1.08	+7.55	+7.46	+2.75	+4.71
Aloe-Beta-70B	83.19	72.15	79.73	88.44	80.88
with CR	+4.38	+5.28	+9.11	+3.01	+5.45
DeepSeek-R1	88.33	73.34	82.48	91.27	83.86
with CR	+4.18	+8.94	+11.94	+3.61	+7.17
		Private models			
(OpenAI) **GPT-4 + Medprompt***	-	79.10	90.20	94.2	-
(Google) **MedPalm-2 + ER***	-	72.30	85.40	89.40	-
(OpenAI) **O1 + TPE***	-	83.90	96.00	95.28	-

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

RAG Paper

TDW on Foundation Models

Theme Development Workshops

July 10th, 2025
Thessaloniki, Greece
(Hybrid event)

Thank you!

dario.garcia@bsc.es

Hugging Face
GitHub
Paper

Hugging Face
Paper

GitHub
Paper

TURTLE

Hugging Face
Paper

Funded by the European Union

ELIAS ELLIOT elsa European Lighthouse on Secure and Safe AI

CENTRE FOR RESEARCH & TECHNOLOGY - HELLAS
Information Technologies Institute