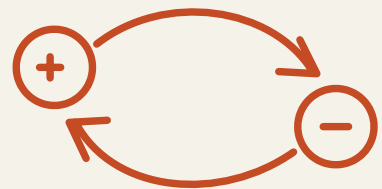
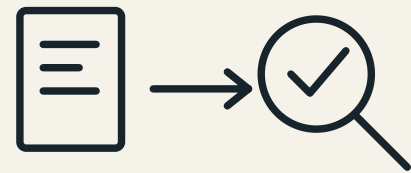


Self-play. Retrieval. Steering. **Evaluation.**



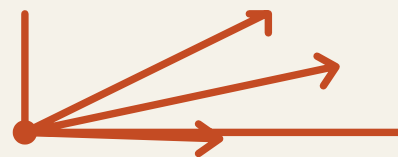
Self-play

Learning without
curated answers



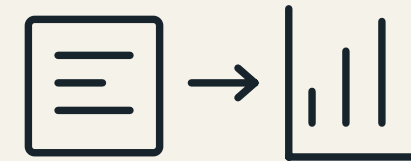
Retrieval

Inside text + image
attention heads



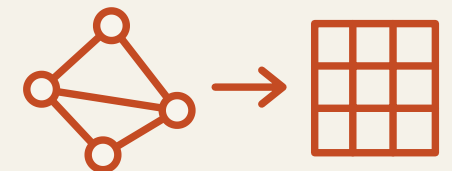
Steering

Composing multiple
behaviours



Evaluation

Beyond headline
benchmark scores



Graph learning

Interfaces for relational
generalisation

Self-play, retrieval, and evaluation.

NeurIPS

SCOPE

Self-play for open-ended tasks
Cyrus Kwan et al.

LOCOS

What retrieval heads actually write
Aryo Gema et al.

Composed image retrieval

When benchmarks admit shortcuts
Matteo Attimonelli et al.

AI psychophysics

Grounding behavioural evaluation
Jonas Leo Mueller, Dario Zanca et al.

EMNLP

Multimodal retrieval heads

Can retrieval heads see images?
Aaron Branson Cigres Li et al.

Hidden VLM instability

Same answer, different representations
Farooq Ahmad Wani et al.

Composing steering vectors

Controlling behaviours without interference
Giwon Hong et al.

LoG

Graph-to-table interfaces

What controls relational generalisation?
Alessandro Cagiano et al.

PREPRINTS

SPSD

Self-play search distillation
Lorenzo Molfetta et al.

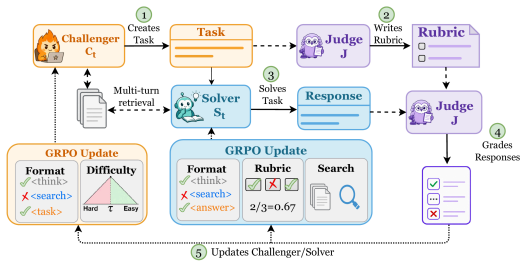
FACE-Eval

Chain-of-thought faithfulness
Aryo Gema et al.

VLM frequency filtering

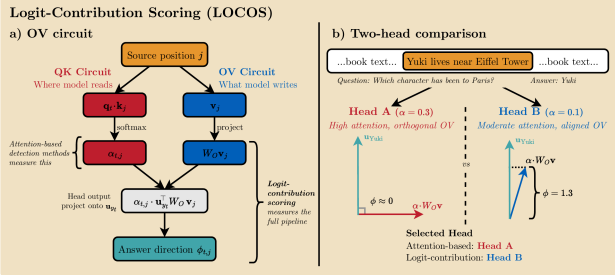
Why verbose prompts help robustness
Farooq Ahmad Wani et al.

NeurIPS
SCOPE



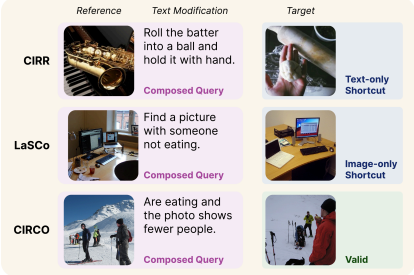
Cyrus Kwan et al.

NeurIPS
LOCOS



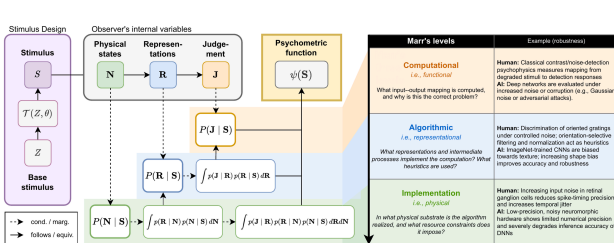
Aryo Gema et al.

NeurIPS
Composed image
retrieval



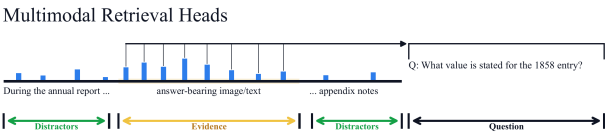
Matteo Attimonelli et al.

NeurIPS
AI psychophysics



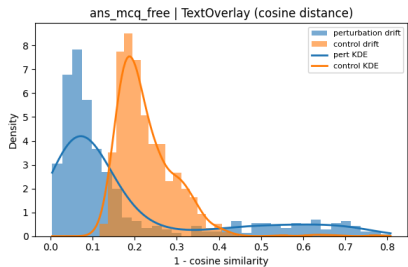
Jonas Leo Mueller, Dario Zanca et al.

EMNLP
Multimodal retrieval
heads



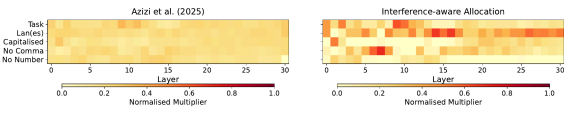
Aaron Branson Cigres Li et al.

EMNLP
Hidden VLM instability



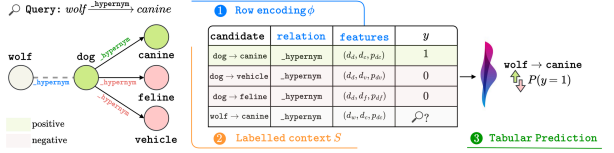
Farooq Ahmad Wani et al.

EMNLP
Composing steering
vectors



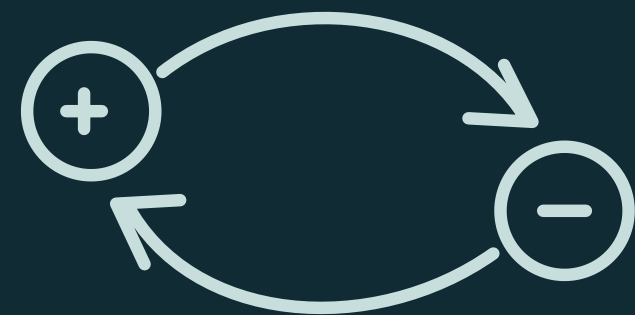
Giwon Hong et al.

LoG
Graph-to-table interfaces



Alessandro Cagiano et al.

Learn



NeurIPS
SCOPE
Self-play for open-ended tasks

EMNLP
Composing steering vectors
Controlling behaviours without interference

LoG
Graph-to-table interfaces
What controls relational generalisation?

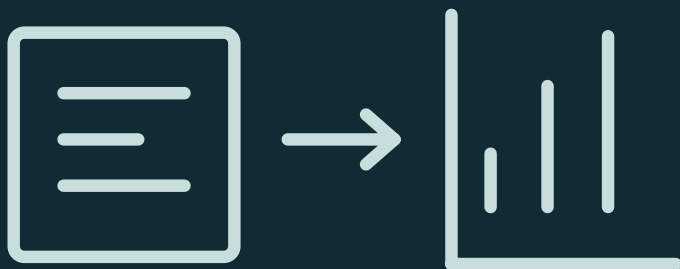
Inspect



NeurIPS
LOCOS
What retrieval heads actually write

EMNLP
Multimodal retrieval heads
Can retrieval heads see images?

Evaluate



NeurIPS
Composed image retrieval
When benchmarks admit shortcuts

NeurIPS
AI psychophysics
Grounding behavioural evaluation

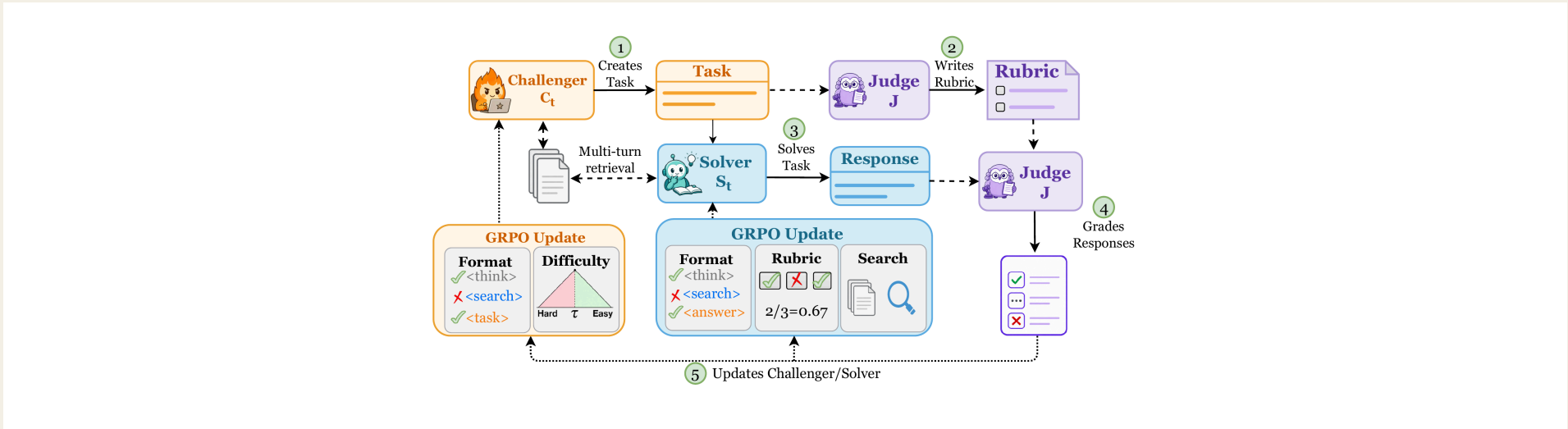
EMNLP
Hidden VLM instability
Same answer, different representations

ML

Learning.
Retrieval.
Control.

SCOPE Cyrus Kwan et al.	NeurIPS
LOCOS Aryo Gema et al.	NeurIPS
Composed image retrieval Matteo Attimonelli et al.	NeurIPS
AI psychophysics Jonas Leo Mueller, Dario Zanca et al.	NeurIPS
Multimodal retrieval heads Aaron Branson Cigres Li et al.	EMNLP
Hidden VLM instability Farooq Ahmad Wani et al.	EMNLP
Composing steering vectors Giwon Hong et al.	EMNLP
Graph-to-table interfaces Alessandro Cagiano et al.	LoG

Learning from self-play. Looking inside models. Questioning the tests.



SCOPE / Cyrus Kwan et al.

NeurIPS

SCOPE
LOCOS
Composed image retrieval
AI psychophysics

EMNLP

Multimodal retrieval heads
Hidden VLM instability
Composing steering vectors

LoG

Graph-to-table interfaces

How models learn. What they retrieve. How we evaluate.

PREPRINTS

SPSD / FACE-Eval

VLM frequency filtering

NeurIPS

SCOPE

LOCOS

Composed image retrieval

AI psychophysics

EMNLP

Multimodal retrieval heads

Hidden VLM instability

Composing steering vectors

LoG

Graph-to-table interfaces

NeurIPS

SCOPE

Cyrus Kwan et al.

Composed image retrieval

Matteo Attimonelli et al.

LOCOS

Aryo Gema et al.

AI psychophysics

Jonas Leo Mueller, Dario Zanca et al.

EMNLP

Multimodal retrieval heads

Aaron Branson Cigres Li et al.

Composing steering vectors

Giwon Hong et al.

Hidden VLM instability

Farooq Ahmad Wani et al.

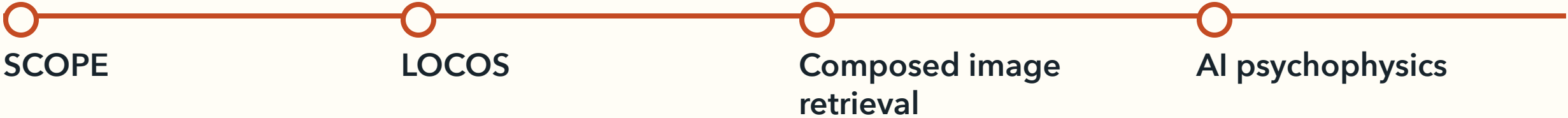
LoG

Graph-to-table interfaces

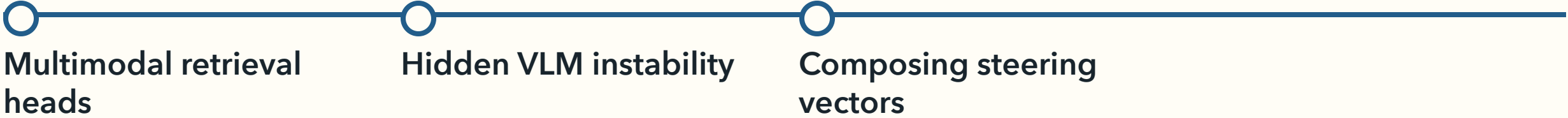
Alessandro Cagiano et al.

Self-play. Retrieval. Evaluation.

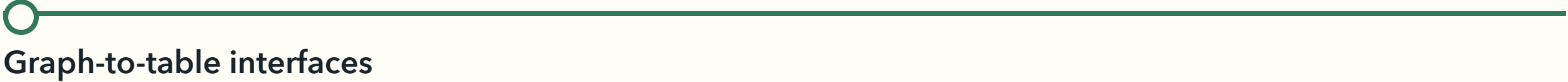
NeurIPS



EMNLP



LoG



PREPRINTS

SPSD / FACE-Eval / VLM frequency filtering

Learning and model behaviour.

NeurIPS**SCOPE**

Self-play for open-ended tasks

Cyrus Kwan et al.

NeurIPS**LOCOS**

What retrieval heads actually write

Aryo Gema et al.

NeurIPS**Composed image retrieval**

When benchmarks admit shortcuts

Matteo Attimonelli et al.

NeurIPS**AI psychophysics**

Grounding behavioural evaluation

Jonas Leo Mueller, Dario Zanca et al.

EMNLP**Multimodal retrieval heads**

Can retrieval heads see images?

Aaron Branson Cigres Li et al.

EMNLP**Hidden VLM instability**

Same answer, different representations

Farooq Ahmad Wani et al.

EMNLP**Composing steering vectors**

Controlling behaviours without interference

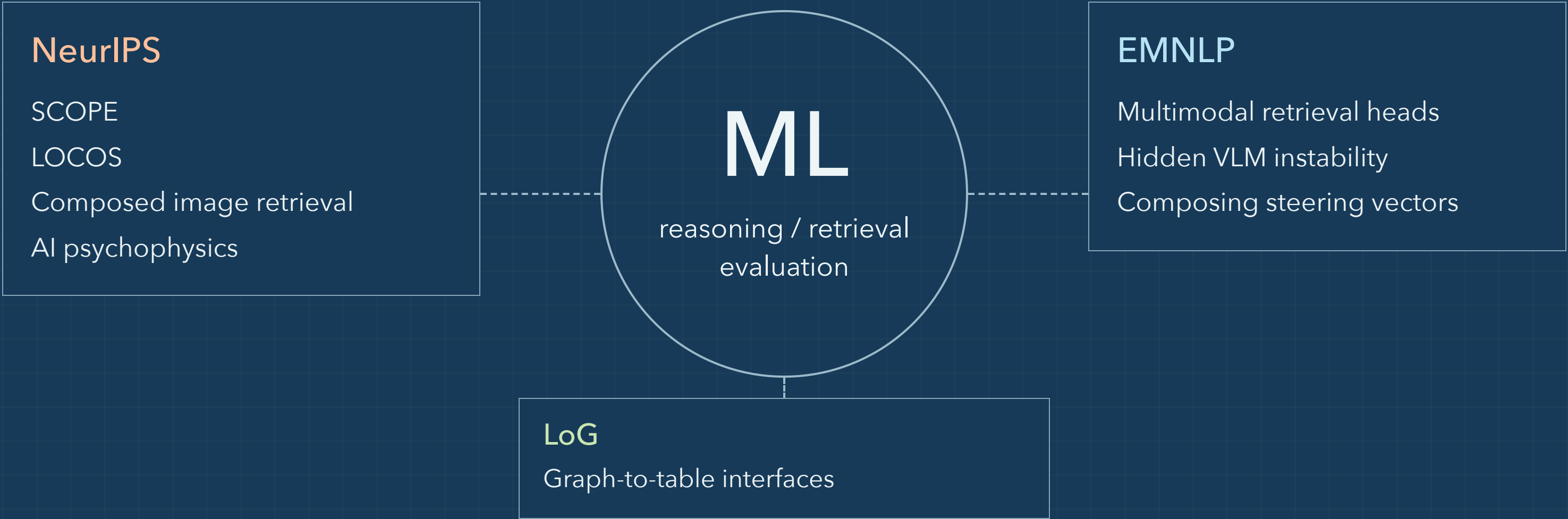
Giwon Hong et al.

LoG**Graph-to-table interfaces**

What controls relational generalisation?

Alessandro Cagiano et al.

SCOPE	Cyrus Kwan et al.	NeurIPS
LOCOS	Aryo Gema et al.	NeurIPS
Composed image retrieval	Matteo Attimonelli et al.	NeurIPS
AI psychophysics	Jonas Leo Mueller, Dario Zanca et al.	NeurIPS
Multimodal retrieval heads	Aaron Branson Cigres Li et al.	EMNLP
Hidden VLM instability	Farooq Ahmad Wani et al.	EMNLP
Composing steering vectors	Giwon Hong et al.	EMNLP
Graph-to-table interfaces	Alessandro Cagiano et al.	LoG



Self-play, retrieval, and steering.

NeurIPS

SCOPE

LOCOS

Composed image retrieval

AI psychophysics

EMNLP

Multimodal retrieval heads

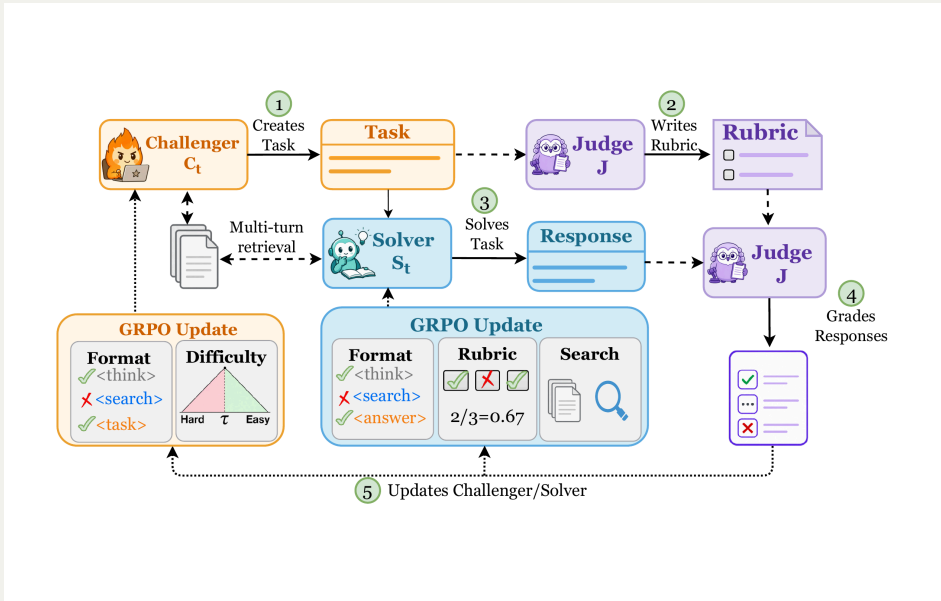
Hidden VLM instability

Composing steering vectors

LoG

Graph-to-table interfaces

Learn / control



SCOPE

NeurIPS

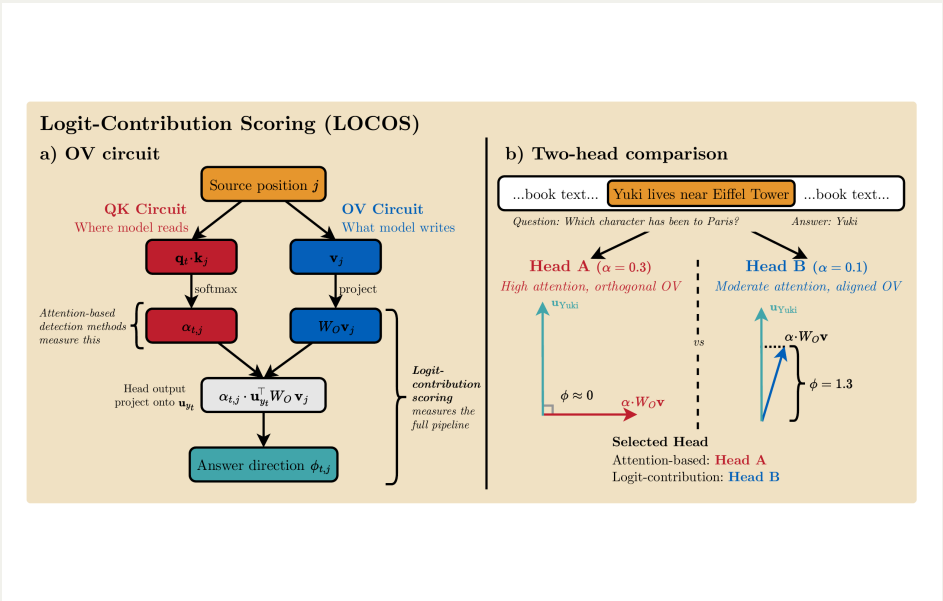
Composing steering vectors

EMNLP

Graph-to-table interfaces

LoG

Inspect



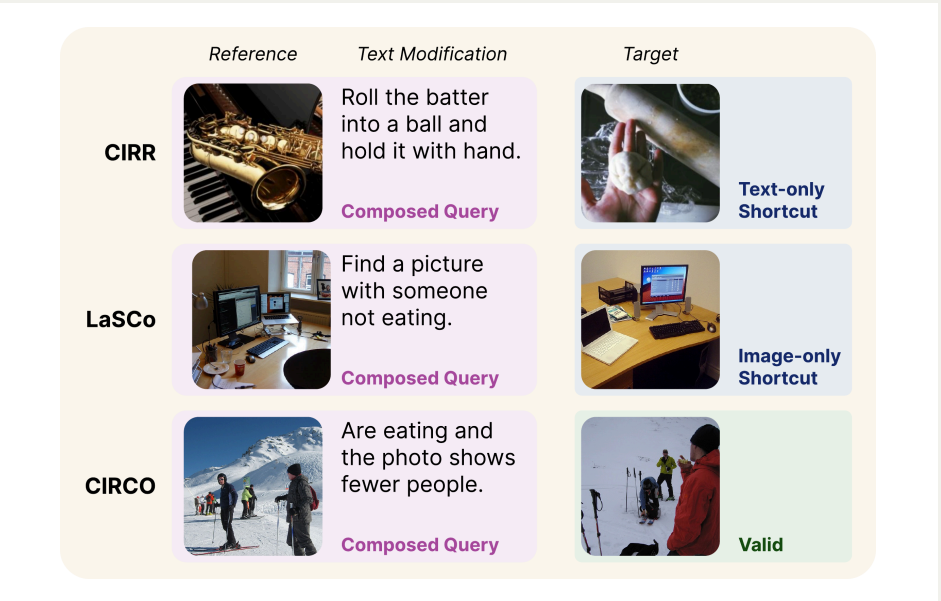
LOCOS

NeurIPS

Multimodal retrieval heads

EMNLP

Evaluate



Composed image retrieval

NeurIPS

AI psychophysics

NeurIPS

Hidden VLM instability

EMNLP

{NeurIPS, EMNLP, LoG}

Self-play, retrieval, steering, evaluation,
and relational generalisation.

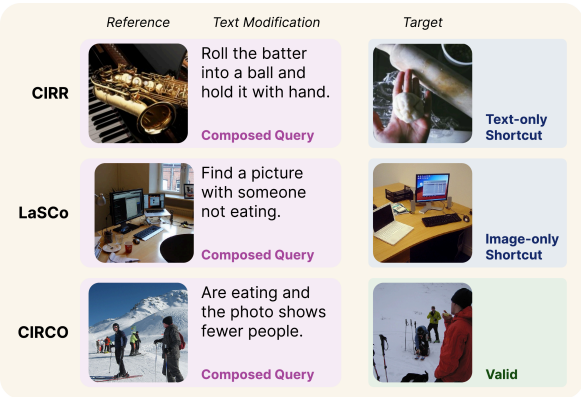
NeurIPS

SCOPE

LOCOS

Composed image retrieval

AI psychophysics

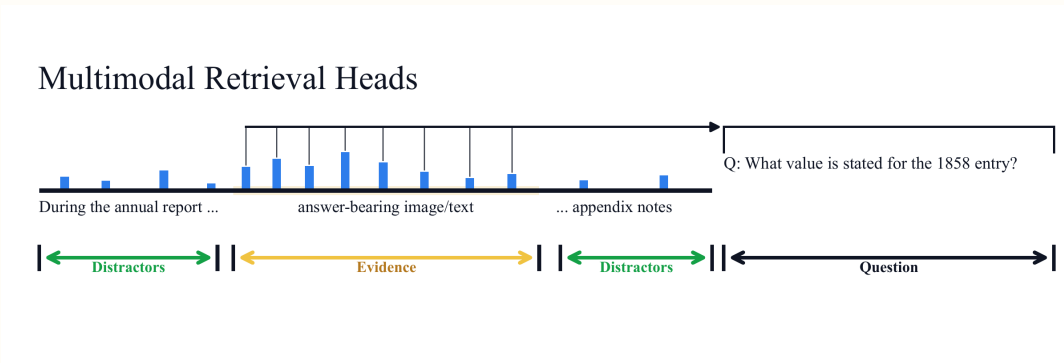


EMNLP

Multimodal retrieval heads

Hidden VLM instability

Composing steering vectors



LoG

Graph-to-table interfaces

