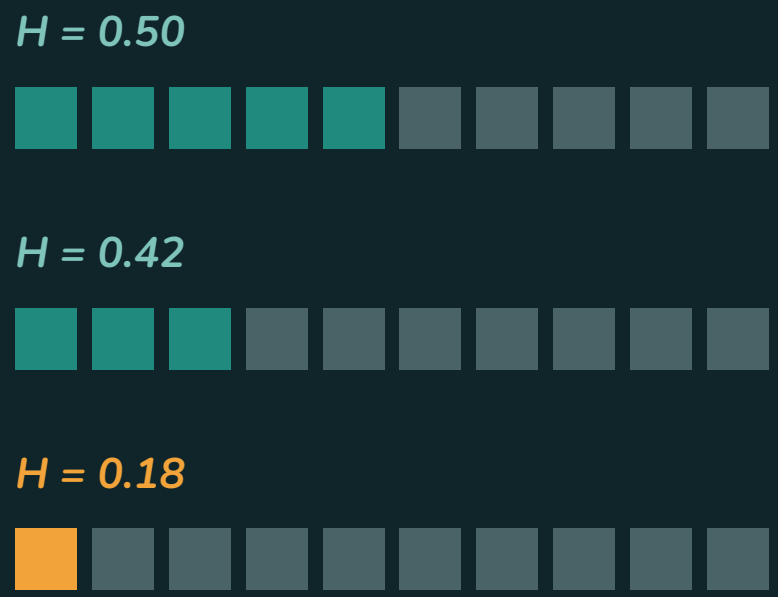
EDMA: Entropy-Driven Multimodal Answering

Quantifying how each modality moves us toward the answer

Emanuele Mezzi · Gertjan Burghouts · Fabio Massacci · Mengyuan Zhang

Vrije Universiteit Amsterdam · TNO



Which Modality Got Us to the Answer?

THE QUESTION

What is the state in the Contour Airlines city destinations that has a tall metal arch on a river?

Table: which cities Contour Airlines flies to

Image: which city has a metal arch

Step-wise prompting

CoT, planning and action chains split the task into steps

Uncertainty-aware methods

Use confidence as one global signal

THE GAP

How much does each modality, and their fusion, narrow down the candidate answers?

Logical Entropy: Counting Distinctions

$$H(P) = 1 - \sum p_i^2$$

- Probability that two random candidates fall in different blocks
- Only counts candidates: no calibrated probabilities needed
- Lower H means a sharper separation

A requirement splits 10 candidates



H = 0.50 maximal uncertainty

...or singles one candidate out



H = 0.18 one candidate stands out

EDMA: Answering as Entropy Reduction

1)

Requirements

The LLM extracts what the answer must satisfy and picks the starting modalities.

2)

Unimodal partitions

Per modality and requirement: split elements into satisfy / not satisfy and compute H.

3)

Cross-modal conditioning

Evidence from one modality filters candidates in another. Every modality takes a turn as target.

4)

Trajectory selection

Pick the trajectory with the lowest final entropy and answer from it.

ΔH = (H unimodal – H multimodal) / H unimodal the entropy change credited to each fusion step

Example: what each modality sees alone

QUESTION

What is the state in the Contour Airlines city destinations that has a tall metal arch on a river?

ANSWER REQUIREMENTS

- ***S*** State
- ***C*** Contour destination
- ***A*** Metal arch



Image · Metal arch

H = 0.38



Table · Contour destination

H = 0.44



Table · State

H = 0.49

The image separates sharply on its own, but the answer is a state in the table. How do we combine them?

Fusing image into table cuts entropy 54%

Final logical entropy per answering trajectory
(lower is better)

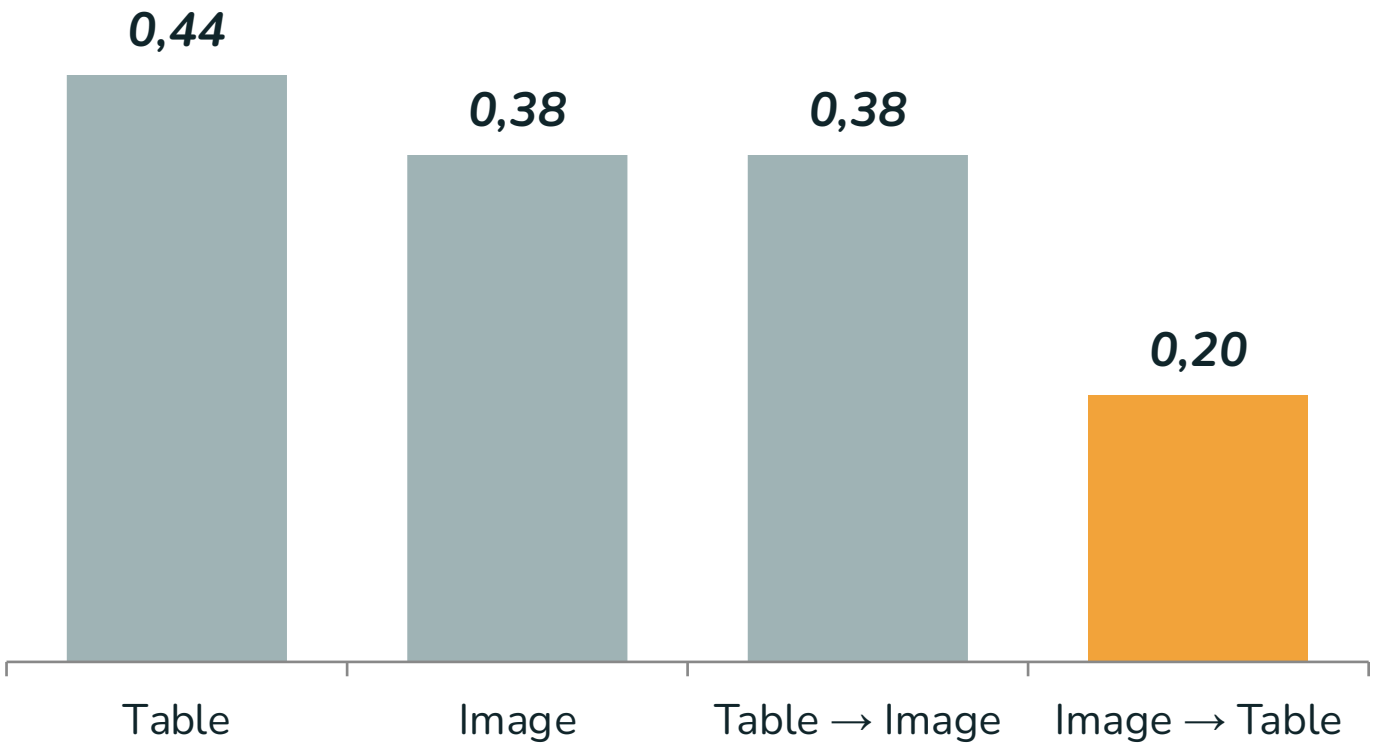


Image → Table

0.44 → 0.20

54% entropy reduction

Table → Image: 0.38 → 0.38, ΔH = 0%. Table evidence adds nothing to the images.

Selected: Image → Table isolates St. Louis, answer: Missouri.

Gains Hold at Matched Compute

EDMA vs. the strongest prompting baseline

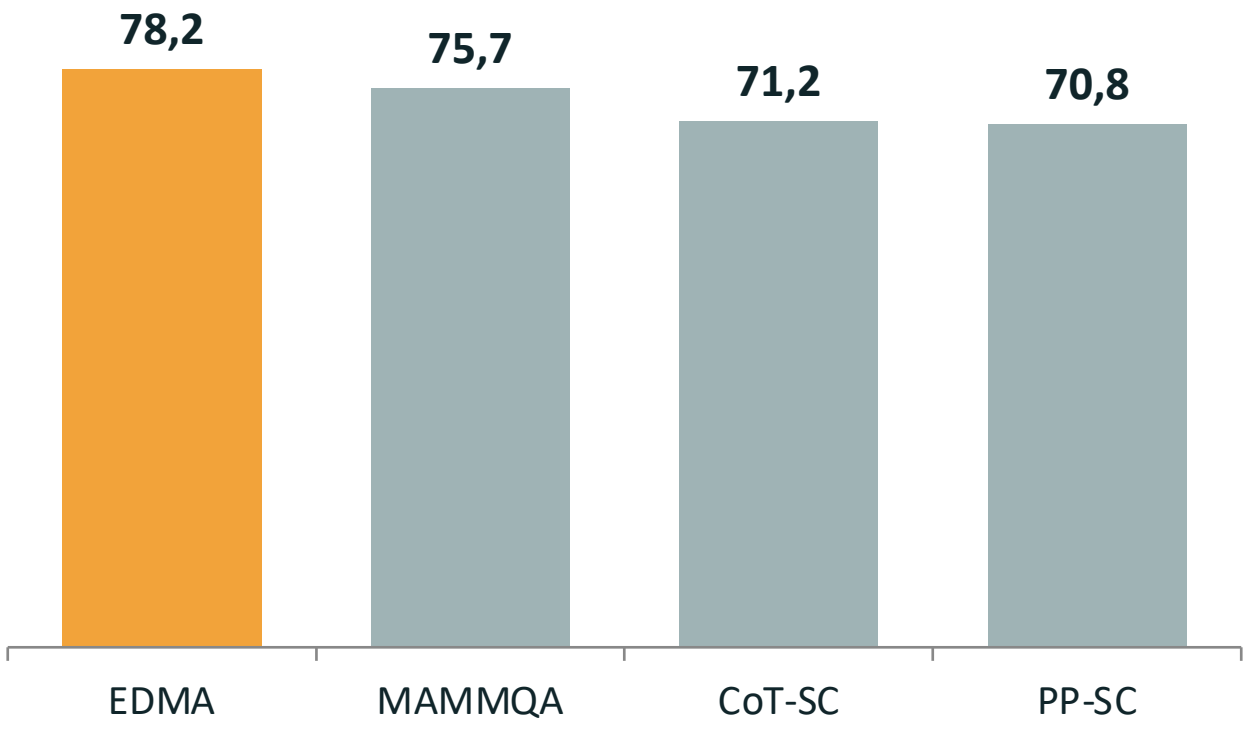
+22.9 EM

+12.8 F1

+4.6 LLM-judge

Average over 6 backbones and 2 benchmarks (MultiModalQA, ManyModalQA).

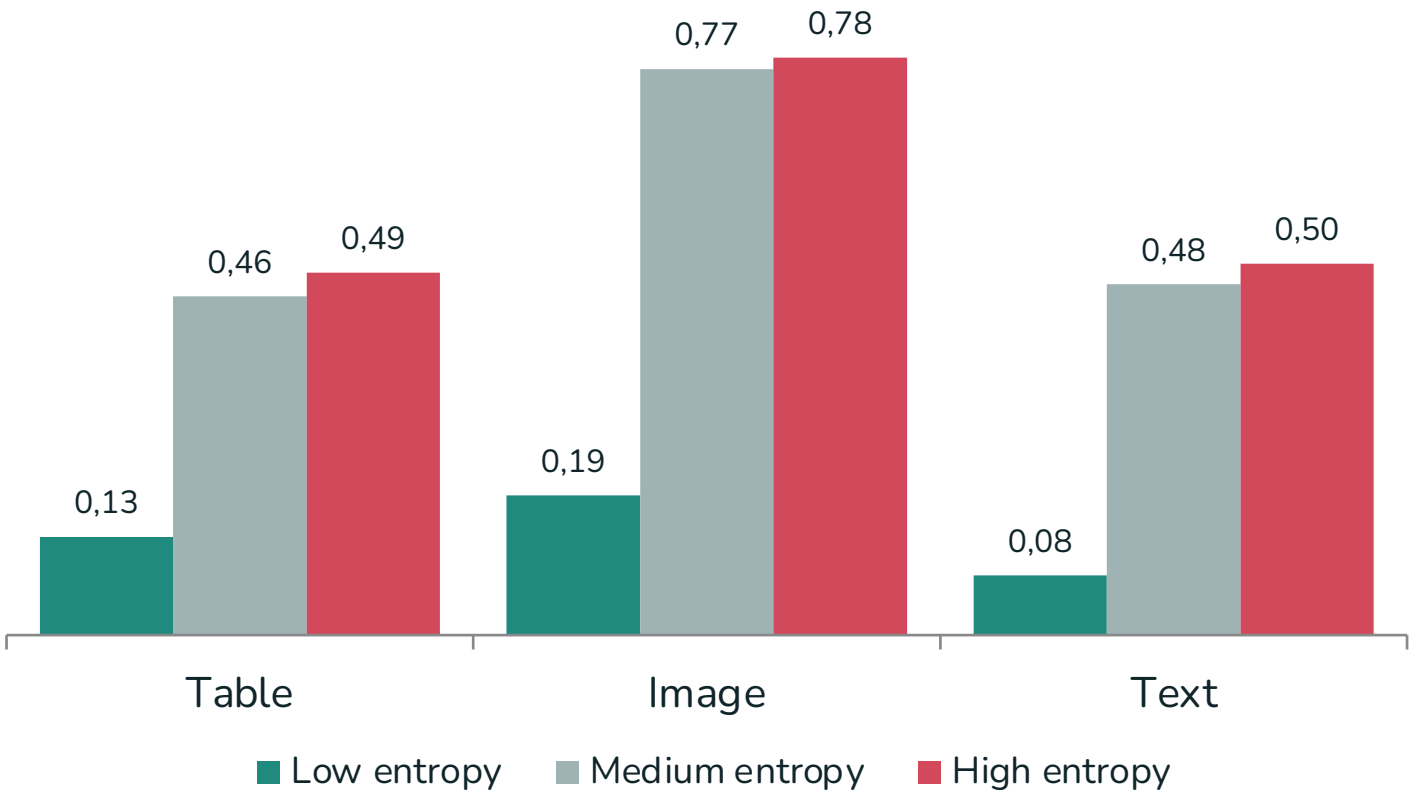
MultiModalQA, LLM-judge, matched-cost baselines



Average over 4 backbones

Lower Entropy, Fewer Wrong Answers

False discovery rate by entropy of the selected trajectory



Wrong Answers Sit at Higher Entropy

Table 0.21 right vs 0.27 wrong

Image 0.21 right vs 0.26 wrong

Text 0.20 right vs 0.24 wrong

Same Question, Forced Trajectory

Lowest, median and highest entropy path: error rate rises monotonically in every modality.

Entropy Makes Answering Traceable

Multimodal answering becomes a trajectory we can measure, trace and select.

WHY IT WORKS (LLM-JUDGE POINTS LOST WHEN REMOVED)

−32.1

random trajectory instead of lowest entropy

−14.4

no cross-modal conditioning

WHAT IT COSTS (VS. SINGLE-PASS PROMPTING)

5–6×

monetary cost

20–46×

latency

Emanuele Mezzi · e.mezzi@vu.nl