DA421M Course Project

ADAPTIVE REASONING ENGINE FOR KNOWLEDGE-BASED VISUAL QUESTION ANSWERING (ARE-VQA)

Abhishek Kumar 220101002
Adarsh Gupta 220101003

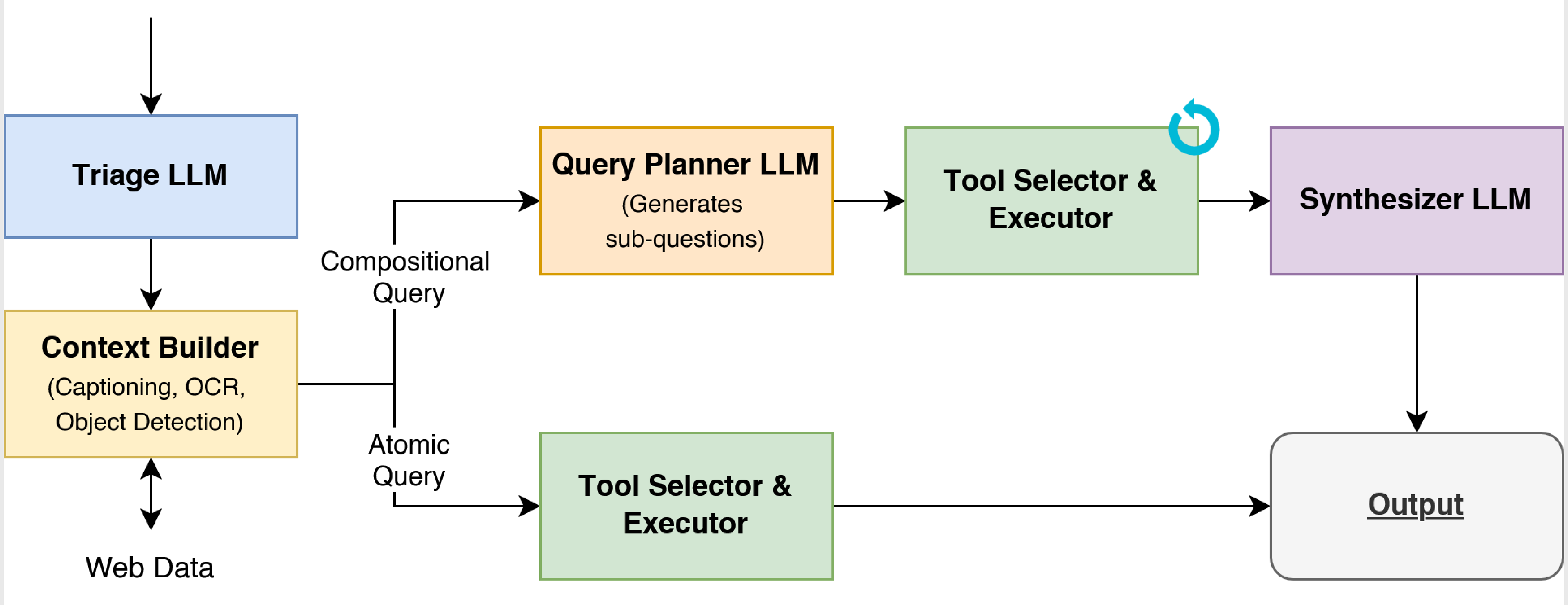
INTRODUCTION

- **Visual Question Answering** (VQA) involves answering natural language questions about images.
- Knowledge-Based VQA (KB-VQA) extends this by requiring reasoning over external world knowledge not present in the image.
- Existing systems are either:
    - Complex pipelines combining visual models with structured databases (e.g. Wikidata)
    - Black-box LLM systems that treat the model as a single reasoning engine without adaptation.
- **Proposed Objective:** Build a modular, adaptive reasoning engine that routes questions intelligently, retrieves necessary evidence, and provides interpretable answers.

MOTIVATION

- Recent KB-VQA focus on using Large Language Models (LLMs) to perform reasoning using textual prompts instead of explicit training. But have drawbacks like :
  - Uniform treatment of all question types (no adaptive routing).
  - Inefficient for simple queries, inadequate for compositional reasoning.
  - Monolithic LLM approaches lack interpretability and adaptability
- To overcome these limitations, we showcase a modular reasoning pipeline that mimics human decision-making. Each stage of ARE-VQA performs a specific reasoning task, analyzing question complexity, retrieving relevant knowledge, and synthesizing multi-step results.

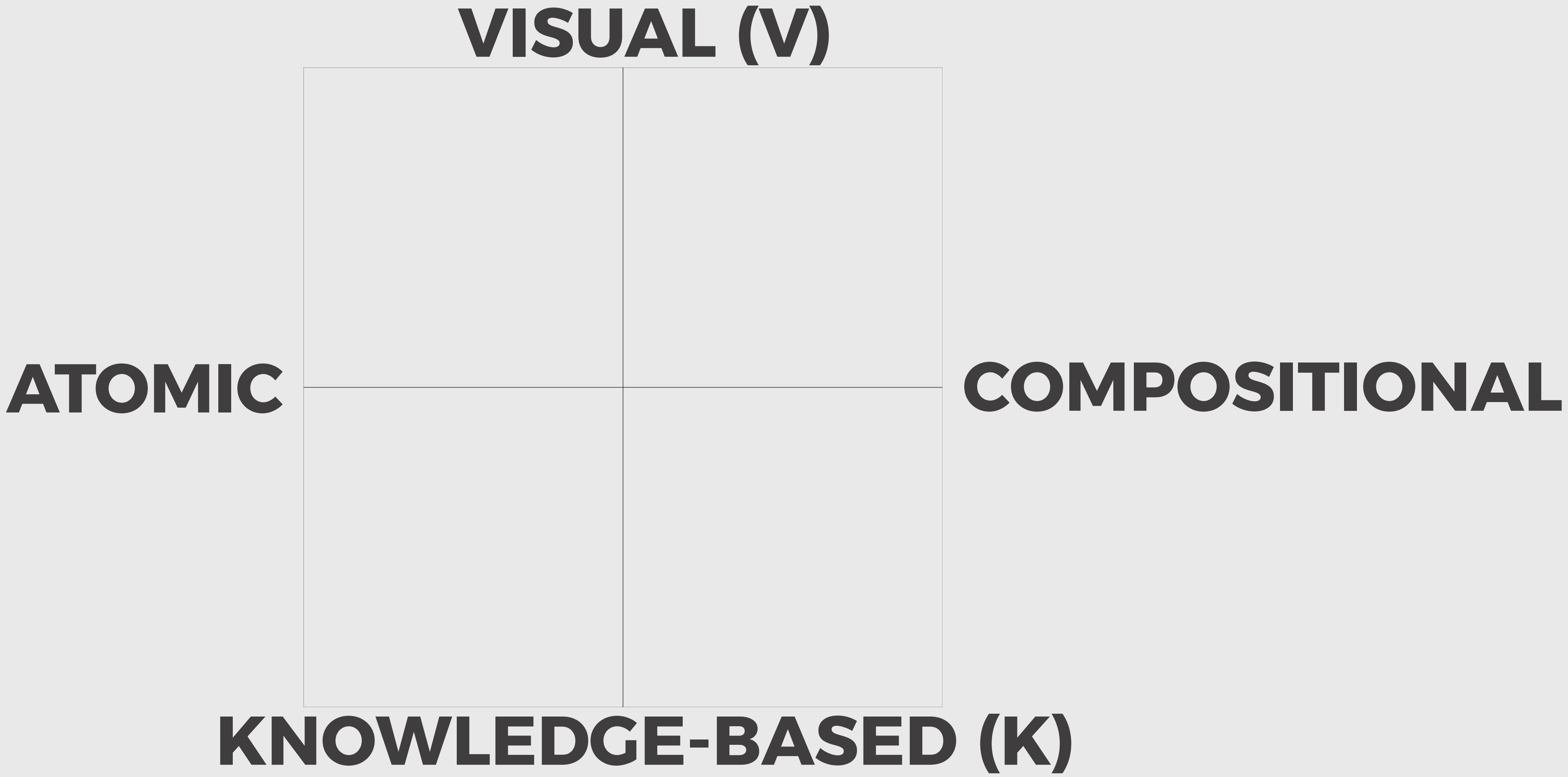
PROPOSED MODEL PIPELINE



We are using five-stage pipeline designed to reason about visual questions in an adaptive and human-like manner.

MODULE 1: TRIAGE LLM

- INPUTS: QUESTION + IMAGE
- OUTPUTS: ONE OF 4 CLASSES: V-A, V-C, K-A, K-C.
- DONE BY A VLM (VISION-LANGUAGE MODEL)



MODULE 2: CONTEXT BUILDER

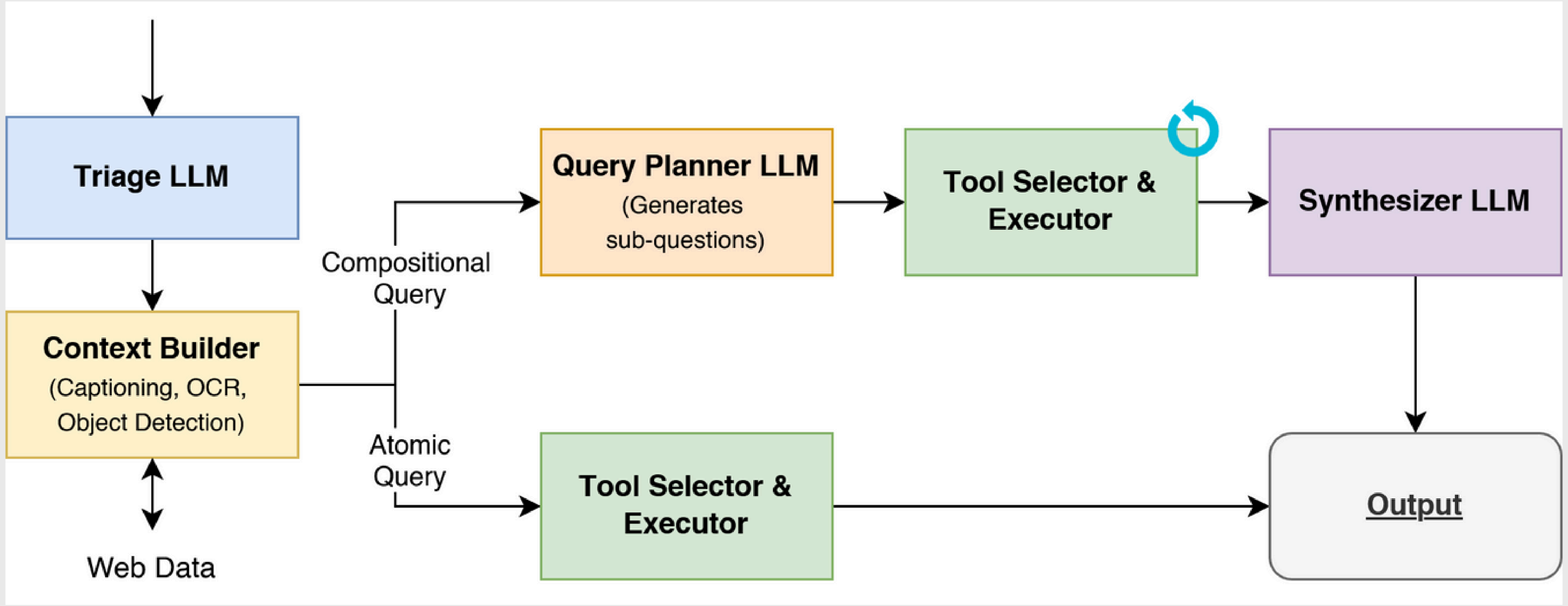
STEP 1) VLM EXTRACTS IMAGE CAPTION
STEP 2) VLM EXTRACTS KEY VISUAL ENTITIES
STEP 3) OCR DONE USING PYTESSERACT
STEP 4) IF QUESTION TAGGED AS KNOWLEDGE BASED:

- LLM (QWEN3:8B) GENERATES A KNOWLEDGE SNIPPET AND A SEARCH QUERY
- SEARCH QUERY RAN ON INTERNET VIA API SEARCH

MODULE 3: QUERY PLANNER

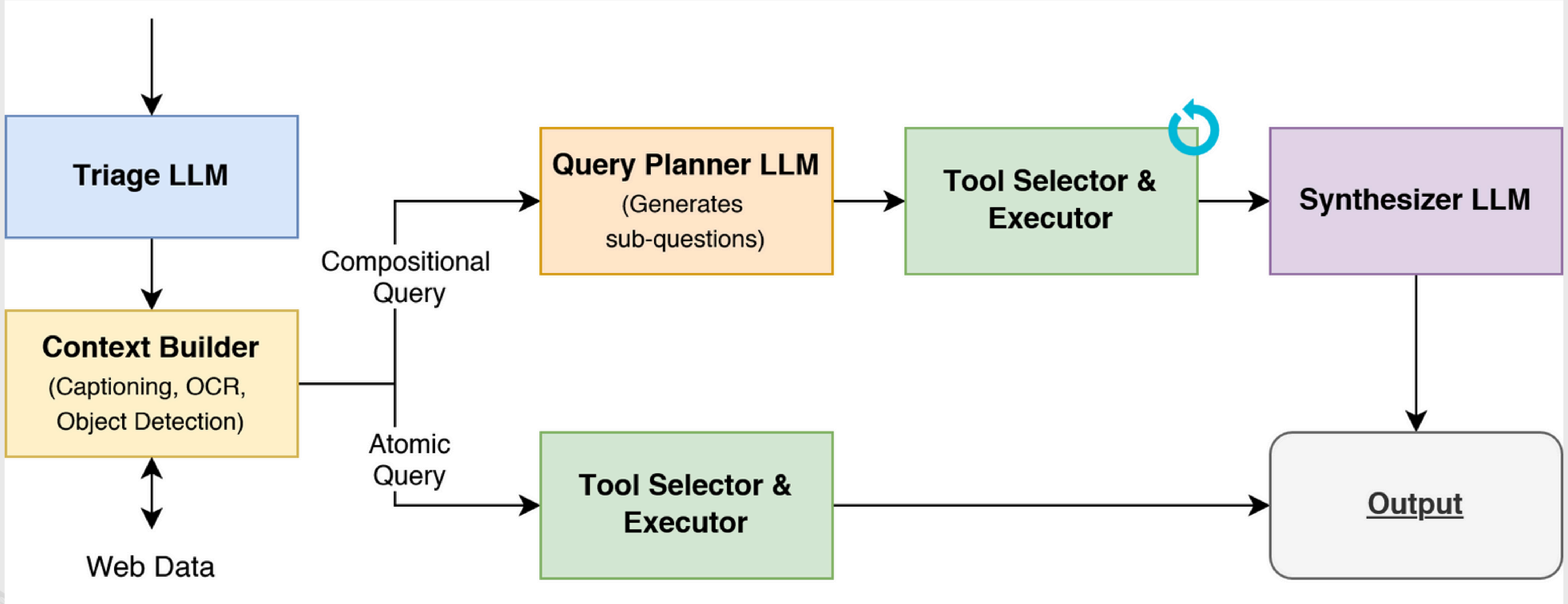
THE QUERY PLANNER IS EXECUTED FOR COMPOSITIONAL QUERIES AND DIVIDES THE COMPOSITIONAL QUERY INTO A SERIES OF ATOMIC QUESTIONS WHICH CAN BE SOLVED ITERATIVELY.

THE PROMPT ENFORCES ENTITY CONTINUATION TO AVOID HALLUCINATIONS



MODULE 4 & 5: TOOL SELECTOR, EXECUTOR & SYNTHESIZER

4) TOOL SELECTOR & EXECUTOR: PICKS THE BEST PROMPT/TOOL (VQA / OCR / KNOWLEDGE-INTEGRATED) AND ANSWERS ATOMIC STEPS.
5) SYNTHESIZER: COMPOSES INTERMEDIATE ANSWERS INTO A FINAL RESPONSE FOR COMPOSITIONAL QUERIES.



DATASET DETAILS

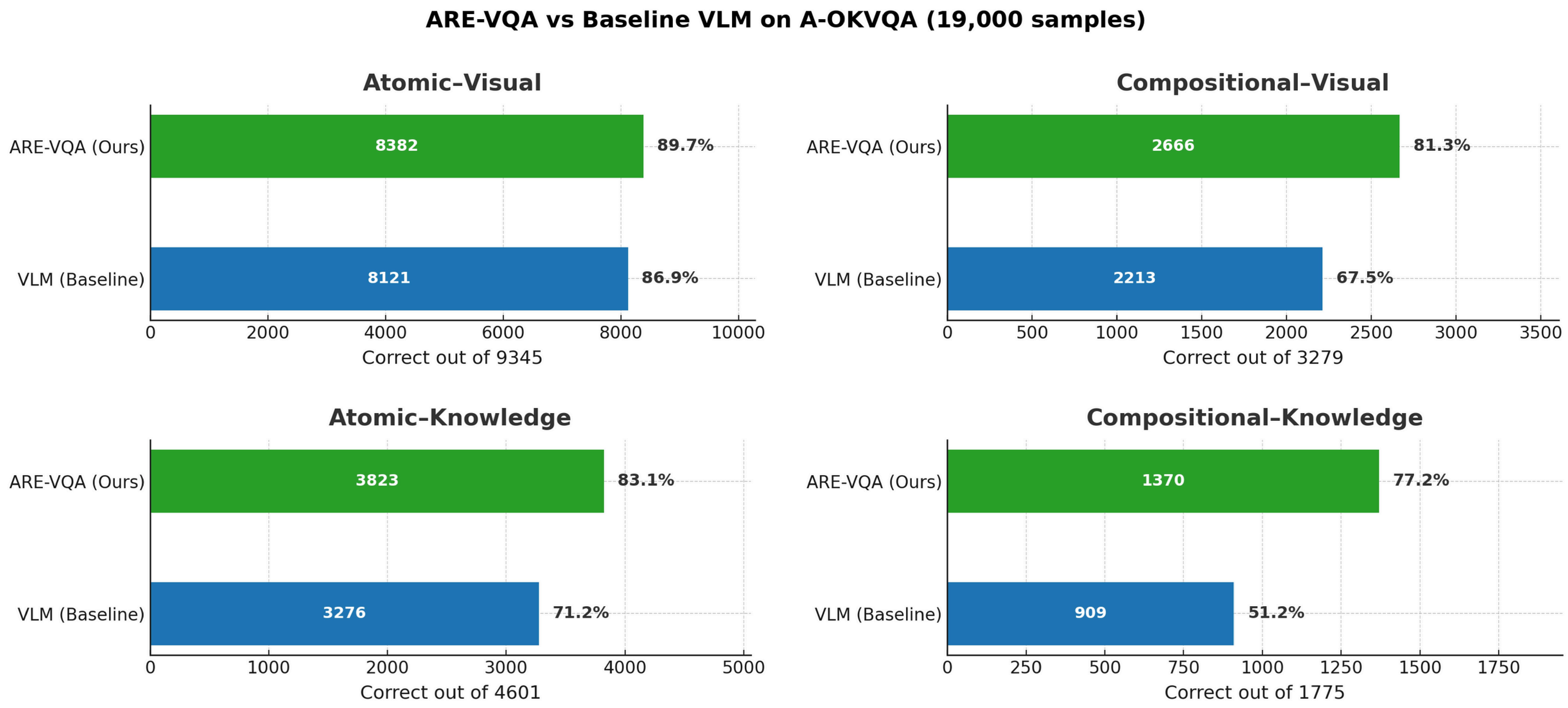
DATASET USED: **A-OKVQA DATASET**

TOTAL SAMPLE COUNT: **19,000**

TRIAGE SPLIT:

- VISUAL–ATOMIC (V-A): 9345
- VISUAL–COMPOSITIONAL (V-C): 3279
- KNOWLEDGE–ATOMIC (K-A): 4601
- KNOWLEDGE–COMPOSITIONAL (K-C): 1775

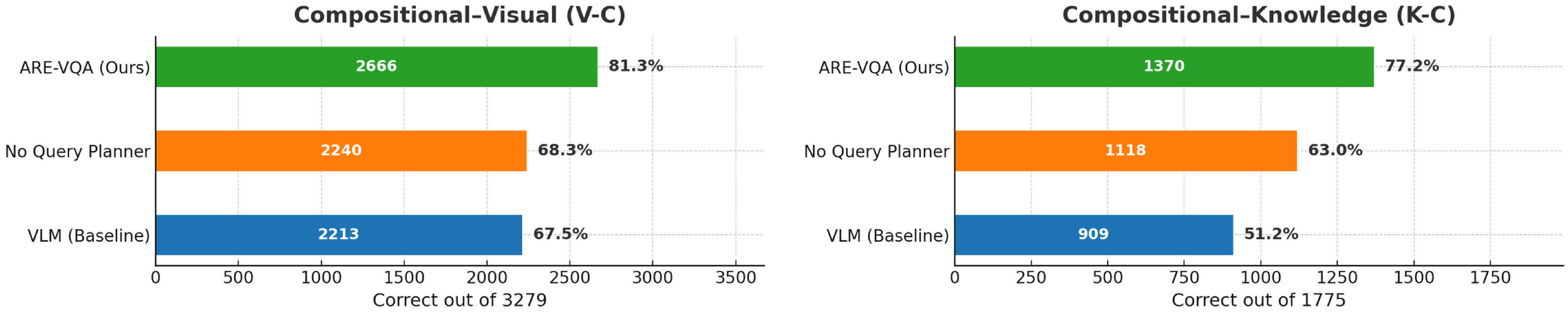
RESULTS



Overall accuracy — VLM (Baseline): 73.6% (13969/19000); ARE-VQA (Ours): 85.9% (16328/19000).

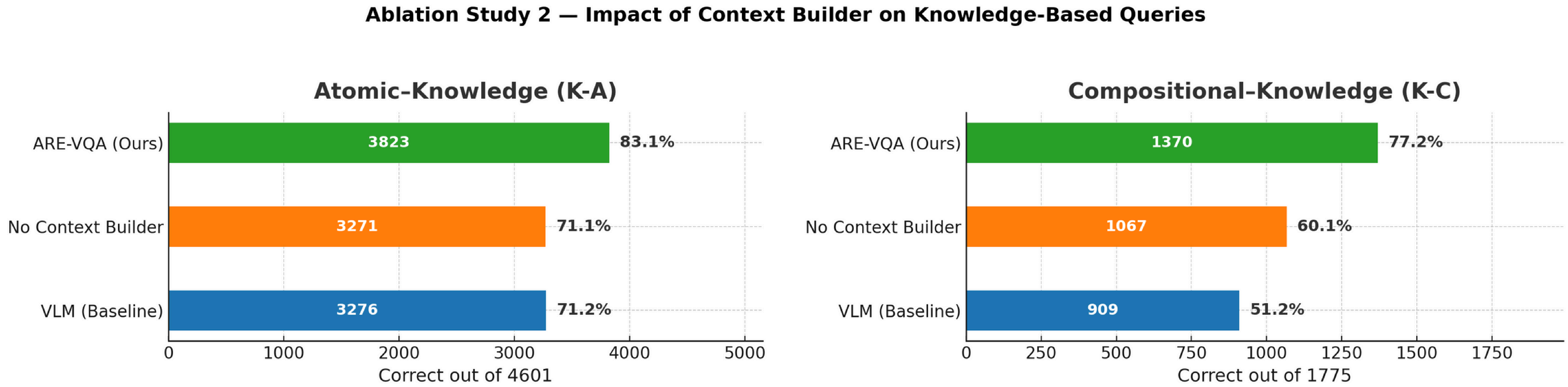
ABLATION STUDY - 1

Ablation Study 1 — Impact of Query Planner on Compositional Queries



Ablation 1 (No Query Planner): Δ vs Baseline — V-C: +0.8 pts; K-C: +11.8 pts. ARE-VQA improves to V-C: 81.3%, K-C: 77.2%.

ABLATION STUDY - 2



Ablation 2 (No Context Builder): Δ vs Baseline — K-A: -0.1 pts; K-C: +8.9 pts. ARE-VQA improves to K-A: 83.1%, K-C: 77.2%.

FUTURE WORK & CONCLUSION

Conclusion:

- ARE-VQA achieves higher accuracy and interpretability through adaptive modular reasoning.
- Strong improvements on knowledge-based and compositional questions validate the pipeline design.

Future Work:

- Enhance triage and tool selection via adaptive learning.
- Test generalization on broader multimodal tasks.
- Integrate cited reasoning for improved transparency.

THANK YOU

GitHub Link (Code + Report):
https://github.com/CoolSunflower/ARE-VQA