CS 668 Applied Large Language Models
Spring Semester, 2026
Doc 12 Part 3 & More Attention
Feb 19, 2026

Part 4

Assignment update for part 4

Don’t need parts 1- 3

Readings

Build a Large Language Model (from Scratch)

QUESTION: What GPA is required for graduate students to remain in good standing?

Chunk 1 | Page 75 | Score: 0.7989
Preview: maintain a cumulative grade point average of at least 2.85 in all
units attempted subsequent to admission to the university.
Students in a graduate d ...

--- LLM Answer ---
Graduate students must maintain at least a 3.0 grade point average in graduate courses taken in the degree program to remain in good standing.

QUESTION: What graduate standing is required to be appointed as a Graduate Teaching Associate (TA)?

EXPECTED: Graduate TAs must be admitted to San Diego State University with classified or conditionally classified graduate standing.

GENERATED: Graduate TAs must be admitted to San Diego State University with classified or conditionally classified graduate standing.

RELEVANCE SCORE: 1.00
TOP CHUNK SCORE: 0.6421
AVG CHUNK SCORE: 0.7167

EVAL_PROMPT = """You are a evaluation assistant. Score the generated answer from 0 to 5:
5 = Perfect: captures all key information from the expected answer
4 = Good: captures most key information, minor details missing
3 = Acceptable: captures the main point but misses some details
2 = Patrial: has some relevant information but misses key points
1 = Poor barely related to the expected answer
0 = Wrong: completely incorrect or irrelevant

Focus on meaning, not exact wording. Respond with ONLY a JSON object: {"score": "<int>", "reason": "<brief explanation>"}"""

Cloud: OpenAI text-embedding-3-small + GPT-4o-mini
Local: e5-small-v2 + Llama 3 (3B)
Checker: qwen2.5:7b

	Cloud	Local
Perfect (5)	58	33
Good (4)	153	136
Acceptable (3)	21	51
Partial (2)	21	32
Poor (1)	8	9
Wrong	3	3

Avg score: 0.7462121212121212

verdict

correct	164
partially_correct	66
not_in_context	27
incorrect	7

Total questions: 264
Score 0.0: 36 (13.64%)
Score 0.5: 107 (40.53%)
Score 1.0: 121 (45.83%)
Average score: 0.6610
Min score: 0.0
Max score: 1.0
Average score: 0.6609848484848485

Evaluating the scores and responses:

I think it does a pretty good job at finding the right answer. I noticed that it gave out a lot of 1's and 0's.

132: How many units should a student take to be considered a full-time graduate student?

Ground truth: 9 units of coursework with class number of 500-999 or enrollment in one of the following classes for the respective program of study: 799A (Thesis), 897 (Doctoral Research), 899 (Dissertation), 894 (Clinical Research)

****

LLM: To be considered a full-time graduate student, a student should take nine units of coursework numbered 500 through 999.

Verdict: partially_correct

Score (0 - 5): 3

Missing Points: ['Inclusion of enrollment in specific classes such as 799A (Thesis), 897 (Doctoral Research), 899 (Dissertation), and 894 (Clinical Research) as qualifying for full-time status.']

Incorrect Points: []

Notes: The model correctly states that nine units of coursework numbered 500 through 999 are required for full-time graduate student status. However, it omits the important detail that enrollment in certain classes (799A, 897, 899, 894) also counts as full-time enrollment for respective graduate programs. This omission makes the answer incomplete.

```
def rephrase_query(original_query):
    rephrase_prompt = (
        "You are an expert at optimizing search queries for an academic vector database. "
        "The following user query failed to return an answer: '{query}'\n\n"
        "Rephrase this query to be more descriptive, formal, and likely to match text "
        "found in a University Graduate Bulletin. Do not answer the question, "
        "just provide the rephrased text."
    )

    prompt_template = ChatPromptTemplate.from_messages([("system", rephrase_prompt)])
    chain = prompt_template | llm

    response = chain.invoke({"query": original_query})
    return response.content.strip()
```

Original Query: Who is considered an international student at SDSU, and what are the basic admission requirements for international graduate or post-baccalaureate applicants?

New Query: Please provide detailed information regarding the definition of an international student at San Diego State University (SDSU), including the criteria used to classify students as international. Additionally, outline the fundamental admission requirements and application procedures for international applicants seeking graduate or post-baccalaureate admission to SDSU.

Original Query: I’m a grad student. What GPA do I need to avoid academic probation?

New Query: Graduate students seeking information regarding the minimum cumulative GPA requirements necessary to maintain good academic standing and avoid probationary status as outlined in the university's academic policies.

Original Ans: You need to maintain a minimum 3.0 grade point average to avoid academic probation.

New Answer: Graduate students must maintain a cumulative grade point average of at least 2.85 in all units attempted subsequent to admission to the university to avoid academic probation.

Testing Question: How long do I have to complete my master’s degree?

Original Query: How long do I have to complete my master’s degree?

New Query: What is the maximum allowable duration for completing a master's degree program according to university policies?

Original Ans: You have seven years from the semester you entered the M.S. program to complete all degree requirements.

New Answer: According to university policies, all requirements for a master's degree must be completed within seven years after initial registration in the courses used toward the degree.

Expected: All requirements for advanced certificates and master's degrees coursework must be completed within six years after initial registration in course(s) used towards the completion of degree requirements.

```
f"""
You are answering questions about SDSU using ONLY the provided sources.

Rules:
- You must base your answer ONLY on the provided passages.
- Do NOT use outside knowledge.
- Do NOT infer or assume missing details.
- Keep the answer to 1–2 sentences.
- After the answer, list the supporting pdf_page numbers.\

Question:
{query}

Sources:
{context}

Return:
1) A short answer (1–2 sentences).
2) Cite pdf_page numbers.
""".strip()
```

Run Queries & Evaluate the Returned Chunk for Relevance

QUERY 1: How many 600+ level CS classes are available?

	Rank	Cosine Similarity	PDF Page	Text (Preview)
0	1	0.60	168	30 units, of which at least 15 units must be i...
1	2	0.59	168	COMP 605/\nCS 605 Scientific Computing **.** ....
		0.58	174	Prerequisite: Computer Science 320 or Linguist...
3	4	0.56	173	CS 503 **. Scientific Database Techniques (3)*...
4	5	0.56	173	Computer Science\n\n\nThe certificate requires...

Collapse Output

QUERY 2: Who are the graduate advisers for Computer Science?

	Rank	Cosine Similarity	PDF Page	Text (Preview)
0	1	0.61	172	**OFFICE** : Geology/Mathematics/Computer Scie...

read from the QuestionsAnswers.csv and generate 5 top results from each Question

From there, we I would send them into an LLM (gpt-4o-mini for this one) and generate a response based on only the information that was provided via the top 8 results.

from there, I i have another llm call that has the question given, the ground truth, and the prediciton and then just the score based on 0,.5,1.

from each question, I can view and make a table of the results and see what was good, what was bad and see how it all went.

Using the same LLM (gpt-4o-mini) for both generation and scoring introduces a bias: the model tends to rate its own output favourably. A more serious evaluation would use a separate, stronger model as the judge, or rely on human annotation.

Out of 25 questions, 6 were Correct, 6 were Partially Correct, and 13 were Incorrect.

```
queries = [
    "How many 600+ level CS classes are available?",
    "Who are the graduate advisers for Computer Science?",
    "What ART section number if Graphic Communication?",
    "What classes will discuss Earthquake magnitude?",
    "What is the lowest GPA needed to get admission into a Master of Science Degree in Exercise Physiology?",
    "What public health class is section 663?",
    "What is the maximum number of Open University or SDSU Global Campus units that can be applie toward a master’s degree before admission?",
    "What graduate standing is required to be appointed as a Graduate Teaching Associate (TA)?",
    "What is the last day to add or drop classes or change grading basis for Spring Semester 2021?",
    "Which office oversees SDSU’s nondiscrimination and nonharassment policy?",
    "How many hours per week may graduate students work as Student Assistants?",
    "What independent doctoral degrees are offered by San Diego State University?",
    "What happens if a student doesn’t pay the university?",
    "What are the unit requirements for a Master of Business Administration Degree for Executives?",
```

Read the file, don’t copy it by hand

You are to score a MODEL_ANSWER compared to the GROUND_TRUTH on a scale of 0.0 to 1.0.

Paraphrasing is okay and extra information is okay as long as it does not contradict GROUND_TRUTH.

Scoring rubric:
- 1.0 = All key facts are correct and complete
- 0.9 = Correct, tiny omission/wording issue
- 0.8 = Mostly correct, minor missing detail
- 0.7 = Mostly correct, but several minor misses
- 0.6 = Mixed, some correct, some wrong
- 0.5 = Some key facts are correct but others are missing or incorrect, about halfway correct
- 0.4 = Mostly wrong with a small correct piece
- 0.3 = Largely wrong with a lot missing
- 0.2 = Barely related
- 0.1 = Almost entirely wrong
- 0.0 = Completely incorrect or contradicts GROUND_TRUTH

Return in the following format only:

Score: <the score>

Ground Truth: The last day to add/drop classes or change grading basis for Spring Semester 2021 was February 2, 2021.

Model Answer: Feb. 2

Correct?: False

```
def AskRag(question):
    db = Chroma(
        persist_directory=CHROMA_PATH,
        embedding_function=OpenAIEmbeddings()
    )

    llm = ChatOpenAI(model_name="gpt-4o-mini", temperature=0)
    query_text = question
    results = db.similarity_search_with_score(query_text, k=3)

    if not results:
        print("No matches found.")

context_text = "\n\n---\n\n".join([doc.page_content for doc, _score in results])

prompt_template = ChatPromptTemplate.from_template(
    """
    Answer the question based only on the following context:
    {context}
    ---
    Question: {question}
    """
)
prompt = prompt_template.format(context=context_text, question=query_text)
response = llm.invoke(prompt)
return response.content
```

```
os.environ["OPENAI_API_KEY"] = os.getenv("OPENAI_API_KEY")
```

```
results = []

for _, row in df.iterrows():
    query = str(row[QUESTION_COL])
    true = str(row[IDEAL_COL])



    ai_response, pages, _ = rag_from_query(query, k=5)

    print("\nAI ANSWER:")
    print(ai_response)



evaluation_prompt = (
    f"User Query: {query}\n\n"
    f"AI Response:\n{ai_response}\n\n"
    f"True Response:\n{true}\n\n"
    "return a score that is only a number between 0.00 and 1.0."
)
```

```
results = []

for _, row in df.iterrows():
    query = str(row[QUESTION_COL])
    true = str(row[IDEAL_COL])

    ai_response, pages, _ = rag_from_query(query, k=5)

    print("\nAI ANSWER:")
    print(ai_response)

evaluation_prompt = (
    f"User Query: {query}\n\n"
    f"AI Response:\n{ai_response}\n\n"
    f"True Response:\n{true}\n\n"
    "return a score that is only a number between 0.00 and 1.0."
)
```

QUERY: How many 600+ level CS classes are available?

Result 1 | distance: 0.7811 | page: 174
CS 696. Selected Topics in Computer Science (3)
Prerequisite: Consent of instructor.
Intensive study in specific areas of computer science. May be repeated with new content. See Class Schedule for specific con - tent. Credit for 596 and 696 applicable to a master’s degree with approval of the graduate adviser.
CS 705. Advanced Parallel Computing (3)
(Same course as Computational Science 705)
Prerequisite: Computer Science 605 [or Computational Science 605].
Libraries, numerical methodology, optimization tools, visual - ization of results, MPI and GPU computing models. Applications conducted on CSRC student cluster and NSF XSEDE computing resources.
CS 605. Scientific Computing (3)
(Same course as Computational Science 605)

Result 2 | distance: 0.7904 | page: 167
30 units, of which at least 15 units must be in 600- and 700-level courses excluding 799A to include:
Required core courses (15 units):
  CS 503 Scientific Database Techniques 3
  OR
  CS 514 Database Theory and
   Implementation 3
  OR
  COMP 607 Computational Database
   Fundamentals 3
  COMP 526 Computational Methods
   for Scientists 3
  COMP 536 Computational Modeling
   for Scientists 3
  OR
  MATH 636 Mathematical Modeling 3
  COMP 605/

Result 3 | distance: 0.8295 | page: 171
600-level electrical engineering course or one of the math- ematics courses listed below in the ALC area of study may replace one 600-level computer science course.

b. Students select two areas of study from the areas listed

Result 3 | distance: 0.8295 | page: 171
600-level electrical engineering course or one of the mathematics courses listed below in the ALC area of study may replace one 600-level computer science course.
b. Students select two areas of study from the areas listed below, and take at least two courses from each area.
c. Three units of Thesis (799A), and an oral presentation and defense.

Result 4 | distance: 0.8639 | page: 174
CS 666. Advanced Distributed Systems (3)
Prerequisite: Computer Science 570.
Design of distributed systems including abstract models, algorithms, and case studies of real-world systems. Group research related to distributed systems.
CS 682. Speech Processing (3)
Prerequisites: Graduate standing, Computer Science 310, Mathematics 254, Statistics 551A.
Algorithms and methods for processing of speech. Feature extraction, human speech production and perception, pattern recognition for acoustic and language modeling as applied to automatic speech and speaker recognition.
CS 696. Selected Topics in Computer Science (3)
Prerequisite: Consent of instructor.
Intensive study in specific areas of computer science. May be

Result 5 | distance: 0.8699 | page: 170
tent. Credit for 596 and 696 applicable to a master’s degree with approval of the graduate adviser.
COMP 705. Advanced Parallel Computing (3)
(Same course as Computer Science 705)
Prerequisite: Computational Science 605 [or Computer Science 605].
Libraries, numerical methodology, optimization tools, visualization of results, MPI and GPU computing models. Applications conducted on CSRC student cluster and NSF XSEDE computing resources.
COMP 797. Research (1–3) Cr/NC/RP
Prerequisite: Six units of graduate level computational science courses.
Research in computational science. Maximum credit six units applicable to a master’s degree.

AI ANSWER:
There are 5 600+ level CS classes available:

1. CS 696. Selected Topics in Computer Science
2. CS 705. Advanced Parallel Computing
3. CS 605. Scientific Computing
4. CS 666. Advanced Distributed Systems
5. CS 682. Speech Processing

EVALUATION SCORE:
0.00

Database of documents for RAG (Part 2)

```
vector_store.add_documents(chunks) # add chunks t
```

```
4]: ['86ffa1f7-482a-4acf-af8e-ee36e487f5eb',
     '3437cbb3-741f-4780-b57a-56a4af27e03f',
     '8c03d5fc-1149-49c8-a26a-825976229993',
     '5bffd1d4-d2d2-4c88-9afc-b75ac514a699',
     '97aac9b3-20c9-4bb7-8ed0-c55c6e98ad95',
     '369f9459-c379-4d68-a947-65b441d297fe',
     '9e7594b2-13fb-460e-b067-fc68374c1f9c',
     '50aabb3a-e483-4bd4-a5af-8657f45b23b9',
     'ee15050a-0ff6-4d87-a9cd-67c92dd151e6',
     'e62969b1-155f-44ad-9bfb-548f70929fb2',
     '01155c00-5ec4-400d-b8fd-018e8d70fe70',
     'c75818f7-726a-4526-83f0-ab49ed3c682f',
     '81534708-8b7f-4eb7-b56e-97841368ae14',
     '270f9cff-09a8-4b2f-bac4-ca42de2e8fa4',
     '09dd4d2d-8da4-4f6e-9c67-e8b0a489c19d',
     '10439bf5-a7ac-44d3-a934-07f4c95413be',
     '7ca788ed-da6f-4685-bda0-754af6c38883',
     'a19e42b2-d431-4e1d-8824-02d03ee2e5d7',
     '0a4ff42a-376d-44b9-b109-9aca44059334',
     '4356ce06-a354-4b33-93d9-1542d045572b',
     '626b54b4-9f35-4d32-8371-4a3343449dba',
     '2b884a94-70a9-4c86-a119-a928aba463ae',
     '7f25642c-69c3-4585-b11c-5da4d81cb50e',
     'de9ac29c-b956-477c-8d64-b0534802675a',
     'e5a561b9-8084-409e-a97f-65859861af38',
     '58930ae1-cd86-43c2-bb29-0e8c748c5600',
     'f0db457f-e37c-4cb9-ae84-0f84f292ff2a',
     '185a45e9-35f0-488e-a9ff-f27c69aab565',
     '25485b7d-67ea-49ad-afaa-8b6534bccc54',
     '18446f98-5e27-449a-a50c-57ed98604429',
     '5e629b86-f4b8-44d2-80d9-40d41e8cbde6',
     'e8945a89-f920-40e9-a2fe-32549262b994',
     '2c42e0a2-8914-425d-9206-be1758c81f27',
     '7582b674-9e13-4ee0-8a48-f5252ecdfc06',
     '5e1eabc8-1424-4c28-954b-30fa8e202d0f',
     '459b7b19-bcba-49e1-9e68-3dca00c97f8f',
     '0867cf7b-e8eb-47bd-977f-b0338ece6437',
     'e2a256d2-6c9f-4988-b40a-9e29b063a68e',
     'a130ec40-f2eb-4ef4-ad84-8630c25ca9b9',
```

Part 3 Connecting LLM

49]:
```python
# RAG: Retrieval + LLM Answer
USE_OPENAI = True

def build_context(hits, max_chars=4000):
    parts = []
    total = 0
    for h in hits:
        header = f"[page={h['page']} | section={h.get('section','')} | score={h['score']:
        block = header + "\n" + h["text"].strip()
        if total + len(block) > max_chars:
            break
        parts.append(block)
        total += len(block)
    return "\n\n---\n\n".join(parts)

def rag_answer(query, top_k=5):
    # Retrieve top k chunks using FAISS DB
```

Evaluation Part 3

running this cell takes about 4-6 minutes bc going through all 264 questions.

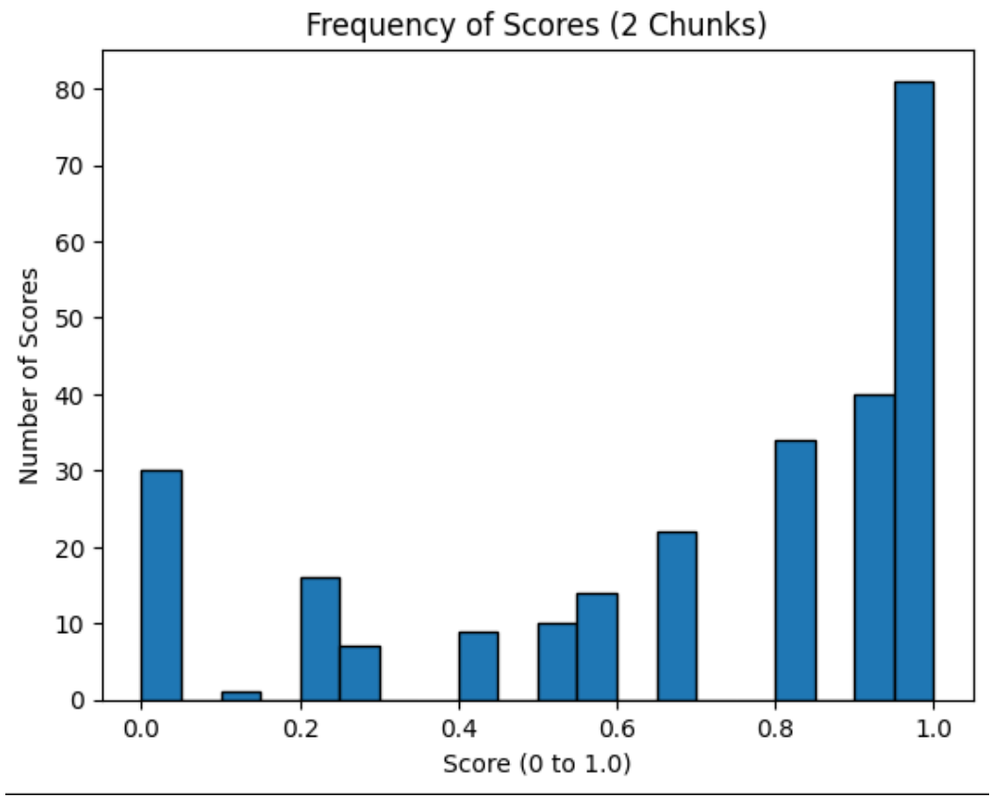
Please answer the question about the SDSU Graduate Bulletin.

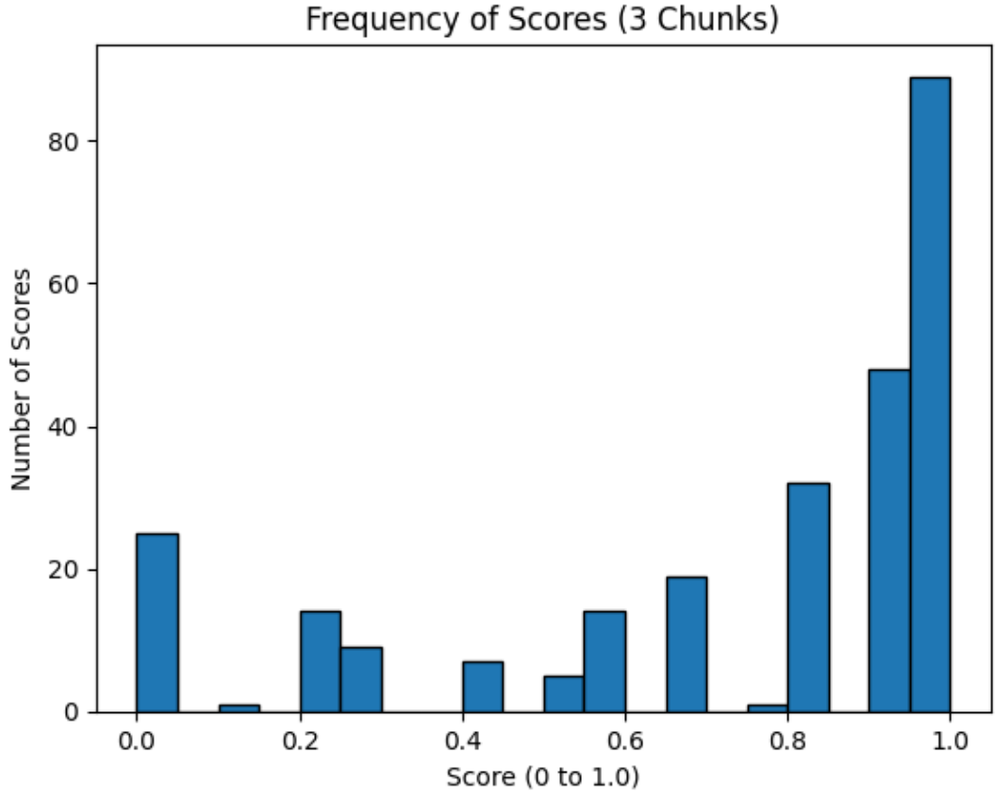
You may ONLY use info from these five chunks retrieved from RAG results to answer the question.

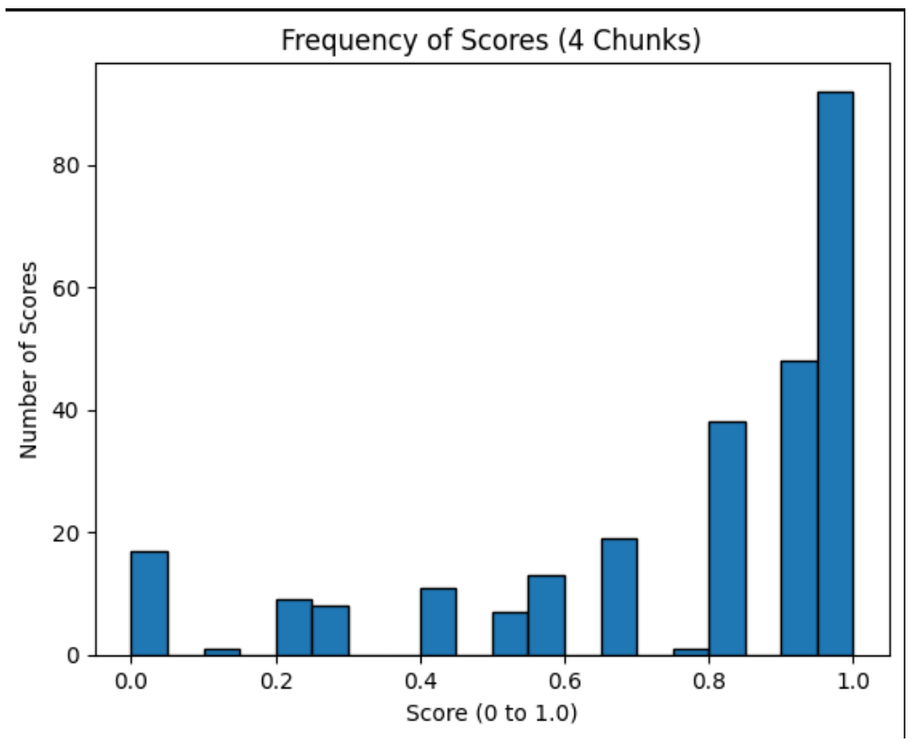
I have also provided their respective retrieval scores.

Rules:
-Only use info from the provided 5 chunks
-Do NOT use outside knowledge
-Keep your answer concise

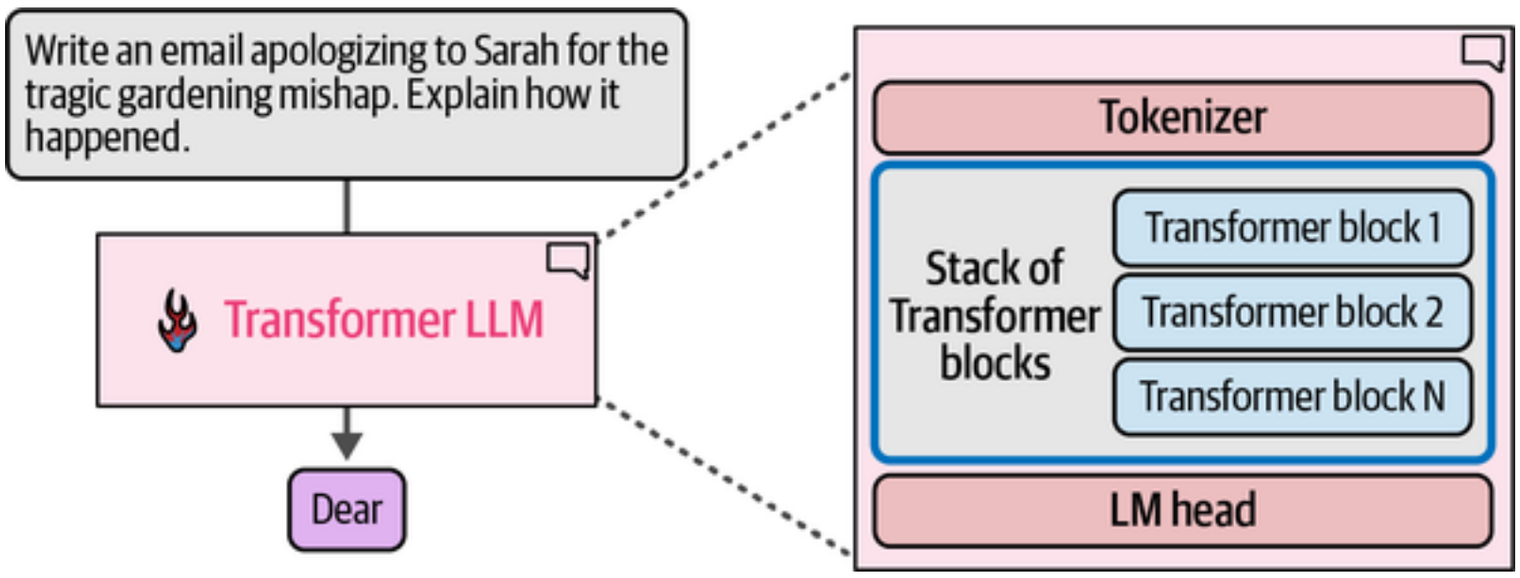
It also doesn't seem to take into account the retrieval scores for each of the chunks. I got this conclusion through the constant usages of "According to Chunk x preceding a large amount of responses".


Frequency of Scores (2 Chunks)
Number of Scores
Score (0 to 1.0)

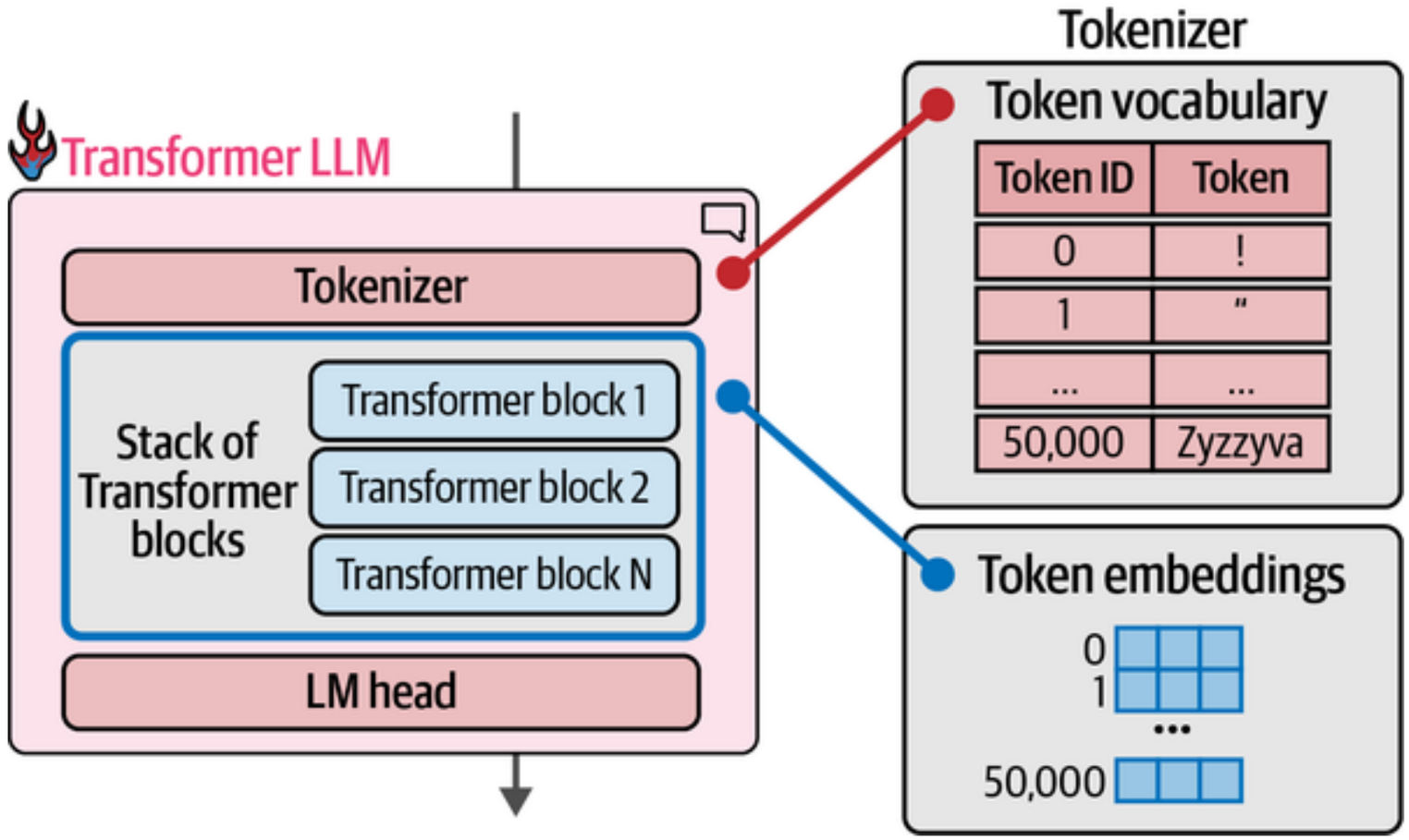

Frequency of Scores (3 Chunks)
Number of Scores
Score (0 to 1.0)


Frequency of Scores (4 Chunks)
Number of Scores
Score (0 to 1.0)

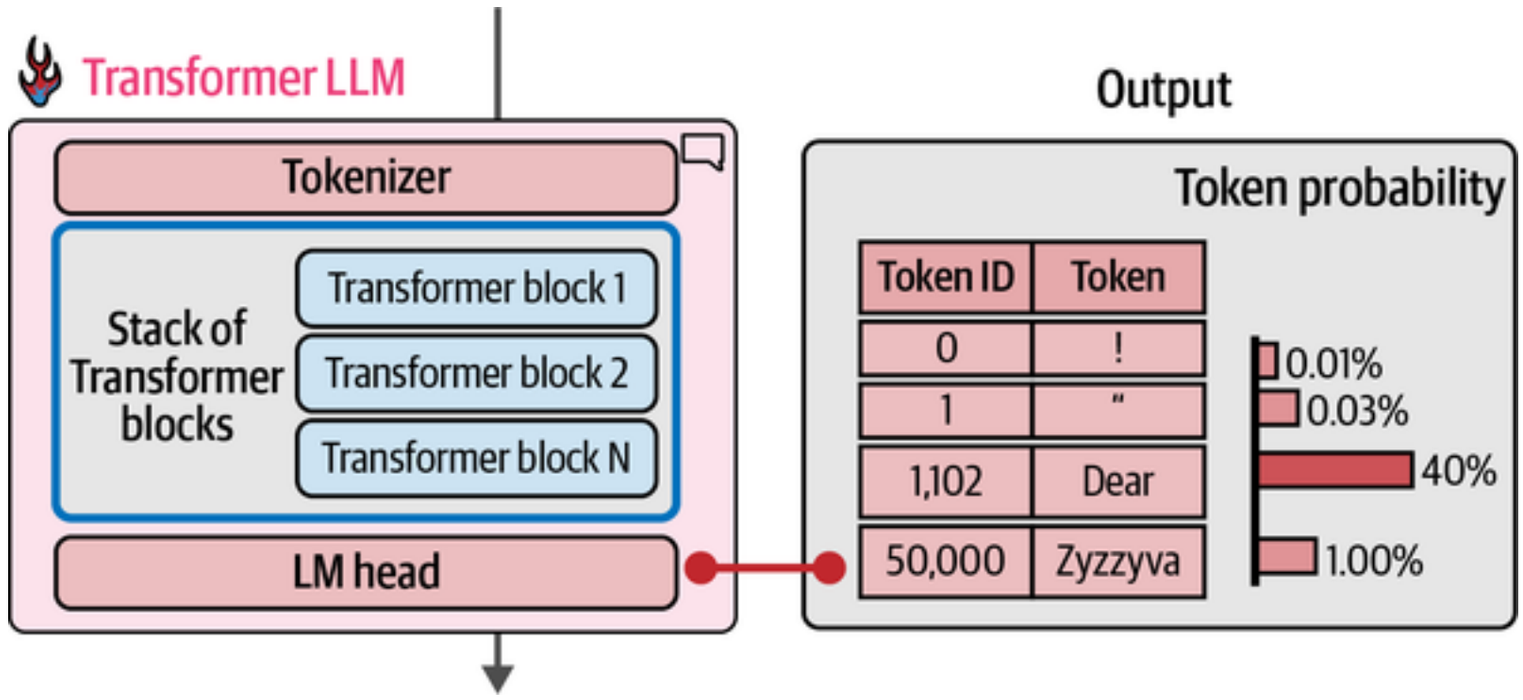
Big Picture - LM Head



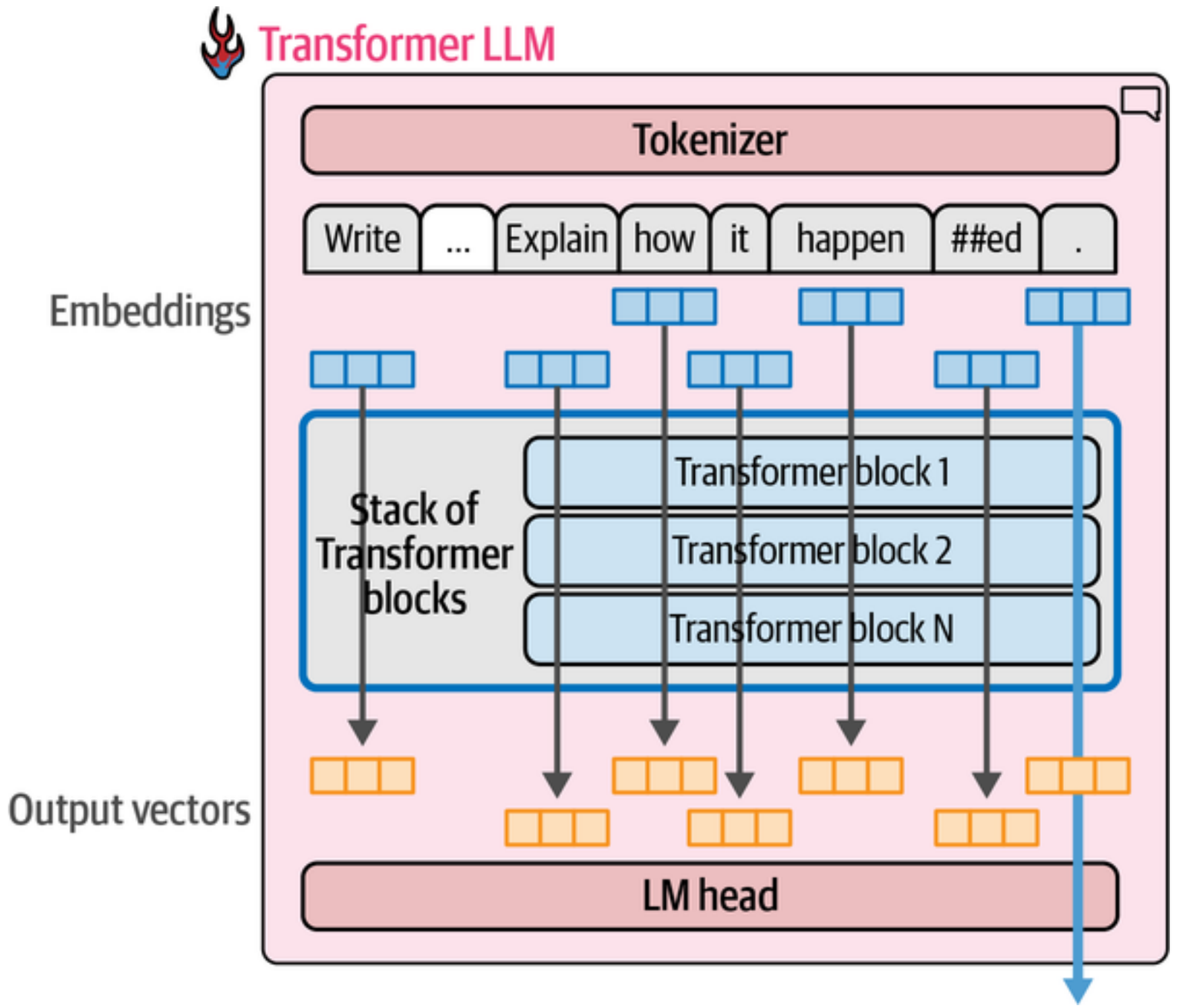
Big Picture - LM Head



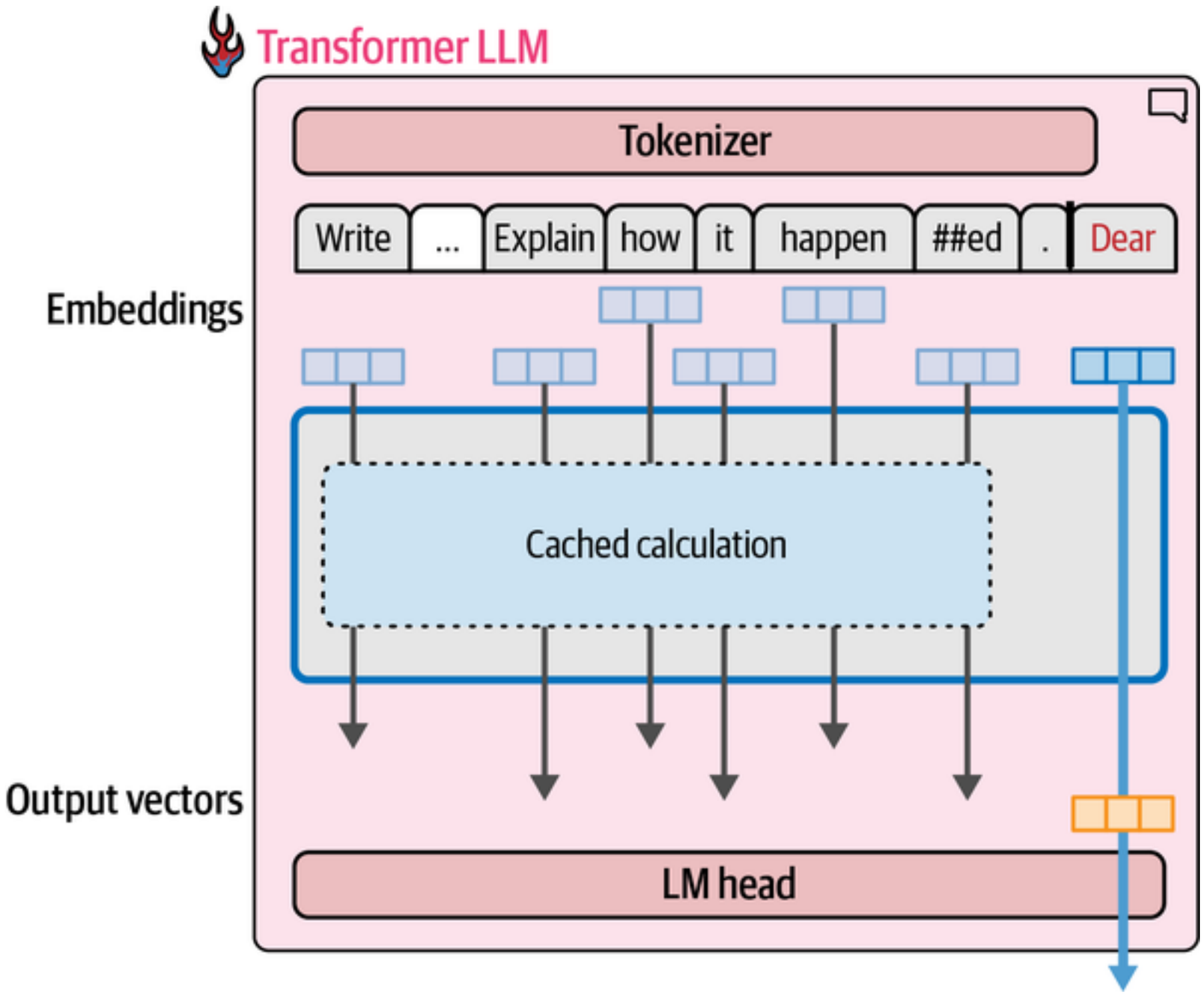
Big Picture - Output

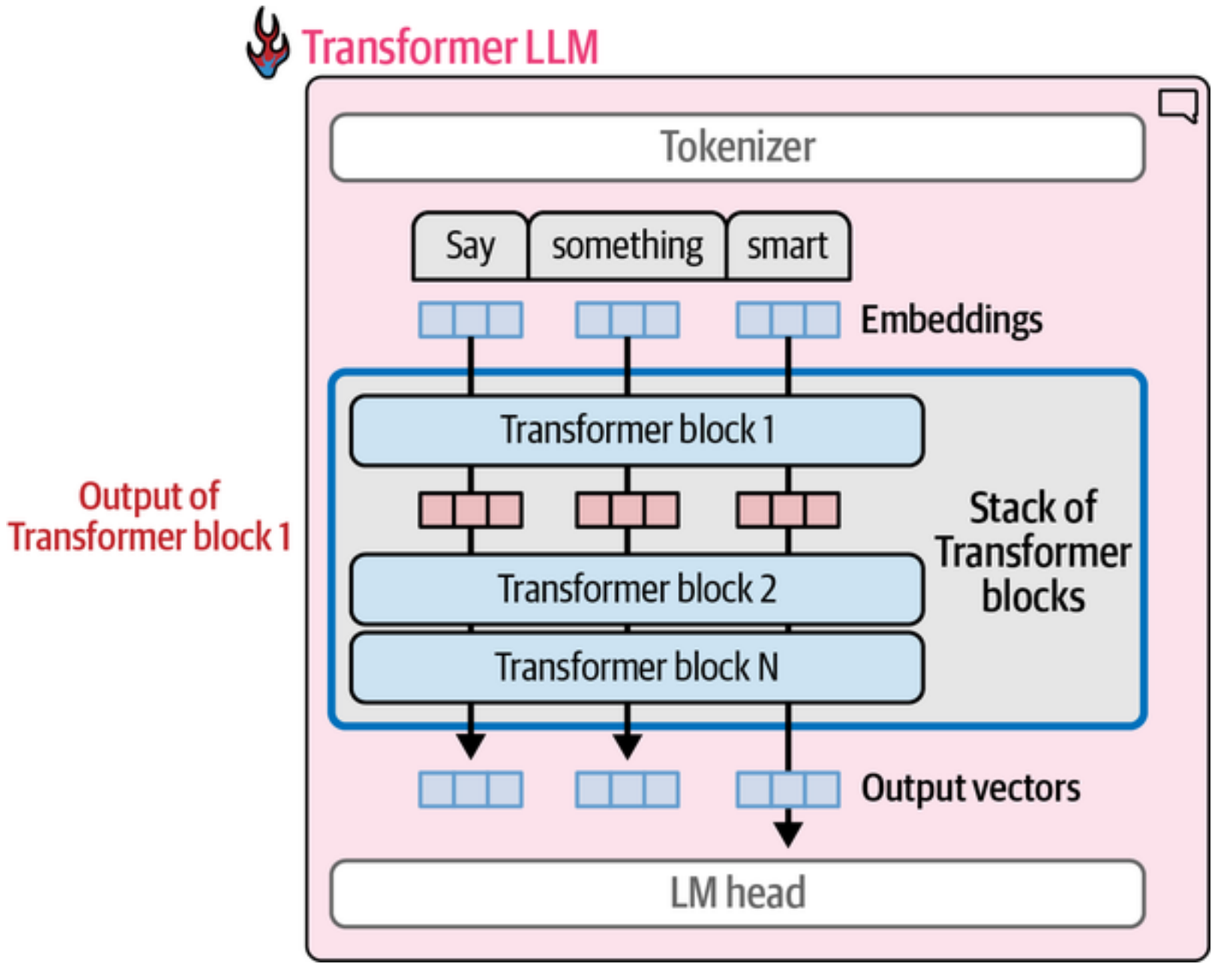


Big Picture - Processing Token in Parallel (sort of)

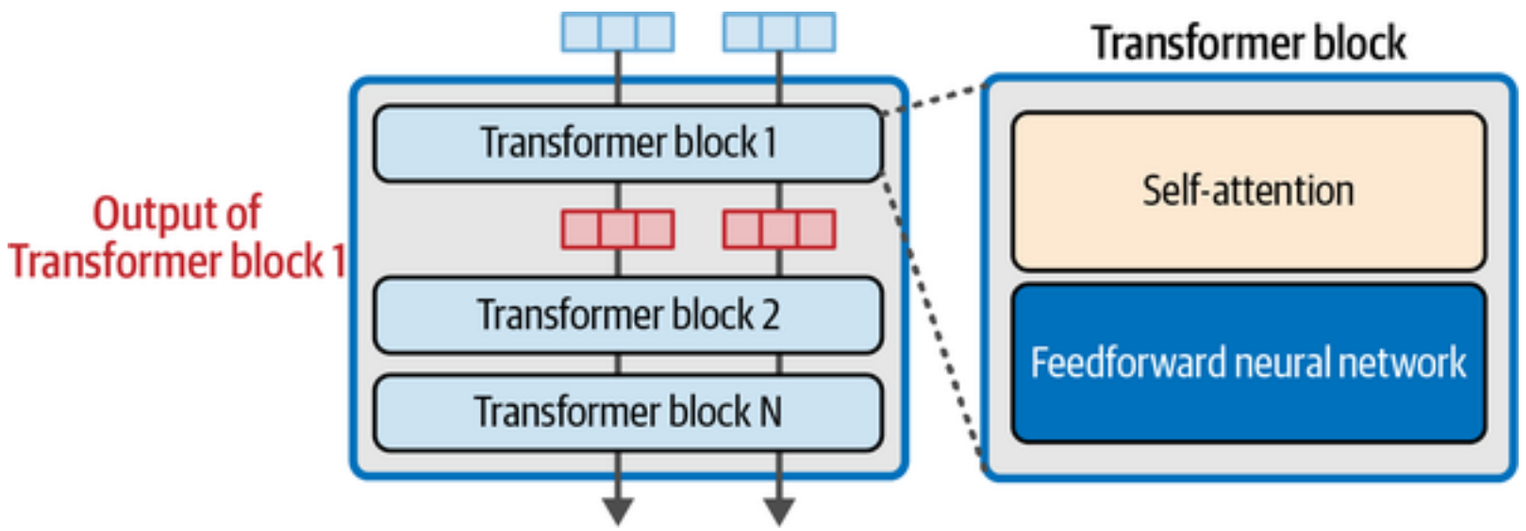


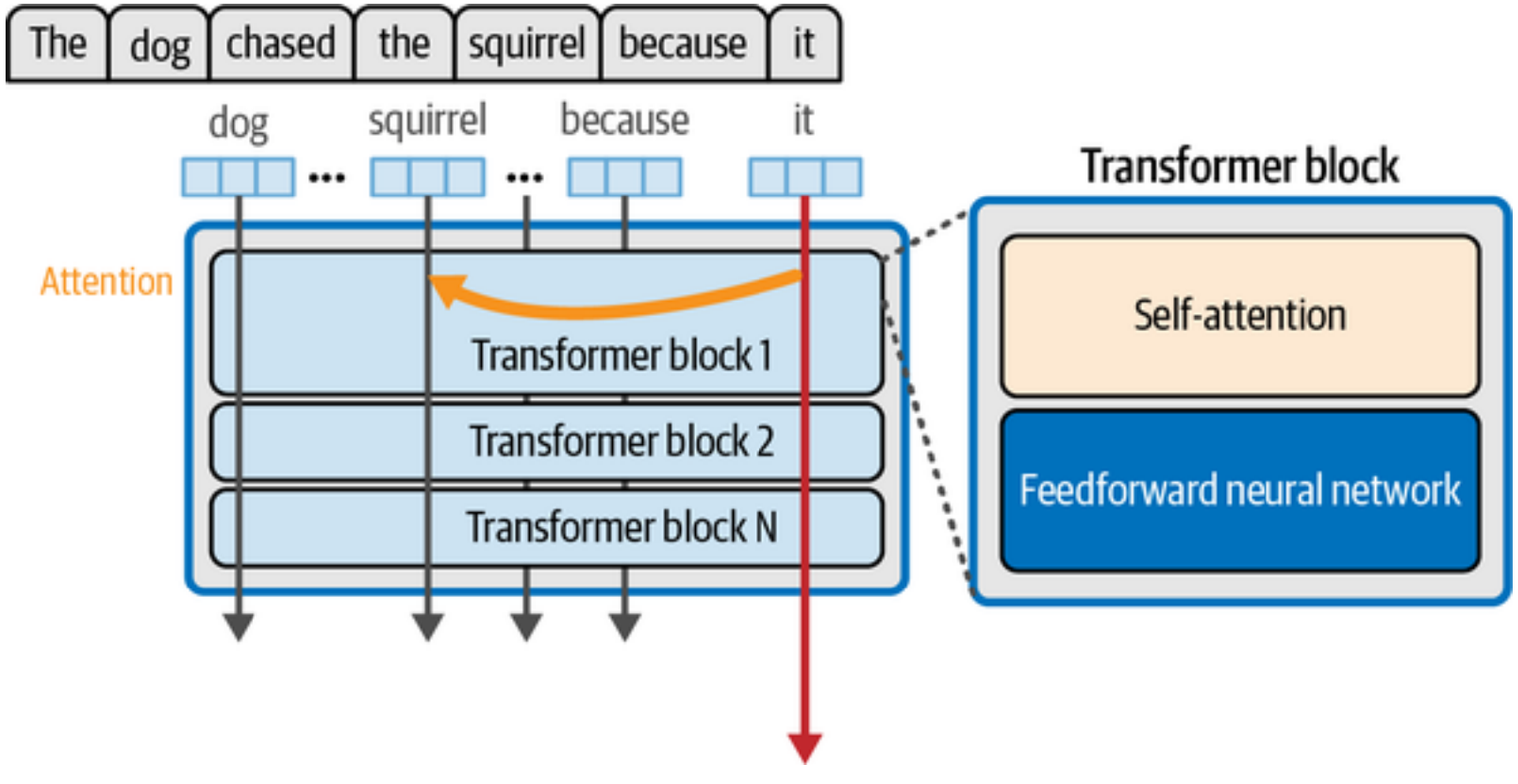
Big Picture - Processing Token in Parallel (sort of)



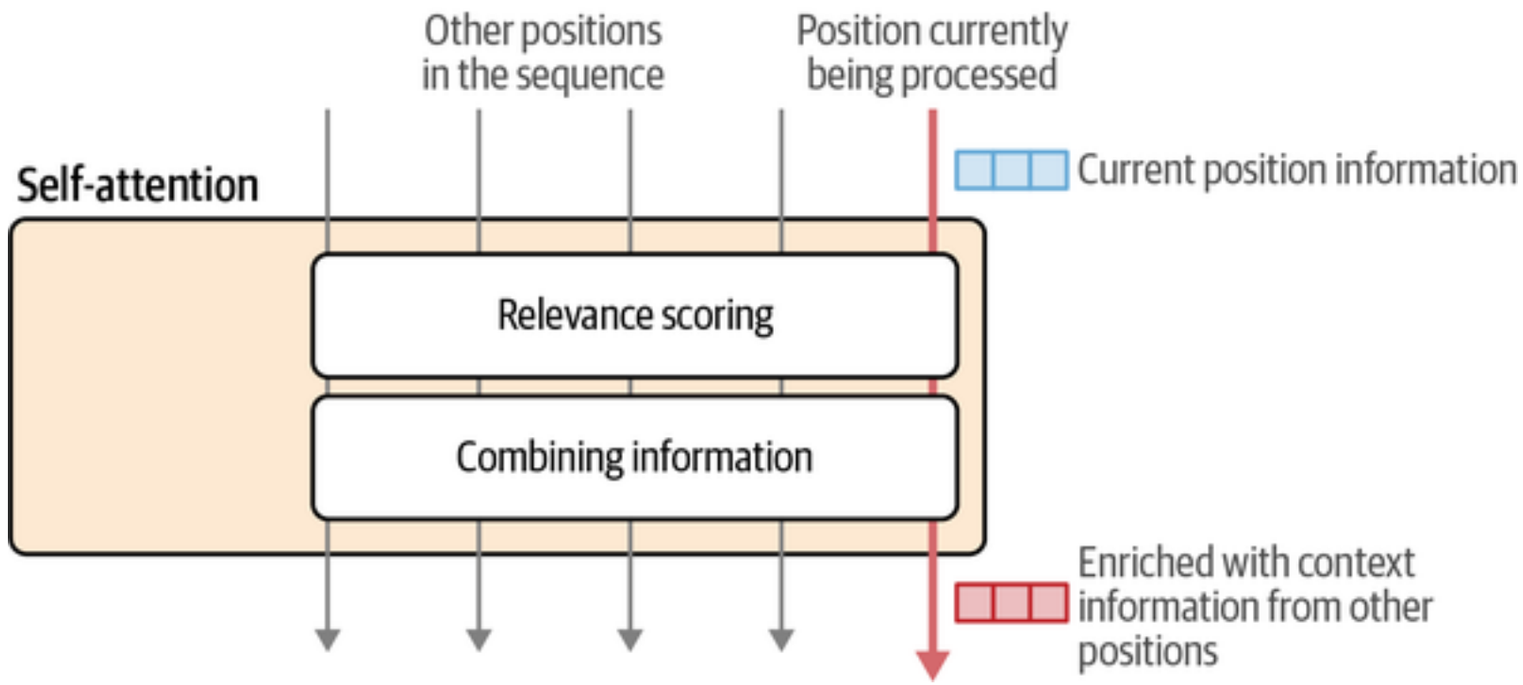

Transformer LLM
Tokenizer
Say
something
smart
Embeddings
Transformer block 1
Transformer block 2
Transformer block N
Stack of Transformer blocks
Output vectors
LM head
Output of Transformer block 1

Block Contents





Attention



Need a way to compute how relevant each previous token is

Combine those computations into output vector

Intuition of attention

Build up the contextual embedding from a word by selectively integrating information from all the neighboring words

We say that a word "attends to" some neighboring words more than others

How attention is calculated

The inputs to the layer are:

The vector representation of the current position or token

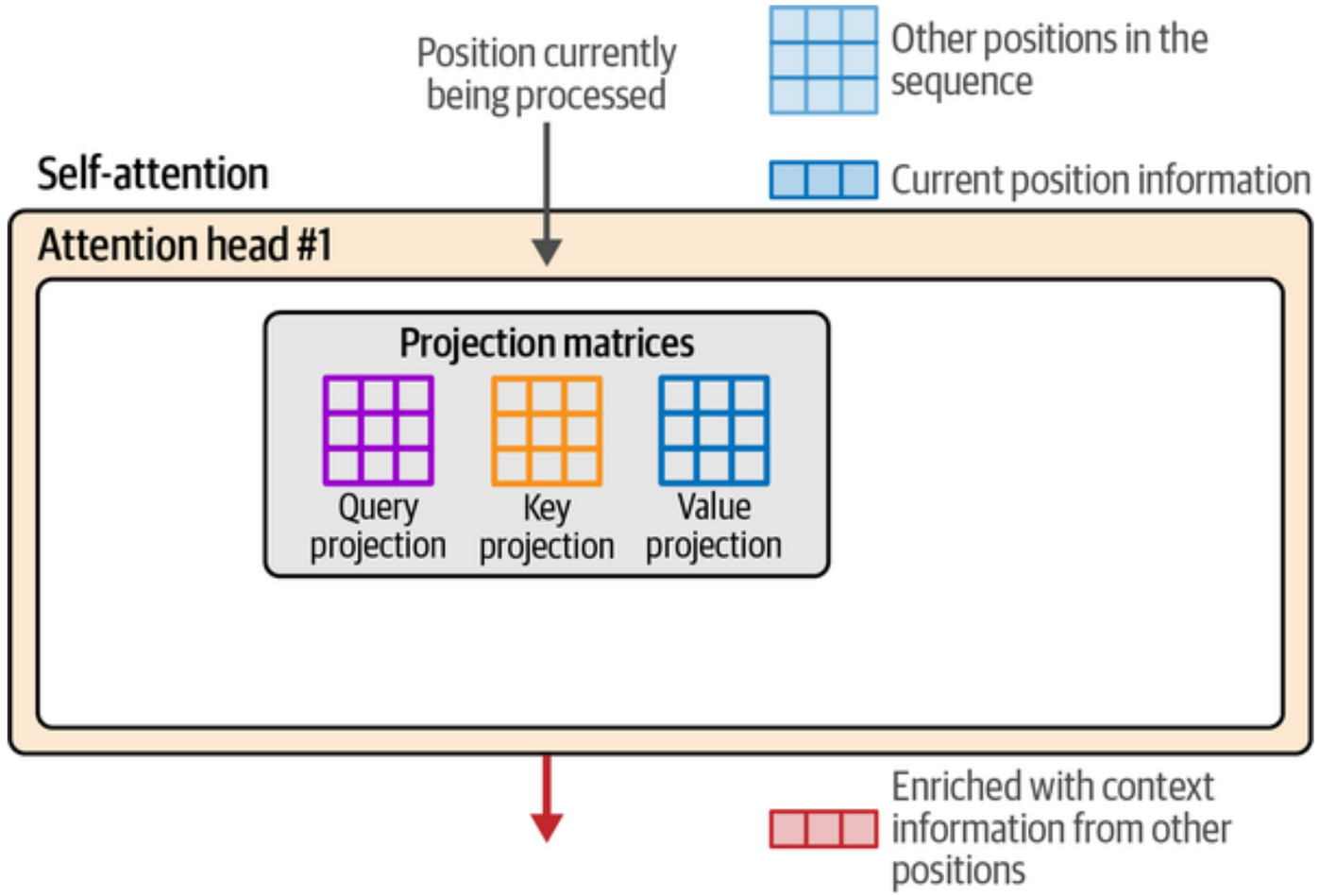
The vector representations of the previous tokens

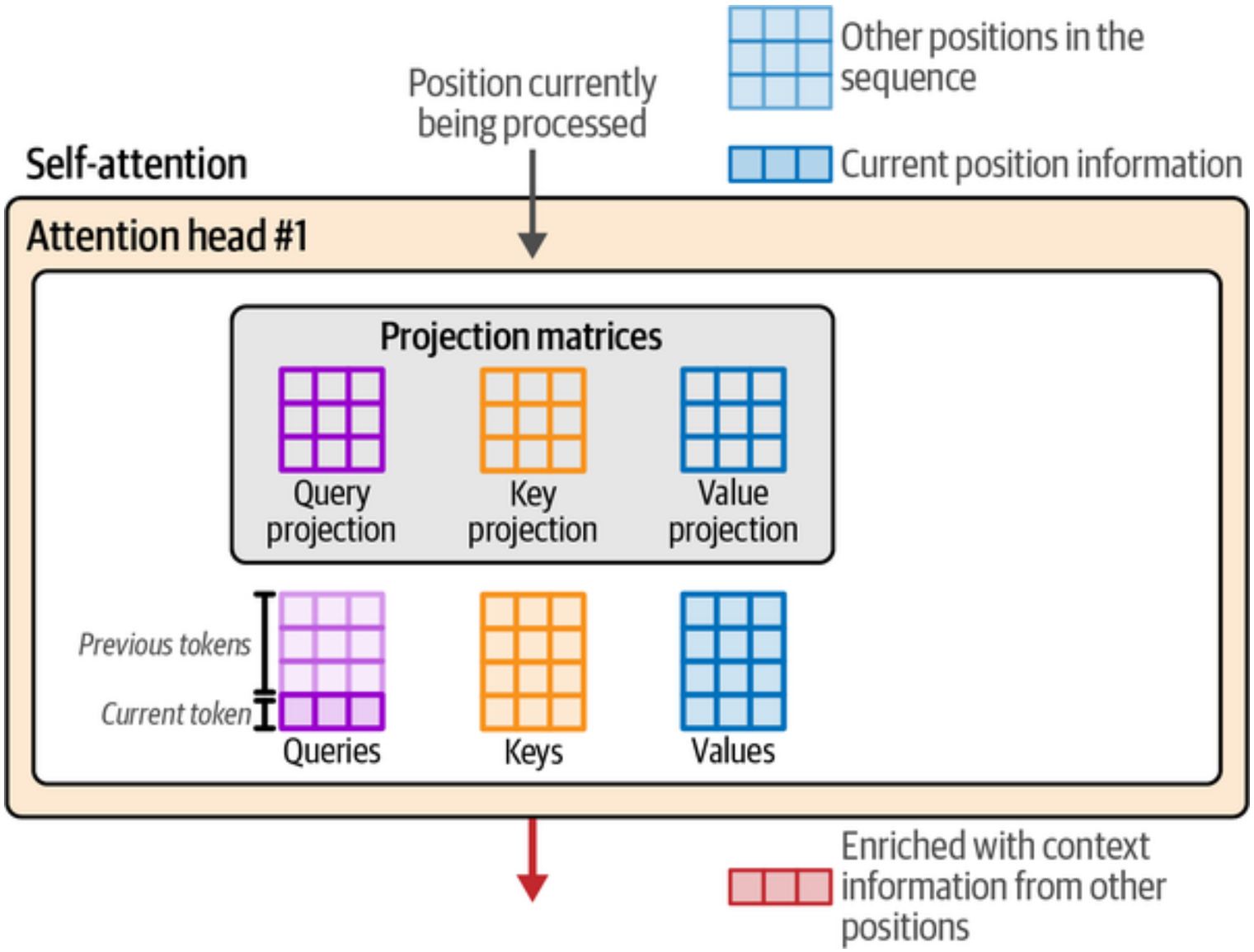
Training Process produces

A query projection matrix

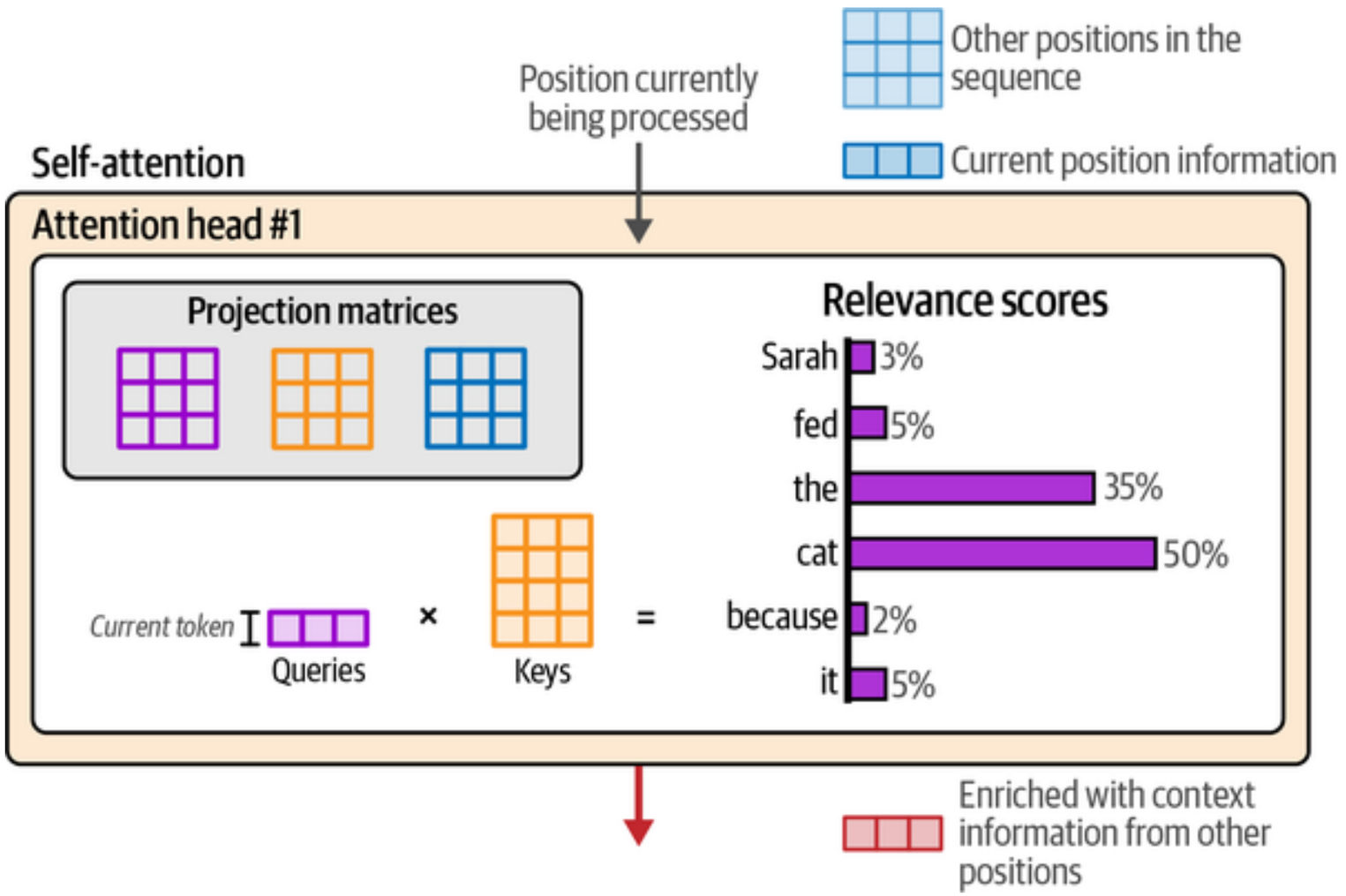
A key projection matrix

A value projection matrix



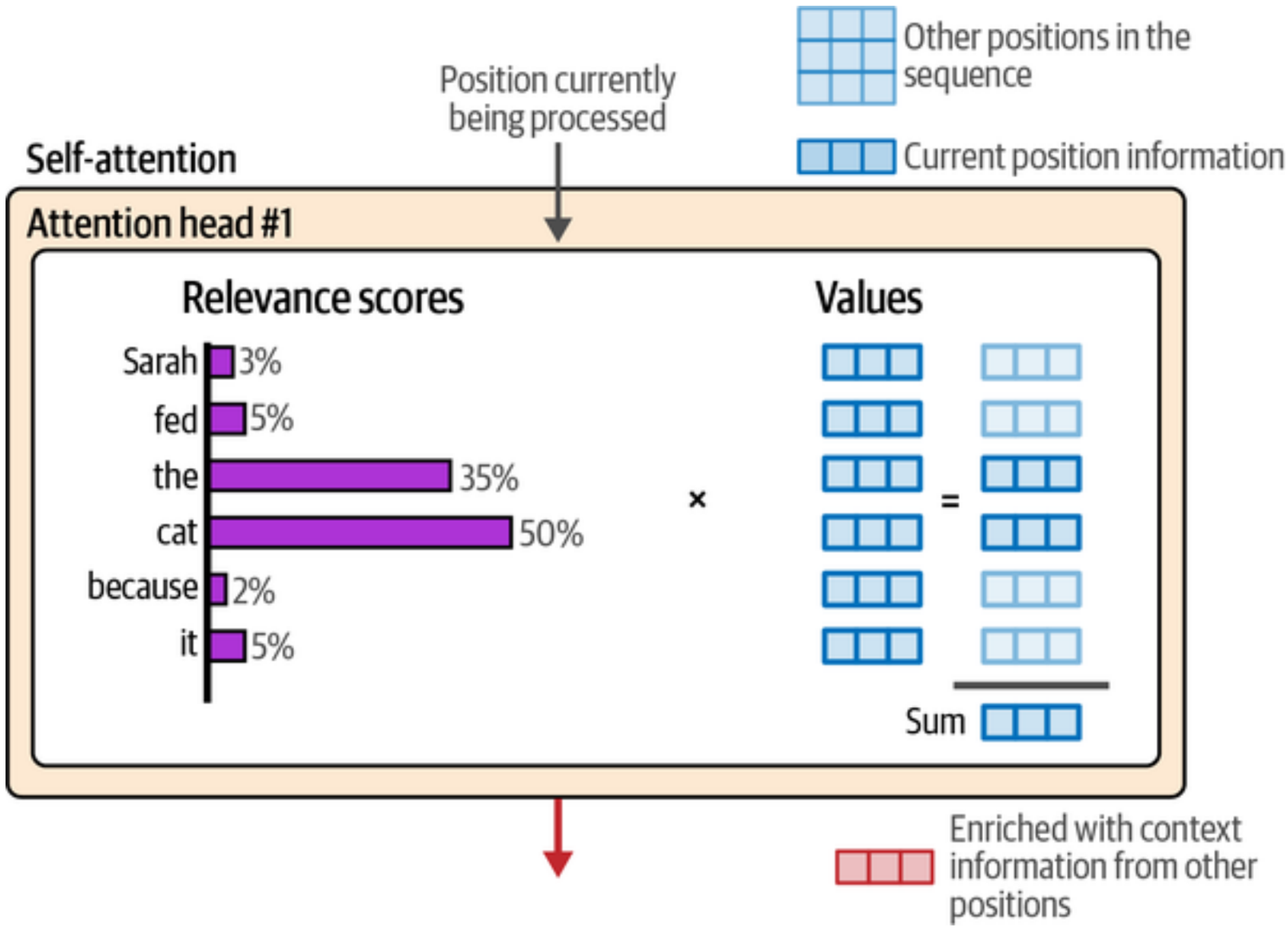


Multiple input by projection matrices



Queries * Keys give Relevance scores

Combining Values



Variations

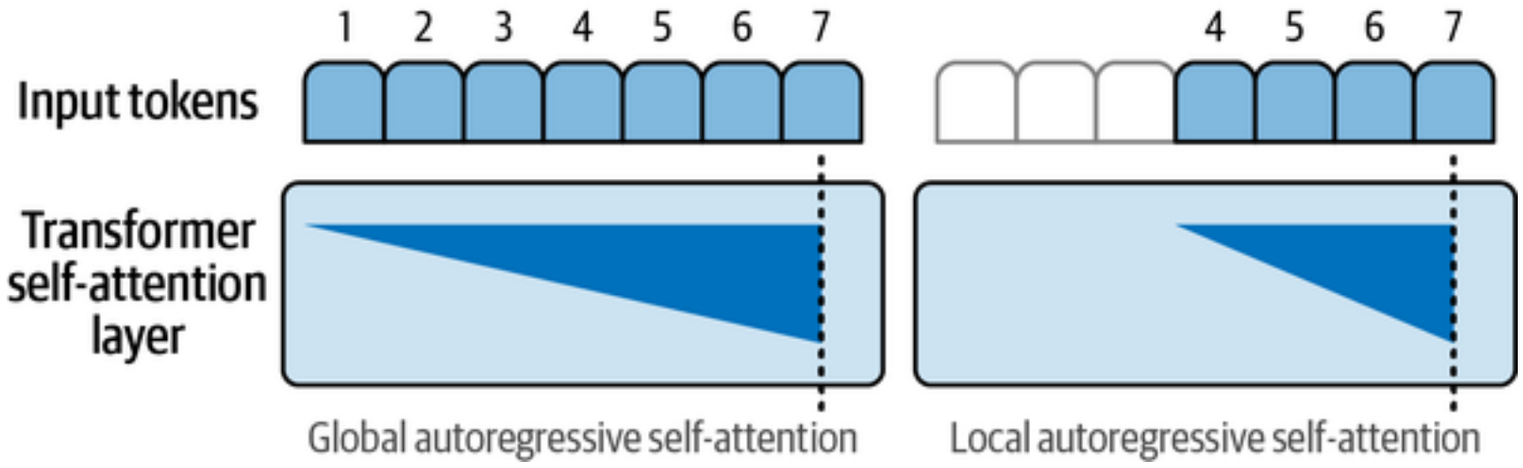
Local Attention

Multi-headed Attention

Grouped-query

Multi-query

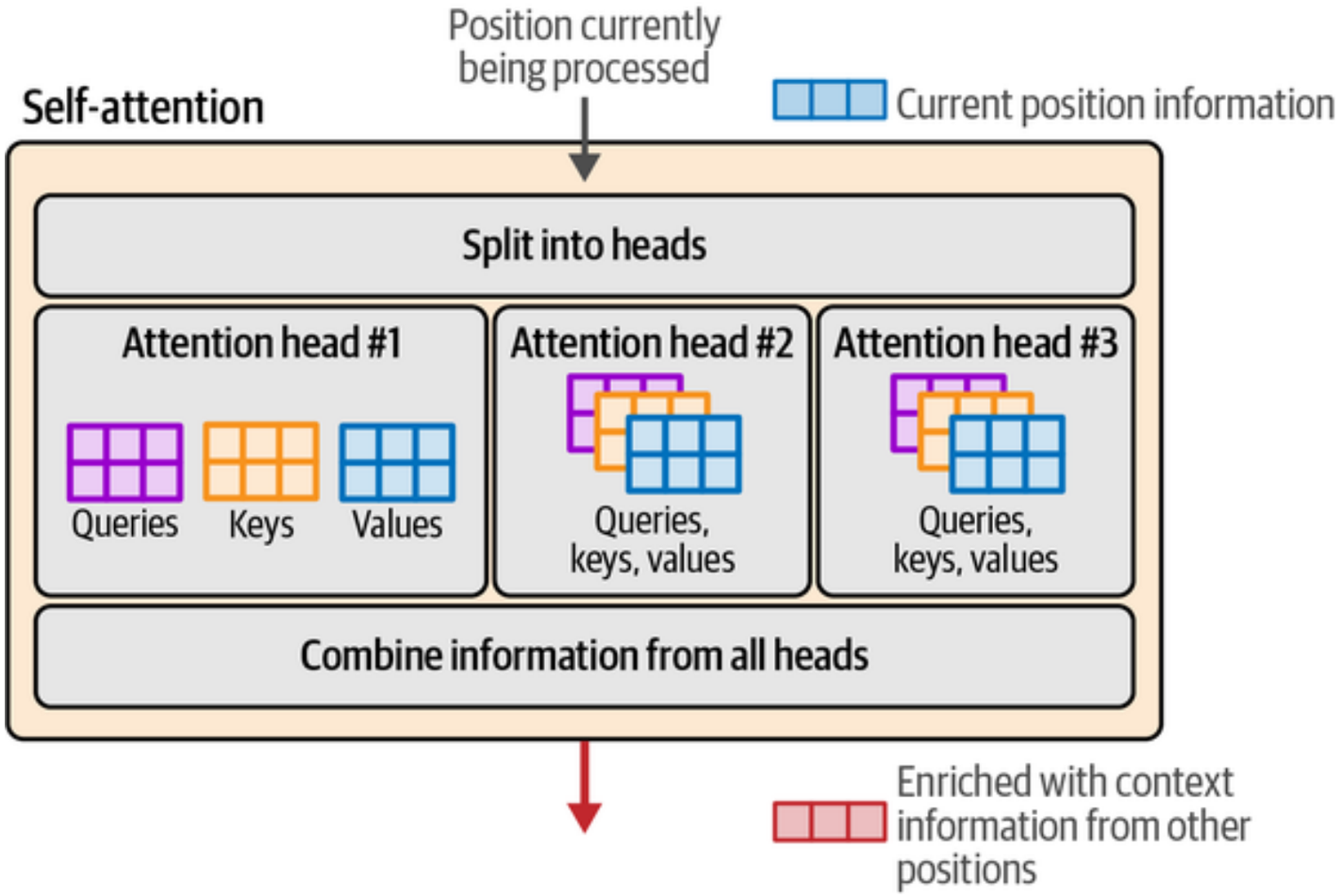
Local Attention



Using multiple heads

- Some heads use local attention

Multiple Heads



Examples

Head 1: Might focus on syntactic relationships

Head 2: Coreference resolution

Head 3: Semantic relationships

Head 4: Long-range dependencies

Number of Heads

Model Size	Attention Heads
8B	32
70B	64
405B	128

Key/Value Heads

- Often fewer
- Llama 3, all versions (8B to 405B) 8 KV heads

Grouped-query