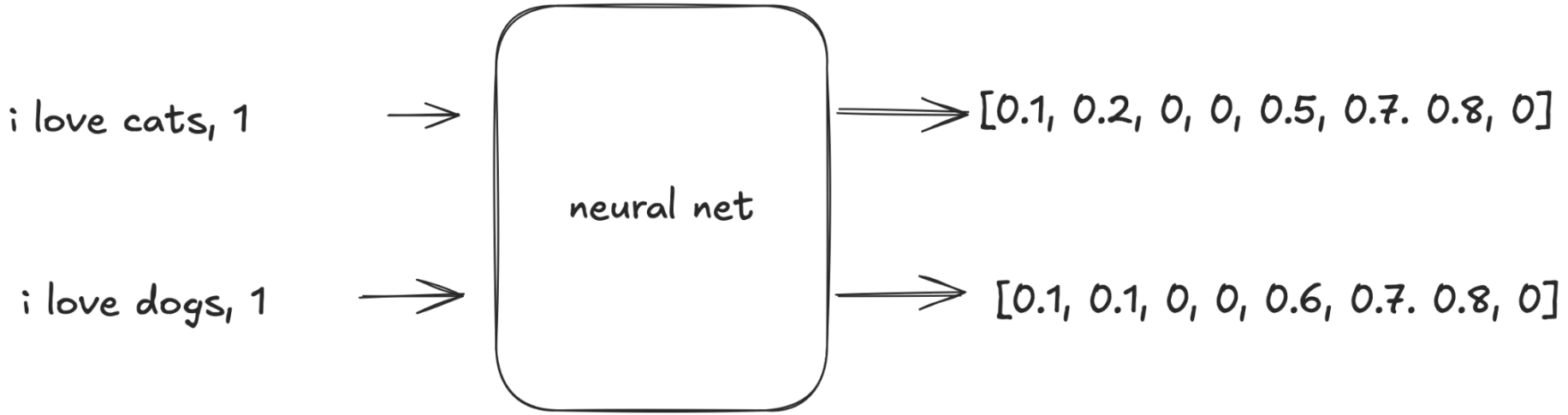
RAG, A2A, Resource-Aware Optimization

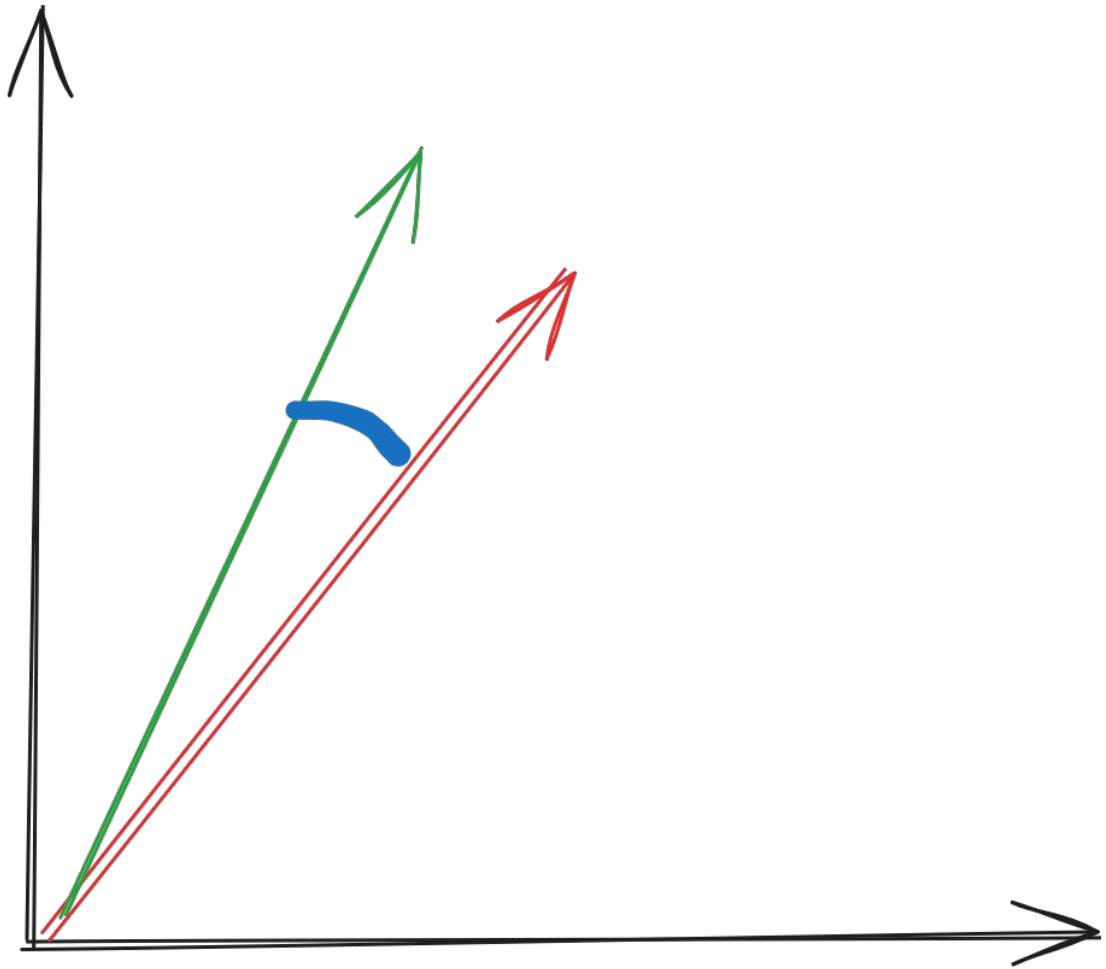
made with ❤️ for “Little ML book club”

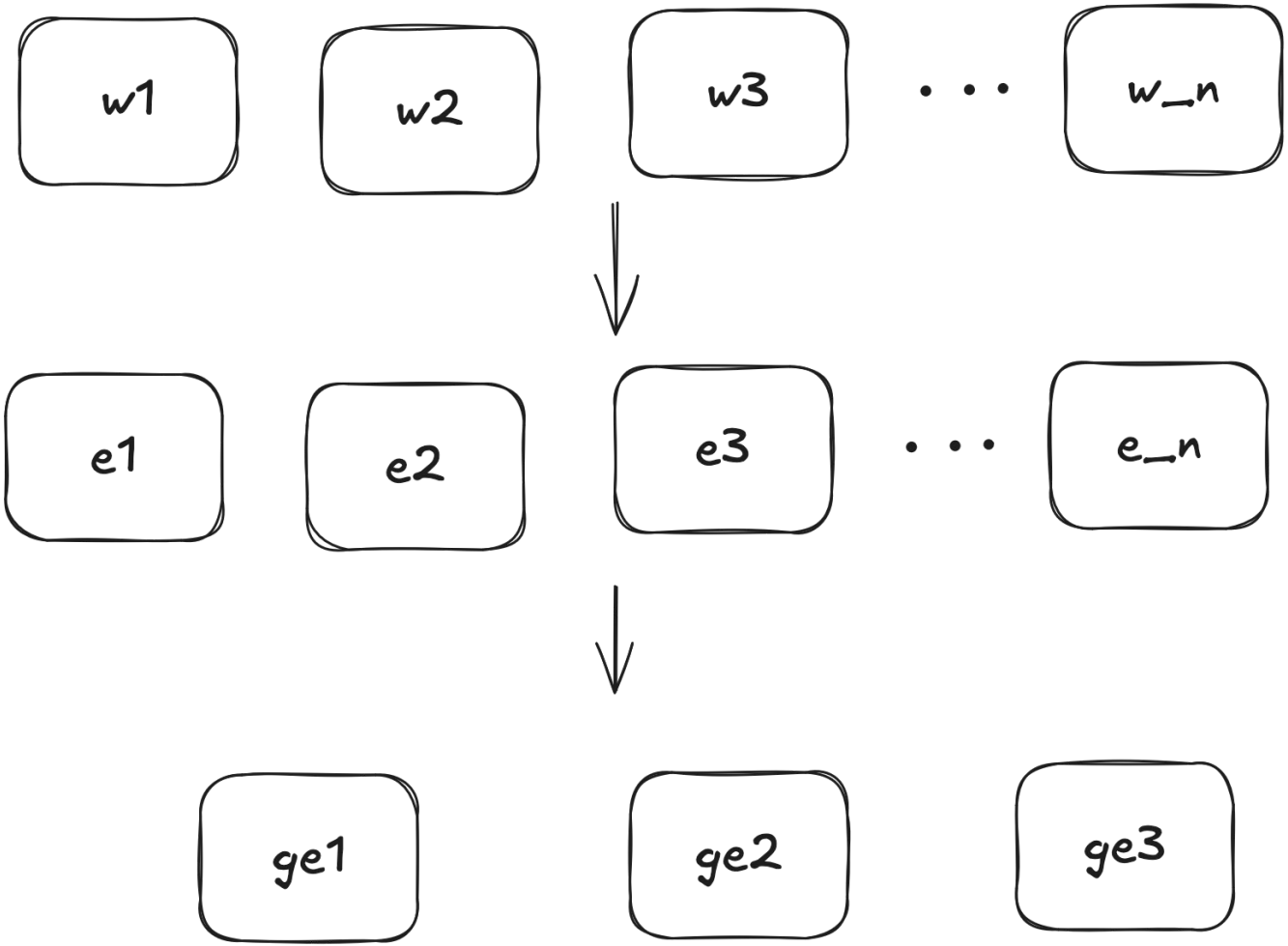
i love cats → [0.1, 0.2, 0, 0, 0.5, 0.7. 0.8, 0]


i love cats, 1
i love dogs, 1
neural net
[0.1, 0.2, 0, 0, 0.5, 0.7. 0.8, 0]
[0.1, 0.1, 0, 0, 0.6, 0.7. 0.8, 0]

[0.1, 0.2, 0, 0, 0.5, 0.7. 0.8, 0]

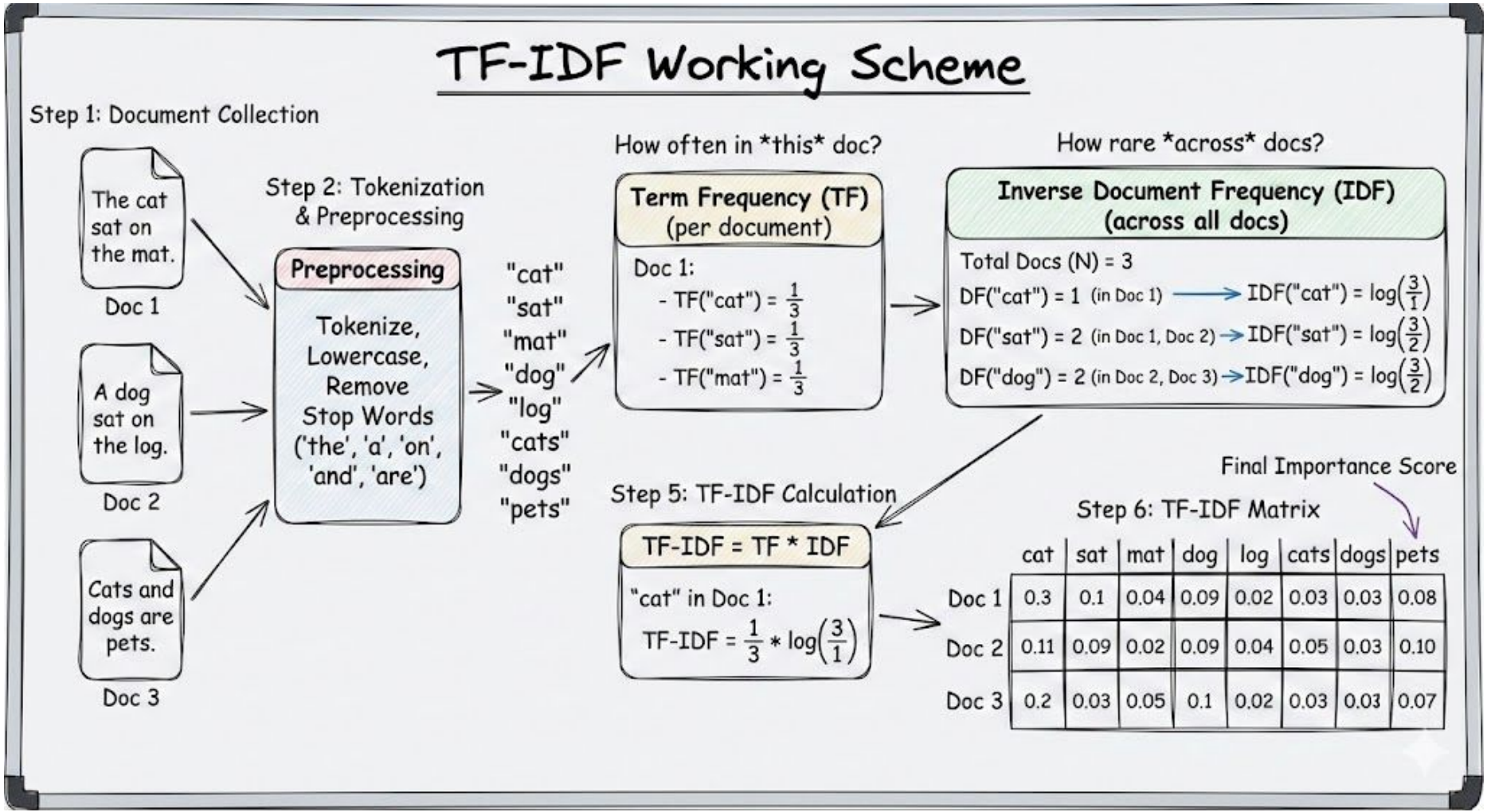
[0.1, 0.1, 0, 0, 0.6, 0.7. 0.8, 0]



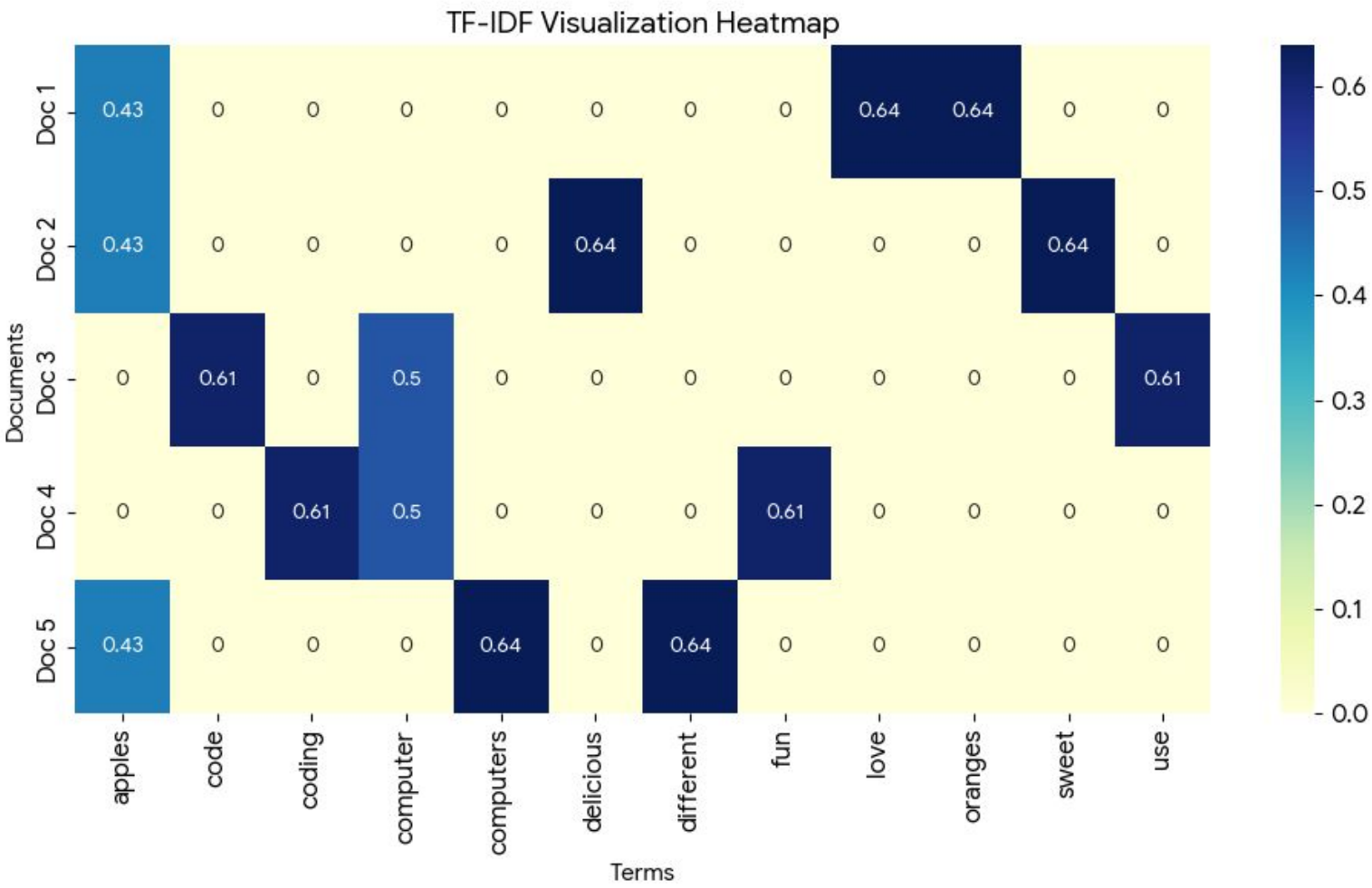

w1
w2
w3
. . .
w_n
e1
e2
e3
. . .
e_n
ge1
ge2
ge3

problem - exact word search

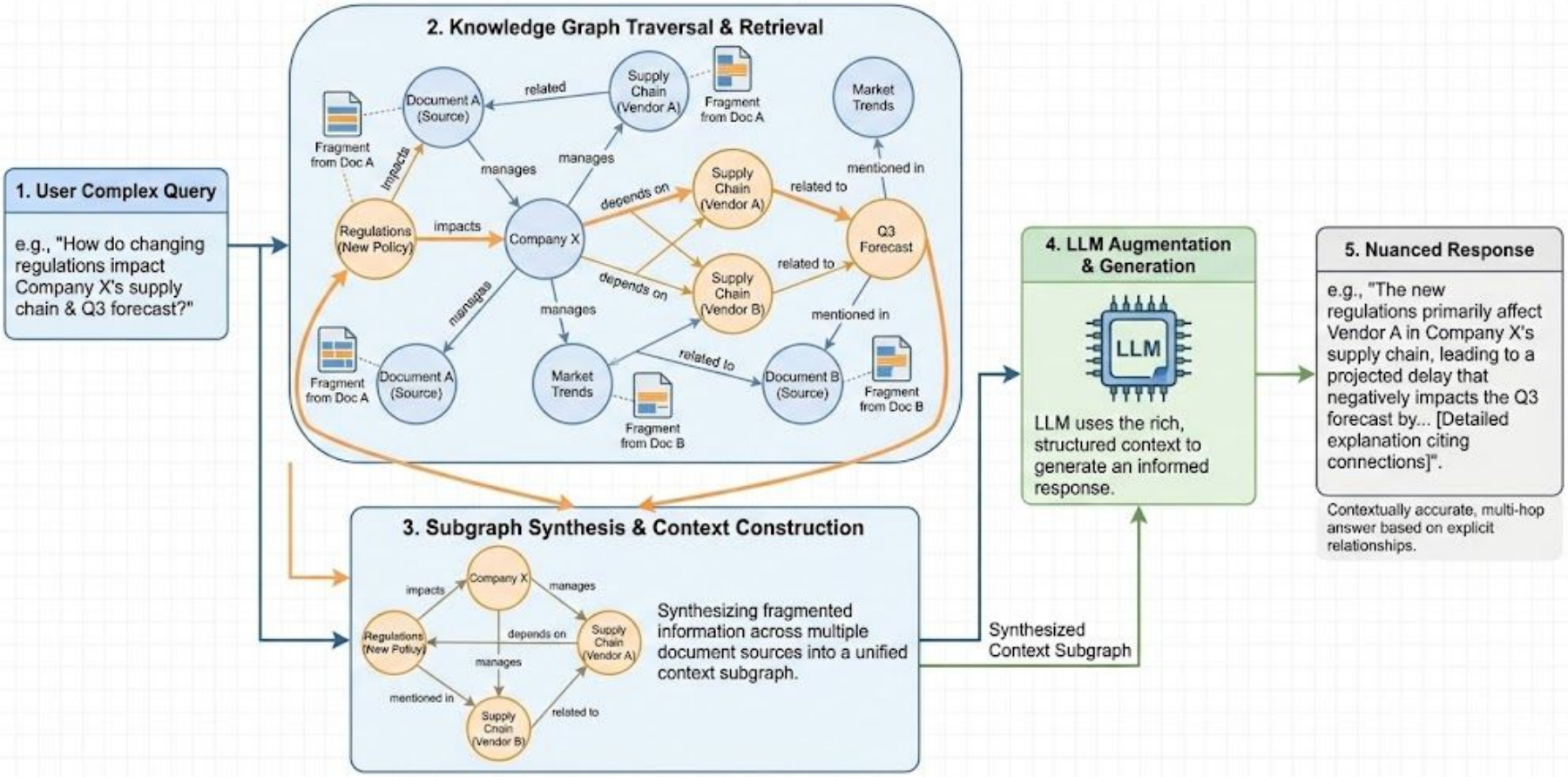
TF-IDF Working Scheme

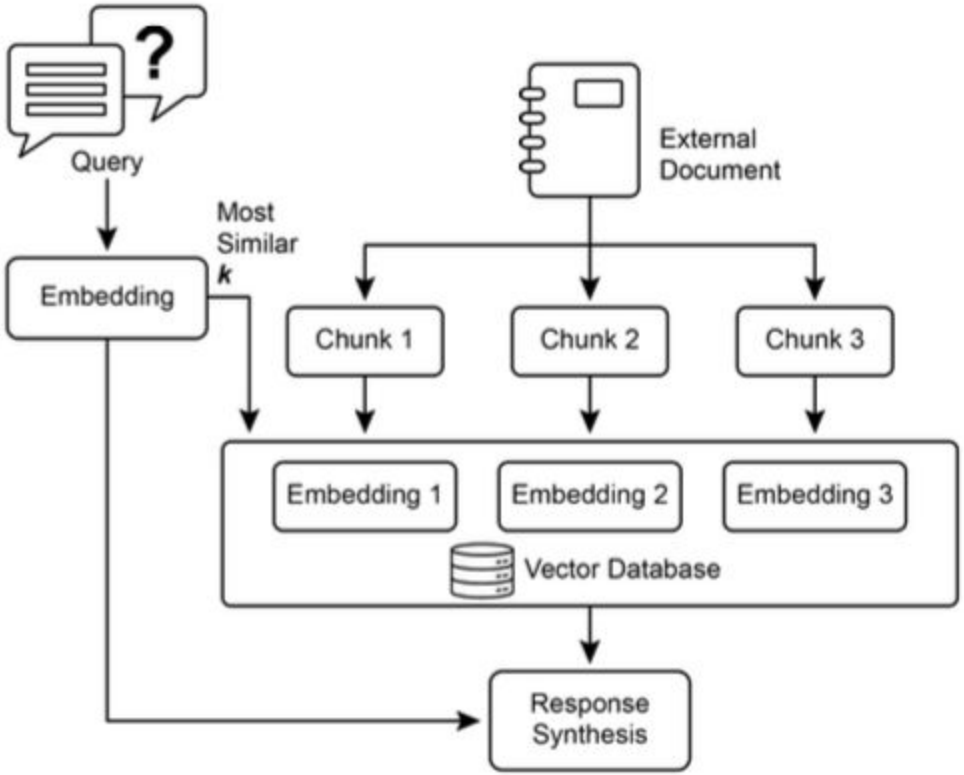


	cat	sat	mat	dog	log	cats	dogs	pets
Doc 1	0.3	0.1	0.04	0.09	0.02	0.03	0.03	0.08
Doc 2	0.11	0.09	0.02	0.09	0.04	0.05	0.03	0.10
Doc 3	0.2	0.03	0.05	0.1	0.02	0.03	0.03	0.07


TF-IDF Visualization Heatmap
Documents
Doc 1
Doc 2
Doc 3
Doc 4
Doc 5
0.43 0 0 0 0 0 0 0 0.64 0.64 0 0
0.43 0 0 0 0 0.64 0 0 0 0 0.64 0
0 0.61 0 0.5 0 0 0 0 0 0 0 0.61
0 0 0.61 0.5 0 0 0 0.61 0 0 0 0
0.43 0 0 0 0.64 0 0.64 0 0 0 0 0
apples
code
coding
computer
computers
delicious
different
fun
love
oranges
sweet
use
Terms
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Graph RAG Working Scheme: Knowledge Graph-Based Retrieval & Synthesis




Query
External
Document
Most
Similar
k
Embedding
Chunk 1
Chunk 2
Chunk 3
Embedding 1
Embedding 2
Embedding 3
Vector Database
Response
Synthesis

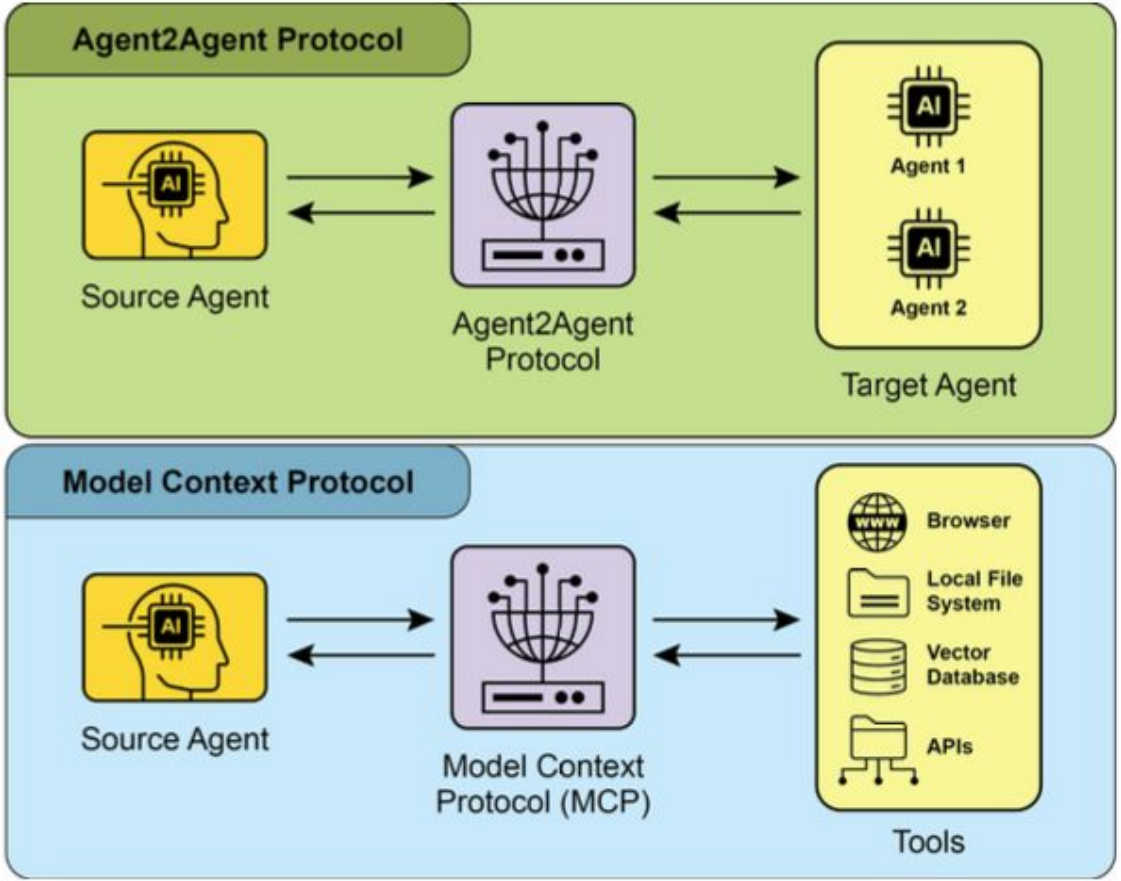


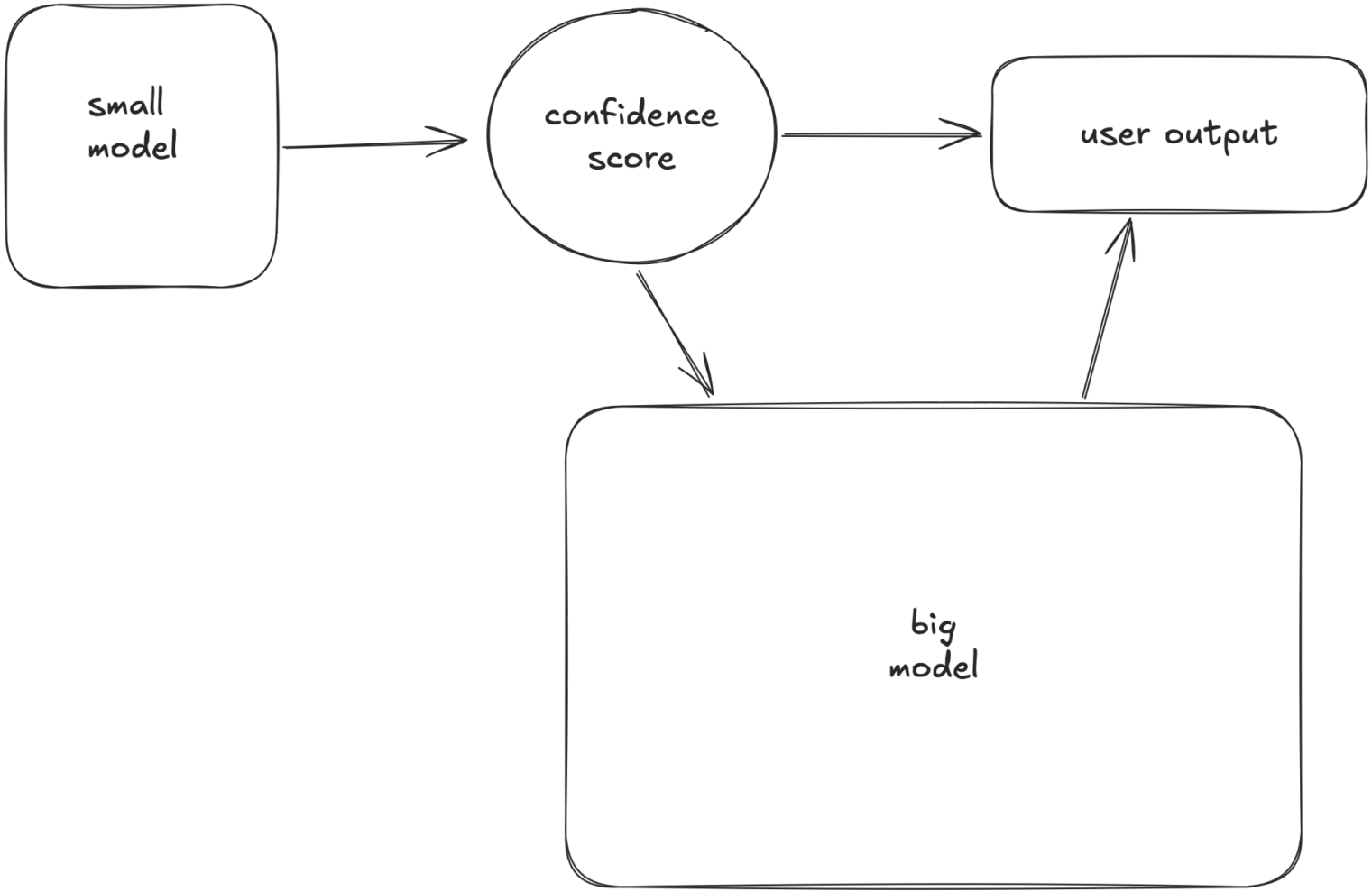
Fig.1: Comparison A2A and MCP Protocols

```
#Synchronous Request Example
{
 "jsonrpc": "2.0",
 "id": "1",
 "method": "sendTask",
 "params": {
   "id": "task-001",
   "sessionId": "session-001",
   "message": {
     "role": "user",
     "parts": [
       {
         "type": "text",
         "text": "What is the exchange rate from USD to EUR?"
       }
     ]
   },
   "acceptedOutputModes": ["text/plain"],
   "historyLength": 5
 }
}
```

```
# Streaming Request Example
{
 "jsonrpc": "2.0",
 "id": "2",
 "method": "sendTaskSubscribe",
 "params": {
   "id": "task-002",
   "sessionId": "session-001",
   "message": {
     "role": "user",
     "parts": [
       {
         "type": "text",
         "text": "What's the exchange rate for JPY to GBP today?"
       }
     ]
   },
   "acceptedOutputModes": ["text/plain"],
   "historyLength": 5
 }
}
```

- Synchronous Request/Response
- Asynchronous Polling
- Streaming Updates (Server-Sent Events - SSE)
- Push Notifications (Webhooks)

- use less data
- use smaller models when possible


small model
confidence score
user output
big model

see you next time
for
“Reasoning Technique, Evaluation and Monitoring, Prioritization, Exploration and Discovery”