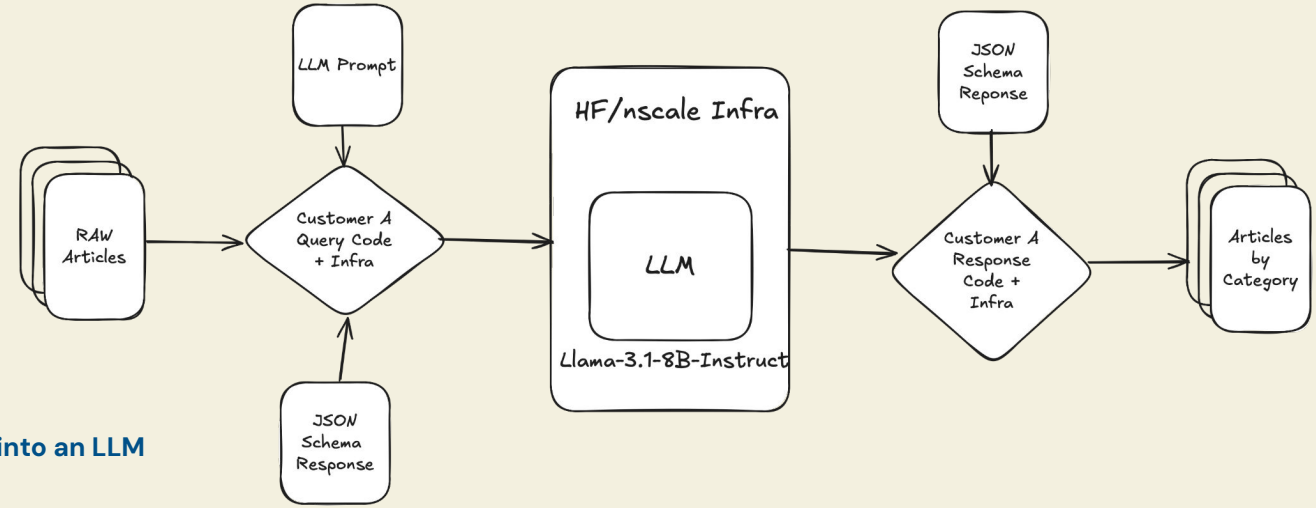
10.28.25

Company A Article Classifier

REDIS Semantic Router POC

1. Current Architecture
2. Pain Points
3. New Architecture
4. Semantic Router
5. Comparative Results
6. Conclusion

Current System



High Level View

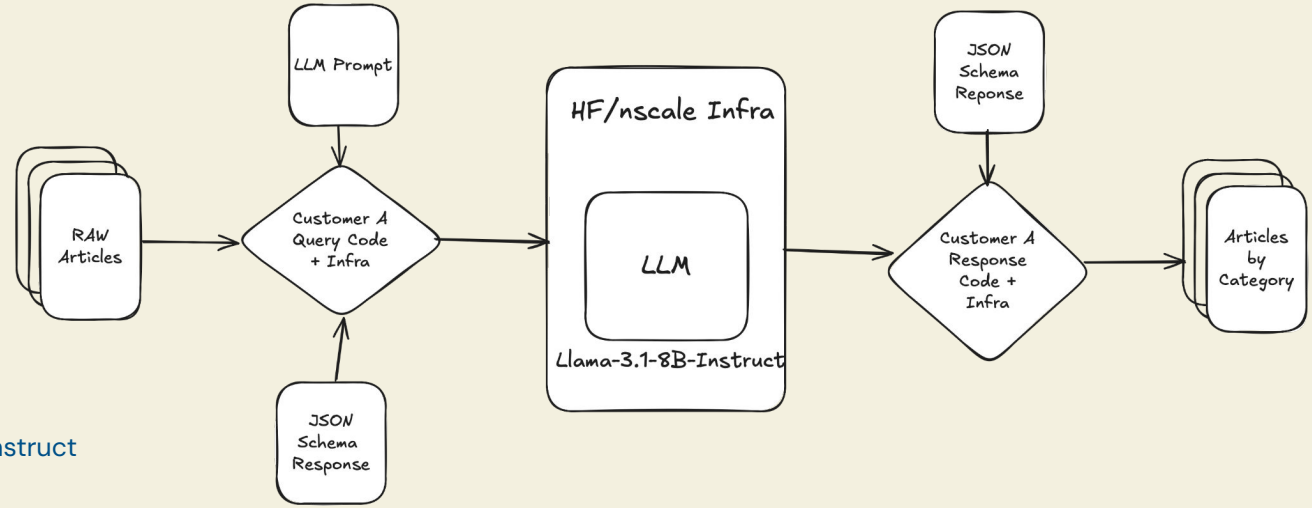
-Articles are formatted and pushed into an LLM prompt

-Categories are provided in the query as a schema response

-Customer A code and infra push query to Hugging Face/nscale infra

-Response contains category (or not)

Pain Point



Infra

-> LLM -> meta-llama/Llama-3.1-8B-Instruct

-> Provider -> Hugging Face / nscale

Pains

-High latency for each LLM calls

-High cost/query

New System

High Level View

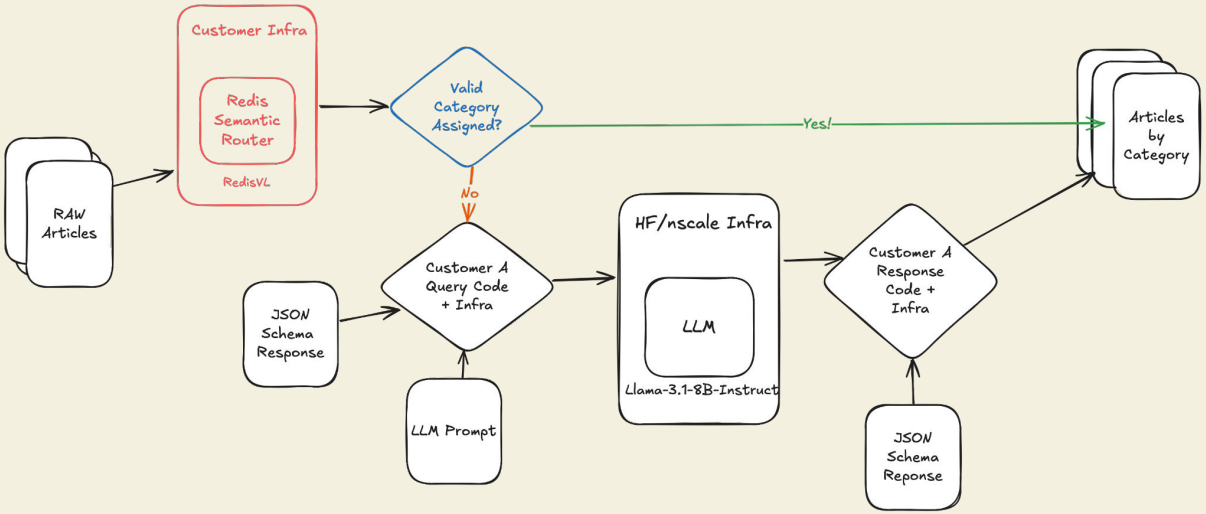
-Articles are formatted and pushed to a Redis Sematic Router

-The Router is hosted on Customer A infra

-The Router generate embedding from the text and store them into Redis

-If it return a valid category -> Success

-If not, it send the article to LLM prompt for expensive inference to LLM

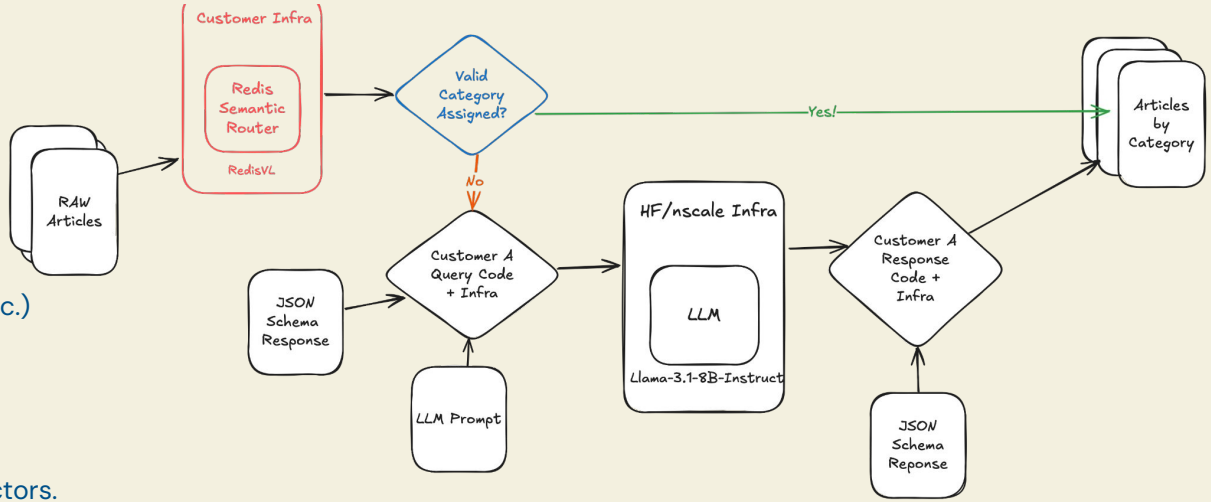


Semantic Router

The Router - Details

1. Each Route = Topic (business, tech, sport, etc.)
   - Stores representative reference embeddings from training articles.
2. At inference time:
   - Embed the new article.
   - Query Redis for its nearest route vectors.
   - If similarity ≥ threshold -> assign category directly.
   - Else -> fallback to LLM for classification.
3. Threshold Optimizer tunes per-route confidence cutoffs

Source: https://docs.redisvl.com/en/latest/user_guide/08_semantic_router.html

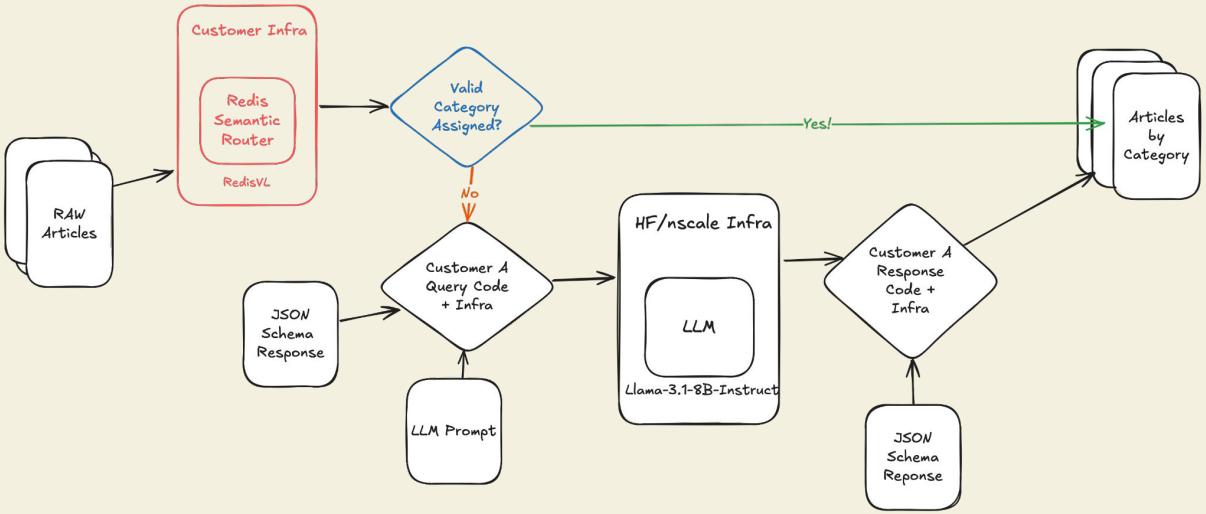


Comparative Results

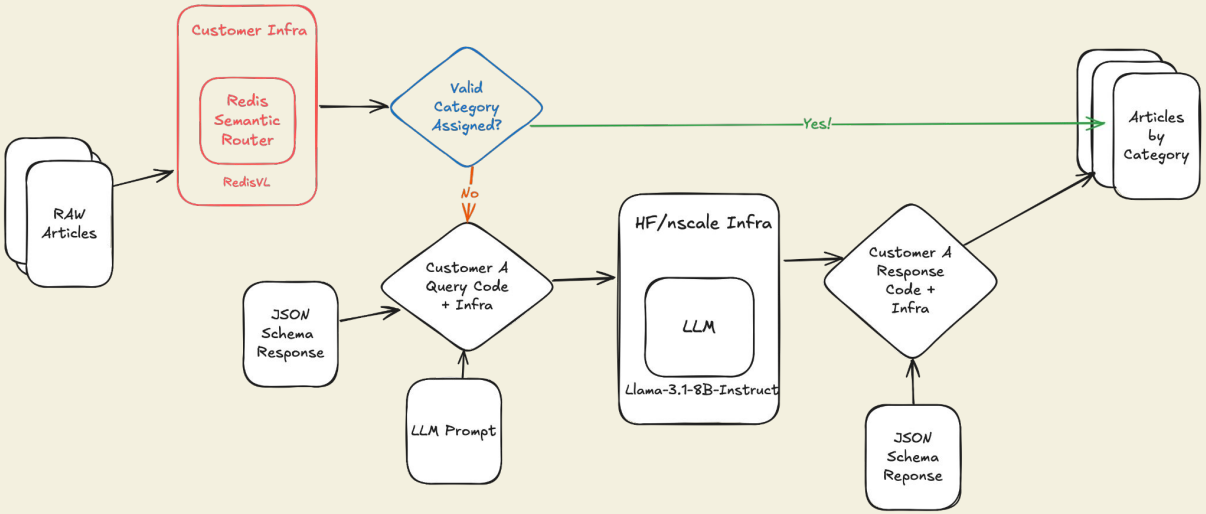
Performance (~100 articles)

Baseline
accuracy=0.78
cost=$0.004993
avg_seconds=1.087
REDIS Semantic Router + LLM
accuracy=0.970
cost=$0.000000
avg_seconds=0.042

Projected cost per ~100k |
baseline=$4.99
router+LLM=$0.00 (no call to HF)



Comparative Results



Performance (~1000 articles)

Baseline
accuracy=0.853
cost=$0.050911
avg_seconds=1.071

REDIS Semantic Router + LLM
accuracy=0.973
cost=$0.000000
avg_seconds=0.040

Projected cost per ~100k
baseline=$5.09
router+LLM=$0.00 (no call to HF)

Conclusion

LLM-only: 100 k daily classifications = expensive + slow

Router-first: 60-90 % routed locally = minimal LLM usage

Result: 3–5× faster, 70–90 % cheaper, same accuracy

https://docs.redisvl.com/en/latest/user_guide/08_semantic_router.html
https://github.com/redis/redis-vl-python
https://scikit-learn.org/stable/modules/neighbors.html