Selection of LLM Model for Perpendo AI

Cloud LLM Models:

Pro's:

1. No Infrastructure Management
2. Instant Scalability
3. Lower Upfront Cost
4. High Availability & Reliability

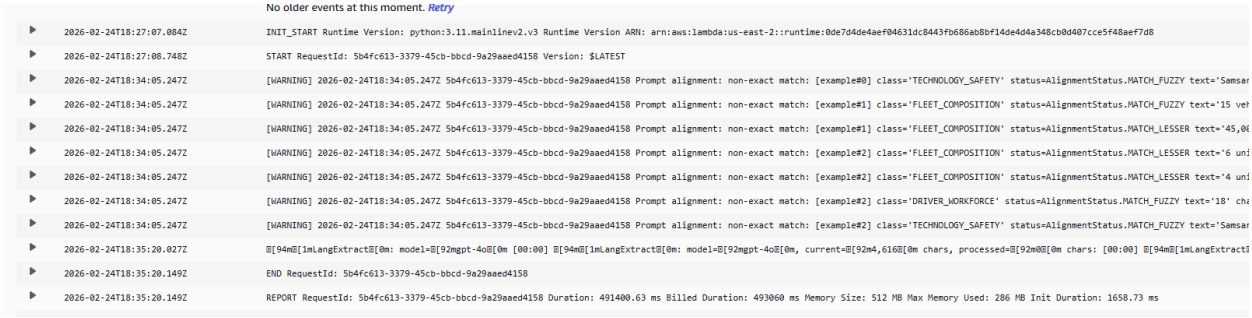
Con's:

1. Ongoing Operational Cost
2. Data Privacy & Compliance Risks
3. Vendor Lock-In
4. Limited Customization

Potential Models to use:

- OpenAI, Claude, Gemini (Very good all around for all tasks, pretty good with lang extract so far)
- Niche Models: Mistral, Qwen (Very good for specific domain tasks, but less known)

Proof of usability:



No older events at this moment. Retry

2026-02-24T18:27:07.084Z	INIT_START Runtime Version: python:3.11.mainlinev2.v3 Runtime Version ARN: arn:aws:lambda:us-east-2::runtime:0de7d4de4aef04631dc8443fb686ab8bf14de4d4a348cb0d407cce5f48aef7d8
2026-02-24T18:27:08.748Z	START RequestId: 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Version: $LATEST
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#0] class='TECHNOLOGY_SAFETY' status=AlignmentStatus.MATCH_FUZZY text='Samsar
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#1] class='FLEET_COMPOSITION' status=AlignmentStatus.MATCH_FUZZY text='15 ve
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#1] class='FLEET_COMPOSITION' status=AlignmentStatus.MATCH_LESSER text='45,0
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#2] class='FLEET_COMPOSITION' status=AlignmentStatus.MATCH_LESSER text='6 uni
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#2] class='FLEET_COMPOSITION' status=AlignmentStatus.MATCH_LESSER text='4 uni
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#2] class='DRIVER_WORKFORCE' status=AlignmentStatus.MATCH_FUZZY text='18' cha
2026-02-24T18:34:05.247Z	[WARNING] 2026-02-24T18:34:05.247Z 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Prompt alignment: non-exact match: [example#2] class='TECHNOLOGY_SAFETY' status=AlignmentStatus.MATCH_FUZZY text='Samsar
2026-02-24T18:35:20.027Z	[94m[1mLangExtract[0m: model=[92mgpt-4o[0m [00:00] [94m[1mLangExtract[0m: model=[92mgpt-4o[0m, current=[92m4,616[0m chars, processed=[92m0[0m chars: [00:00] [94m[1mLangExtract
2026-02-24T18:35:20.149Z	END RequestId: 5b4fc613-3379-45cb-bbcd-9a29aaed4158
2026-02-24T18:35:20.149Z	REPORT RequestId: 5b4fc613-3379-45cb-bbcd-9a29aaed4158 Duration: 491400.63 ms Billed Duration: 493060 ms Memory Size: 512 MB Max Memory Used: 286 MB Init Duration: 1658.73 ms

https://us-east-2.console.aws.amazon.com/cloudwatch/home?region=us-east-2#logsV2:log-groups/log-group/$252Faws$252Flambda$252FBackendStack-dev-PdfHandlerLambdaE161543F-aQspPunQalZu/log-events/2026$252F02$252F24$252F$255B$2524LATEST$255D0c6aa9490de74b95a259bfbb7acec10b

Stats of Cloud LLM:

Average time : 8-12 ~ minutes (depending on size of document)
Average cost: 0.08 - 0.16 cents per request (depends on size)

Self-Hosted LLM Models:

Pro's:

1. Full Data Control & Privacy
2. No Vendor Lock-In
3. Lower Long-Term Cost at Scale
4. Full Customization
5. Performance Optimization

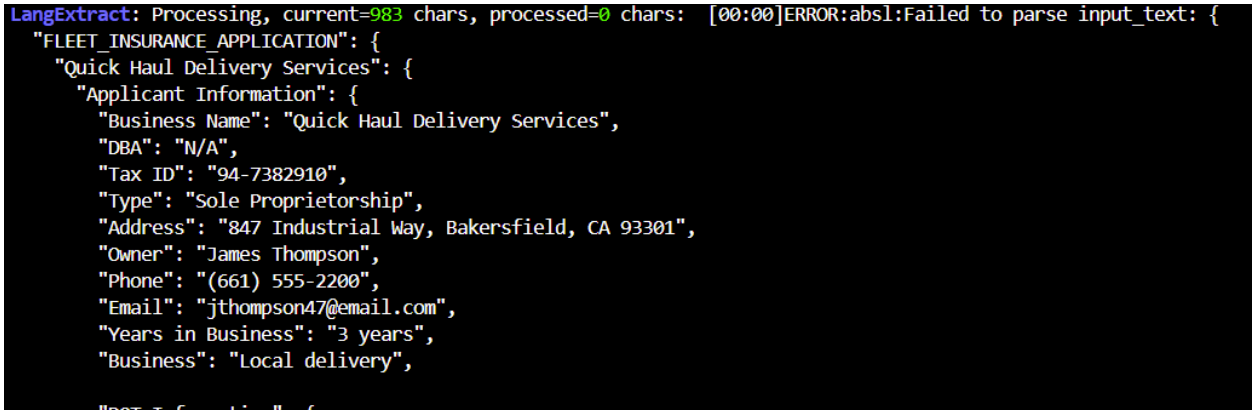
Con's:

1. High Upfront Hardware Costs
2. Ongoing Infrastructure & Ops Complexity
3. Scalability Challenges
4. Maintenance Burden
5. Security Is Now Your Responsibility

Potential Models to use:

- LLaMA 3 (various sizes — e.g., 7B, 34B, 405B)
- Gemma (e.g., Gemma 2 or Gemma 3)
- Falcon (e.g., Falcon-7B, Falcon-40B)
- GLiNER

Proof of usability:
https://youtu.be/HHhcVOxklgE?si=YHizQzhIGWG7D0UW
GitHub - 913165/langextract_ollama



Stats of Local LLM:

- Average Cost per request: 0
- Average time per request: 10-15 minutes