Build vs Buy - LLMAsAService.io

We built LLMAsAService.io from the ground up to support our needs in another LLM enabled SaaS application. This document lists the features we needed to build to safely and reliably offer LLM features. If you want to focus on building cool features using LLMs rather than building the infrastructure to serve those requests, consider LLM as a Service.

Vendor & Model Management

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Multiple Vendor/Model Management	• Create and manage multiple vendor APIs • Add and update models as vendors update their model versions	To avoid downtime for your application or website, and to be able to use the "right" model for your business needs, create any number of vendor model services, and group those into different strengths and purposes.
Unified Coding API	• Normalize API coding interface to specific vendor API which are annoyingly different	Write your code against our unified API interface either as HTTP requests, or using our customer components. Your code will not need to change as new models are added over time. We give you pre-build React hooks, UI components and http APIs.
Secure API Key Management	• Secure encrypted storage of API keys for vendor APIs • Safe retrieval of API keys during inference time	Centrally manage your API keys in one interface. LLMAsAService encrypts and stores API keys which can be rotated and managed in one control panel. These API keys NEVER get coded into your applications or websites,source control, they are only accessed by our service.
Codeless New Model Onboarding	• Allow emerging models to be tested against real prior prompts to confirm quality • Put new models into service without code changes or redeployment	Testing and using new models is a streamlined, codefree process. You can check quality by comparing the response of the new model against prior prompts, and when you are confident send traffic to the new model.

Reliability

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Multiple vendor failover	• Detect failed API call and try another vendor	LLMAsAService tries alternative services if one vendor fails for any reason. This is totally transparent to customer calls. We also allow you to create “Chaos Monkeys” that intentionally fail at random times so you can ensure reliability when vendors have downtime.
Vendor Token Quotas/Limits	• Detect and manage token limits & requests per second failures	Vendors allocate each caller a token and request per second quota. If your application exceeds this quota during peak periods, LLMAsAService fails-over to another vendor automatically.
Call Error Logging	• Detect and alert when APIs return errors	Vendor API calls are logged, including errors when they fail. This helps understand and resolve any configuration issues quickly.
Application Firewall/DDoS Protection	• Block repetitive calls that are indicative of denial of service attacks	LLMAsAservice implements firewall protections against numerous attacks including denial of service.

Safety (Prompt and Response Trustworthiness)

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
System Instructions (brand and policy enforcement)	• Manage system instructions for brand and safety policy enforcement • Inject system instructions into all prompts sent externally	Define System instructions universally or at an individual service group level. These are automatically injected into EVERY prompt transparently to the developer. This ensures every LLM response has the right instructions for brand, legal policy and tone.
PII (Personal Identifiable Information)	• Detect personally identifiable information • Redact or reject prompt with PII to avoid external disclosure	A pre-trained ML model analyzes every prompt for financial, credentials, and other personal information and gives you the choice to redact (tokenize), or block those prompts. Tokenization temporarily replaces PII with random strings, and then transparently reverts those in the

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
		response, keeping PII away from LLM vendors.
Toxic Prompt Rejection	• Detect prompts that contain toxic language • Block and log these prompts and customer/users for policy review	A pre-trained LLM model analyzes every prompt for Insults, hate speech, harassment or abuse, profanity, violence or threat, sexual:, and graphic language. These prompts can be blocked when they exceed a defined threshold, and logged for review.
Blocked Topics	• Specify a set of topics and phrases to be blocked • Block and log any prompt or customer/user who attempts to use these	Curate a list of topics and words that MUST be blocked. This can include sensitive words, known hate speech, or internal product names. Any prompt containing these words will be blocked and logged for review.

Customer Management

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Create and Manage Customers	• Create customers on-demand or in-advance of their first use • Keep track of customer usage and tier level	Control new customer creation based on first call, API, or by manual creation. Manage customer tier levels and data residency requirements, and other data.
Customer Token Budgeting / Replenishment	• Give new customers an initial token allocation • Replenish that token allocation daily/weekly/monthly	Avoid cost overruns either through malicious abuse or customer exuberance. Allow customer trial access with an initial token quota, and define the replenishment rules. Query customers running low on tokens to upsell or remedy.
Customer Cost Tracking	• See what customers are costing you the most	See what customers/enterprise users are using your services the most in dollars.
Customer Management API (and Zapier add-on)	• Create an API to integrate CRM and product purchases with LLM feature access	Public API and zapier add-on allow integration with other business tools like Stripe or CRM systems. Trigger specific processes when new customers make a first call or customers are running low on tokens.

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Geo Constrain (EU customers got to EU hosted services)	• Track customer geography and data residency requirements • Route customers to the right vendor hosting geography	Route customers with data residency requirements to the right vendor API geographies. Important for any E?U customers you might have.

Logging and Observability

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Real-time 30 day Call Logging	• Show recent calls for monitoring and cost analysis • View failed calls to identify vendor downtime or API key problems • Delete calls after pre-defined retention period	Calls within the last 30 days are always visible for cost, status, quality and abuse review. The retention period can be set for automatically purging this data based on your internal retention policies.
Nightly/Weekly/Monthly Log File Generation	• Create log files with call details • Delete log files after pre-defined retention period	Downloadable JSON formatted log files with call details can be produced nightly, weekly or monthly. The retention period for these logs can be set to automatically purge these files based on your internal retention policies.
Calls by country	• Monitor country of origin for calls to detect threats	World map of call locations help you spot excessive calls coming from countries you don't serve allowing you to block those customers.

Cost Control and Optimization

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Auto-routing based on prompt complexity	• Detect the complexity of prompts and route them to less expensive models unless required.	Save on average 60% vendor costs by smart routing to different strength models based on a pre-trained ML model. Capture examples of preferred models using your own examples to better tune this algorithm in the future.
Costs by vendor & Customers	• Rollup the cost by LLM vendor • Rollup the cost by customer	Avoid surprises by seeing real time 30 day costs by both LLM vendor and customer. Understand and manage spend patterns early.

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Response caching	• Cache responses for identical requests within a given timeframe	Control the caching time in minutes where identical prompts get the same response. Avoids LLM vendor costs (and response time) where users send the same prompt multiple times (often in applications where summaries are generated)
Customer token quotas	• Track token usage by customer and block calls after a given quote is used	Whether intentionally or not, block requests after a customer has exhausted their quota to avoid cost overruns.

Developer Tools

Feature	Build: Spec to reach parity with us	Buy: LLMAsAService.io
Unified Call API (code once run against any vendor)	• Single vendor agnostic API interface	Changing model and vendor shouldn't need a code change or redeployment. We allow your developers to have ONE coding interface, and defer the decisions about vendor and model.
React Hook (useLLM)	• Simple single line react/next/javascript hook component	A NPM package allows a single line of code to make secure and safe calls to any LLM.
UI Chat Panel	• Fully function multi-turn chat UI component with full styling control	A NPM fully functional UI component of the quality you would expect for chat style LLM integration in your websites and application.
Embedding Code Builders	• iFrame embedding for custom agents for any website or wordpress instance	A simple embedding wizard helps build agents. Simple copy and paste iFrame code into any website or wordpress instance.

Conclusion

Reliably and safely deploying LLM features and agents is a lot of small but important features. Free your developers from building this infrastructure by using LLMAsAService.io, and let them focus on how LLM features can improve your customers' lives. For more information, see LLMAsAService.io.