viettel
IDC

Datasheet

Viettel Model as a Service (vMaaS)



Datasheet

Viettel Model as a Service (vMaaS)

OVERVIEW

Viettel Model as a Service (vMaaS) provides AI models as API-based services, supporting a wide range of tasks such as language, image, audio, and embedding processing through an OpenAI-compatible API. The service enables enterprises to quickly integrate AI into their applications, products, or operational workflows without investing in GPU/NPU infrastructure or building a dedicated MLOps team.

vMaaS is fully operated on Viettel IDC's data center infrastructure in Vietnam, powered by Qualcomm AI Inference Suite and dedicated AI accelerators optimized for inference. This helps improve processing performance, optimize operating costs, and support scalable AI deployment.

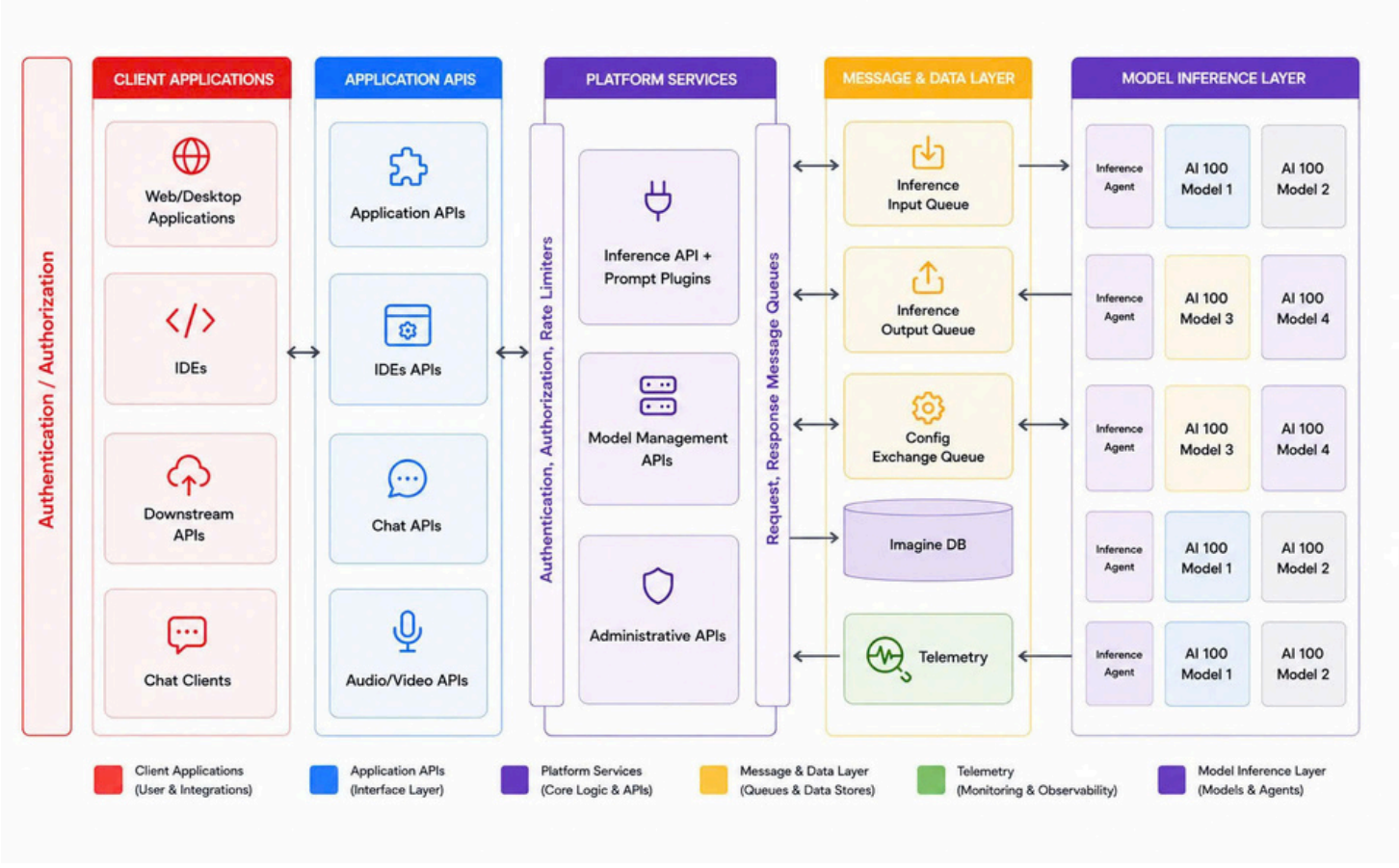

Figure 1. Overall architecture of the Viettel Model as a Service platform

Datasheet

Viettel Model as a Service (vMaaS)

HIGHLIGHTS

AI Models Optimized for Vietnam

vMaaS supports AI models tailored to Vietnamese language context and practical deployment needs in Vietnam. The platform can also be extended to industry-specific use cases such as BFSI, retail, logistics, healthcare, and the public sector.

Cost-Efficient & Easy to Deploy

With OpenAI-compatible APIs and a flexible service model, vMaaS helps enterprises integrate AI into existing systems without investing in dedicated GPU/NPU infrastructure or a separate MLOps team. Powered by Qualcomm AI Inference Suite and AI accelerators optimized for inference, the platform delivers high processing performance with optimized operating costs.

Data Security & Operated in Vietnam

vMaaS is fully deployed on Viettel IDC's data center infrastructure in Vietnam, helping enterprises ensure data sovereignty, comply with local data regulations, and strengthen information security. Customer data is not used to train third-party models.

Expandable AI Ecosystem

Beyond AI Model APIs, vMaaS is designed to evolve into a broader AI ecosystem with services such as MLOps, Vector Database, Knowledge Base-as-a-Service, and Agent Builder—helping enterprises scale from experimentation to enterprise deployment.

Datasheet

Viettel Model as a Service (vMaaS)

CORE FEATURES

vMaaS provides a comprehensive feature set that enables enterprises to build, deploy, manage, and scale AI applications through OpenAI-compatible APIs.

AI Applications

Provides user-facing AI capabilities, enabling users to interact with and visually observe results from popular model types such as LLMs, text-to-image models, and embedding models.

Developer Tools

Provides REST APIs and SDKs compatible with OpenAI standards, enabling rapid integration into existing applications. Includes technical documentation, code samples, and test environments for developers.

Model Management

Enables enterprises to manage model catalogs, configure usage limits, and deploy custom models through BYOM (Bring Your Own Model).

Account & Usage Management

Supports API key management, usage monitoring, token consumption reporting, and billing or credit management.

Support & Operations

Provides documentation, FAQs, and technical support from Viettel IDC to help ensure stable AI deployment and operations.

Datasheet

Viettel Model as a Service (vMaaS)

CUSTOMER BENEFITS

Aspect	Before vMaaS	With vMaaS
AI Deployment Time	Enterprises may need months to invest in GPU/NPU infrastructure, deploy inference systems, and integrate AI models.	With API-based access, enterprises can deploy AI applications quickly without upfront AI infrastructure investment or a dedicated operations team.
Initial Investment	Requires significant CAPEX for AI servers, power, cooling, software, and MLOps teams.	Uses a service-based model with flexible consumption packages, reducing the need for upfront AI infrastructure investment.
Data Sovereignty & Security	Data may be processed or stored on overseas platforms, creating challenges for compliance and data control.	Data is processed and hosted in Vietnam on Viettel IDC infrastructure, supporting security and regulatory requirements.
Multi-model Integration	Enterprises often have to work with multiple providers, SDKs, API standards, and pricing models.	A single OpenAI-compatible API enables access to multiple AI models on one unified platform.
Vietnamese Language & Industry Relevance	International models may not fully understand Vietnamese language context or local business needs.	Supports AI models optimized for Vietnamese use cases and can be extended for industry-specific requirements.
AI Expansion Roadmap	Deploying RAG, Agents, or MLOps often requires additional platforms and repeated integration efforts.	Enterprises can expand from MaaS to KBaaS, Agent Builder, and MLOps within the Viettel IDC AI ecosystem.

Datasheet

Viettel Model as a Service (vMaaS)

AI MODEL CATALOG

09 models
high-performance models ready to use

High throughput
optimized for Qualcomm NPU infrastructure

Continuous updates
continuous roadmap expansion and updates

Category	Model	Supported languages	Key highlights
Large Language Model (LLM)	**Qwen/Qwen3-32B**	Supports 100+ languages and dialects	A high-performance language model designed for complex tasks such as multi-step reasoning, in-depth document analysis, coding assistance, advanced RAG, and tool-based workflow automation.
	Qwen/Qwen3-14B	Supports 100+ languages and dialects	Balances performance and cost efficiency, making it suitable for chatbots, text analysis, content summarization, reasoning, and enterprise tasks that require high accuracy.
	Meta-llama/Llama-3.1-8B	English, German, French, Italian, Portuguese, Hindi, Spanish, Thai	A multilingual long-context model suitable for text summarization, Q&A, tool use, reasoning, and coding assistance within the 8B model class.
	Aisingapore/Llama-SEA-LION-v3.5-8B-R	Vietnamese, Chinese, English, Filipino, Indonesian, Javanese, Khmer, Lao, Malay, Sundanese, Tamil, Thai	Optimized for Southeast Asian languages and cultural contexts, supporting local-language understanding, cross-language translation, and multilingual conversations within a single session.
	Qwen/Qwen3-4B-Instruct-2507	Supports 100+ languages and dialects	A compact and efficient language model suitable for instruction following, logical reasoning, reading comprehension, mathematics, coding, and tool use.

Datasheet

Viettel Model as a Service (vMaaS)

AI MODEL CATALOG

RAG-ready models
Optimized for RAG and knowledge-based Q&A

Multimodal support
Process text, images, and documents

Practical use cases
Built for diverse enterprise AI needs

Category	Model	Supported languages	Key highlights
Image Generation Model	**stabilityai/sdxl-turbo**	English	A text-to-image generation model with fast response speed, suitable for creating illustrations, creative content, and near real-time image generation use cases.
Embedding Model	**BAAI/bge-m3**	Supports 100+ languages	Converts text into semantic vectors and supports long input of up to 8,192 tokens. Suitable for semantic search, RAG systems, enterprise knowledge-base Q&A, and duplicate content detection.
Reranker Model	**BAAI/bge-reranker-v2-m3**	Supports 100+ languages	Evaluates and re-ranks the relevance between a query and text passages. Commonly used in RAG pipelines to prioritize the most accurate information before generating responses.
Vision-Language Model (VLM)	**Qwen/Qwen2.5-VL-7B**	Supports 29+ languages, including English and Vietnamese	A vision-language model that supports image recognition, image description, visual reasoning, and Q&A based on image content. Suitable for scanned documents, invoices, forms, image analysis, and automated image content moderation.

Datasheet

Viettel Model as a Service (vMaaS)

PACKAGE SPECIFICATIONS

vMaaS offers flexible service packages based on monthly token usage, suitable for different customer groups from experimentation and application development to enterprise-scale AI deployment.

Package	Monthly Token Quota	Recommended Users
Starter Dev	5M tokens	Individual developers, test teams, or small-scale PoC projects.
Freelance Dev	10M tokens	Freelancers, small application development teams, or personal automation use cases.
SME	50M tokens	Small and medium-sized enterprises integrating AI into products, applications, or internal workflows.
ISV	100M tokens	AI agencies, ISVs, or solution providers developing AI solutions for end customers.
Customize	On demand	Large enterprises or corporations with high token usage, custom resource requirements, or BYOM deployment needs.
Add-on Token	+1M tokens	Additional token package for customers who need to extend usage beyond their main package limit.

Technical Notes

- Tokens are purchased by package and can be used across the available models on the vMaaS platform.
- New models added in future releases may be updated into the platform's available model catalog.
- For specific requirements, customers can choose the Customize package based on usage scale, resource requirements, or custom model deployment needs.

Datasheet

Viettel Model as a Service (vMaaS)

USE CASES

vMaaS supports a wide range of practical AI use cases, helping enterprises accelerate deployment and create business value.

Use case	Short Description	Recommended Models
AI Customer Service Chatbot	Virtual assistant that automatically answers customer questions 24/7 across websites, applications, Zalo OA, and other service channels. Helps reduce workload for call centers and customer support teams.	Qwen3-32B / Qwen3-14B
Internal RAG / Knowledge Base	Builds a Q&A system based on internal knowledge sources such as processes, policies, product documentation, and operating guidelines. Supports accurate search and source-grounded answers.	Qwen3-4B-Instruct / Qwen3-32B / BGE-M3 / BGE-Reranker-V2-M3
Document Summarization & Analysis	Summarizes contracts, financial reports, meeting minutes, and long documents. Supports key insight extraction, multi-document comparison, and decision-support analysis.	Qwen3-14B / Llama3.1-8B
Coding Assistant	Supports code generation, code suggestions, code review, bug detection, and IDE integration through OpenAI-compatible APIs.	Qwen3-32B / Qwen3-14B
Marketing Content Generation	Supports blog writing, social posts, email marketing, product descriptions, and content variations across different styles, contexts, and target audiences.	Qwen3-32B / Qwen3-14B / Llama3.1-8B
Process Automation (RPA + AI)	Classifies emails, extracts information from forms and documents, and calls tools or functions to complete steps in internal workflows.	Qwen3-14B / Qwen3-4B-Instruct

Datasheet

Viettel Model as a Service (vMaaS)

WHY VIETTEL IDC?

- **A Leading Cloud and Data Center Provider in Vietnam**
  Established in 2008, Viettel IDC is one of Vietnam's pioneering providers of cloud and data center services. With a diverse ecosystem of more than 50 services covering compute, storage, networking, security, monitoring, and digital infrastructure, Viettel IDC provides a trusted foundation for enterprise innovation and digital transformation.



- **High-quality Services**
  Our service infrastructure is designed, operated, and delivered in accordance with international standards:
  - Data centers compliant with ANSI/TIA-942-B:2017 Rated 3.
  - Cloud infrastructure built and developed based on international standards such as NIST and CNCF.
  - Experienced security specialists with recognized certifications.
  - Service availability of up to 99.99%.

- **Comprehensive Certifications**
  Viettel IDC offers one of the most comprehensive ecosystems of certifications and standards in Vietnam for data center and cloud services, helping customers enhance trust, compliance, and service quality assurance.

- **Modern Technology, Aligned with Future Trends**
  Viettel IDC is committed to adopting advanced and practical technologies for customers. With innovation as a guiding principle, we continuously select, deploy, and update leading technology solutions across our cloud, data center, and AI service ecosystem.

- **Experienced Experts and Engineers**
  Our 24/7 technical support team consists of experienced engineers and specialists with strong expertise across multiple technology domains, enabling us to support customers effectively in solving complex business and technical challenges.