Welcome to Dify!

Dify is an open-source large language model (LLM) application development platform. It combines the concepts of Backend-as-a-Service and LLMOps to enable developers to quickly build production-grade generative AI applications. Even non-technical personnel can participate in the definition and data operations of AI applications.

By integrating the key technology stacks required for building LLM applications, including support for hundreds of models, an intuitive Prompt orchestration interface, high-quality RAG engines, and a flexible Agent framework, while providing a set of easy-to-use interfaces and APIs, Dify saves developers a lot of time reinventing the wheel, allowing them to focus on innovation and business needs.

Why Use Dify?

You can think of libraries like LangChain as toolboxes with hammers, nails, etc. In comparison, Dify provides a more production-ready, complete solution - think of Dify as a scaffolding system with refined engineering design and software testing.

Importantly, Dify is **open source**, co-created by a professional full-time team and community. You can self-deploy capabilities similar to Assistants API and GPTs based on any model, maintaining full control over your data with flexible security, all on an easy-to-use interface.

> Our community users summarize their evaluation of Dify's products as simple, restrained, and rapid iteration.
> - Lu Yu, Dify.AI CEO

We hope the above information and this guide can help you understand this product. We believe Dify is made for you.

What Can Dify Do?

ℹ The name Dify comes from Define + Modify, referring to defining and continuously improving your AI applications. It's made for you.

Startups - Quickly turn your AI ideas into reality, accelerating both success and failure. In the real world, dozens of teams have already built MVPs to get funding or win customer orders through Dify.

Integrate LLMs into existing businesses - Enhance capabilities of current apps by introducing LLMs. Access Dify's RESTful APIs to decouple Prompts from business logic. Use Dify's management interface to track data, costs and usage while continuously improving performance.

Enterprise LLM infrastructure - Some banks and internet companies are deploying Dify as an internal LLM gateway, accelerating the adoption of GenAI technologies while enabling centralized governance.

Explore LLM capabilities - Even as a tech enthusiast, you can easily practice Prompt engineering and Agent technologies through Dify. Over 60,000 developers have built their first app on Dify even before GPTs came out.

Technical Spec

For those already familiar with LLM application tech stacks, this document serves as a shortcut to understand Dify's unique advantages

> We adopt transparent policies around product specifications to ensure decisions are made based on complete understanding. Such transparency not only benefits your technical selection, but also promotes deeper comprehension within the community for active contributions.

Project Basics

Established	March 2023
Open Source License	Apache License 2.0 with commercial licensing
Official R&D Team	Over 10 full-time employees
Community Contributors	Over 120 people
Backend Technology	Python/Flask/PostgreSQL
Frontend Technology	Next.js
Codebase Size	Over 130,000 lines
Release Frequency	Average once per week

Technical Features

LLM Inference Engines	Dify Runtime (LangChain removed since v0.4)
Commercial Models Supported	**10+**, including OpenAI and Anthropic Onboard new mainstream models within 48 hours
MaaS Vendor Supported	**7**, Hugging Face, Replicate, AWS Bedrock, NVIDIA, GroqCloud, together.ai,, OpenRouter
Local Model Inference Runtimes Supported	**6**, Xoribits (recommended), OpenLLM, LocalAI, ChatGLM,Ollama, NVIDIA TIS
OpenAI Interface Standard Model Integration Supported	∞
Multimodal Capabilities	ASR Models Rich-text models up to GPT-4V specs
Built-in App Types	Text generation, Conversational, Agent, Workflow, Group(Q2 2024)
Prompt-as-a-Service Orchestration	Visual orchestration interface widely praised, modify Prompts and preview effects in one place. **Orchestration Modes** Simple orchestration Assistant orchestration Flow orchestration Multi-Agent orchestration(Q2 2024) **Prompt Variable Types** String Radio enum

	External API File (Q2 2024)
Agentic Workflow Features	Industry-leading visual workflow orchestration interface, live-editing node debugging, modular DSL, and native code runtime, designed for building more complex, reliable, and stable LLM applications. Supported Nodes LLM Knowledge Retrieval Question Classifier IF/ELSE CODE Template HTTP Request Tool
RAG Features	Industry-first visual knowledge base management interface, supporting snippet previews and recall testing. **Indexing Methods** Keywords Text vectors LLM-assisted question-snippet model **Retrieval Methods** Keywords Text similarity matching Hybrid Search N choose 1 Multi-path recall **Recall Optimization** Re-rank models
ETL Capabilities	Automated cleaning for TXT, Markdown, PDF, HTML, DOC, CSV formats. Unstructured service enables maximum support. Sync Notion docs as knowledge bases.
Vector Databases Supported	Qdrant (recommended), Weaviate, Zilliz
Agent Technologies	ReAct, Function Call. **Tooling Support** Invoke OpenAI Plugin standard tools Directly load OpenAPI Specification APIs as tools **Built-in Tools** 30+ tools(As of Q1 2024)
Logging	Supported, annotations based on logs
Annotation Reply	Based on human-annotated Q&As, used for similarity-based replies. Exportable as data format for model fine-tuning.
Content Moderation	OpenAI Moderation or external APIs
Team Collaboration	Workspaces, multi-member management
API Specs	RESTful, most features covered

Deployment Methods	Docker, Helm

Model Providers

Dify supports the below model providers out-of-box:

Provider	LLM	Embedding	Rerank
OpenAI	✓(🛠)(👓)	✓	
Anthropic	✓		
Azure OpenAI	✓(🛠)(👓)	✓	
Google	✓(👓)		
Cohere	✓	✓	✓
Bedrock	✓		
together.ai	✓		
Ollama	✓	✓	
Replicate	✓	✓	
Hugging Face	✓	✓	
Zhipu AI	✓(🛠)(👓)	✓	
Baichuan	✓	✓	
Spark	✓		
Minimax	✓(🛠)	✓	
Tongyi	✓		
Wenxin	✓		
Moonshot AI	✓(🛠)		
deepseek	✓(🛠)		
Jina		✓	✓
ChatGLM	✓(🛠)		
Xinference	✓(🛠)(👓)	✓	✓
OpenLLM	✓	✓	
LocalAI	✓	✓	
OpenAI API-Compatible	✓	✓	

where (🛠) denotes Function Calling and (👓) denotes support for vision.

This table is continuously updated. We also keep track of model providers requested by community members here. If you'd like to see a model provider not listed above, please consider contributing by making a PR. To learn more, check out our Contributing Guide.