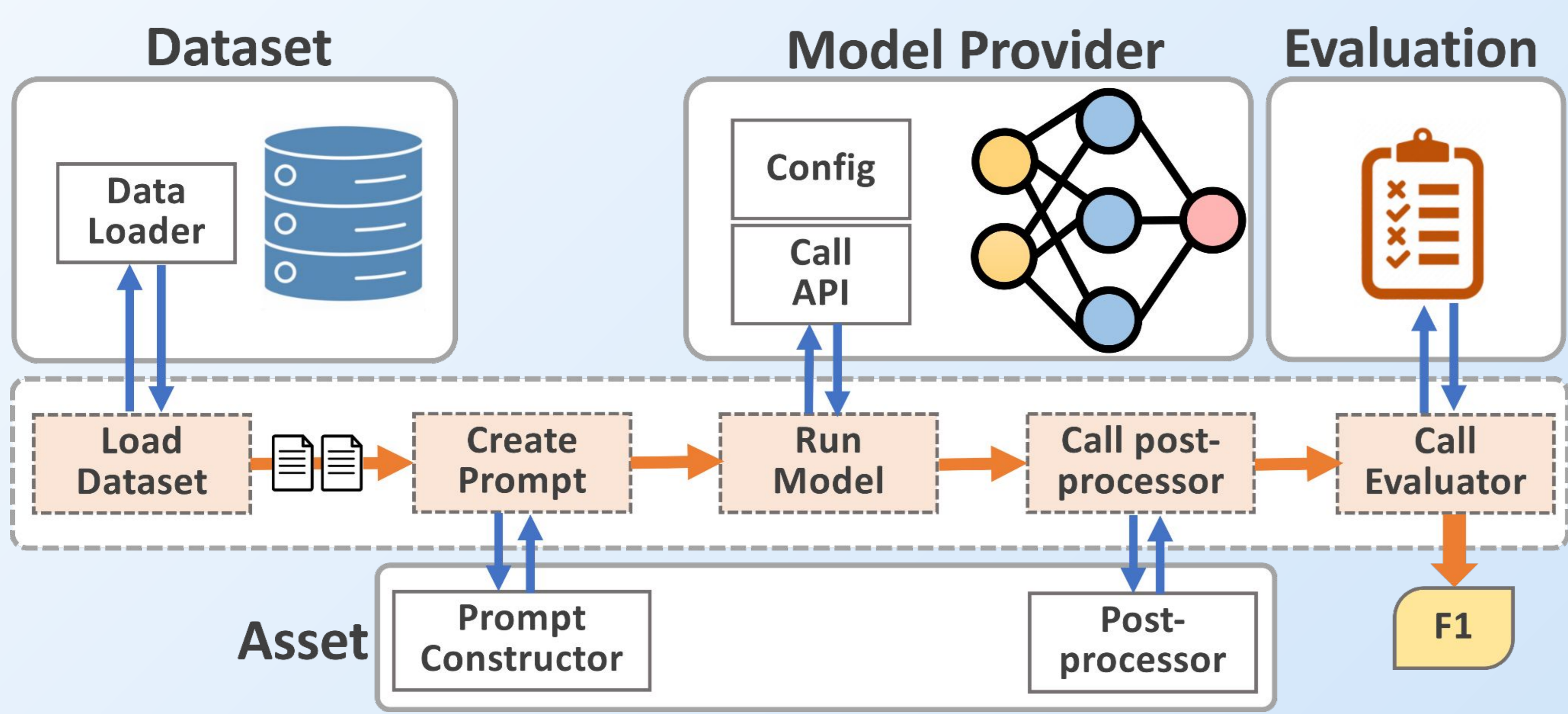


LLMeBench


A Flexible Framework for Accelerating LLMs Benchmarking

Fahim Dalvi Maram Hasanain Sabri Boughorbel Basel Mousi Samir Abdaljalil
 Nizi Nazar Ahmed Abdelali Shammur Absar Chowdhury
 Hamdy Mubarak Ahmed Ali Majd Hawasly Nadir Durrani Firoj Alam
 {faimaduddin, fialam}@hbku.edu.qa

LLMeBench is a language-agnostic large language benchmarking framework that has been used to evaluate over 50 tasks and datasets across a wide range of LLMs



Getting started



<https://llmebench.qcri.org>
 Check out the docs, examples, tutorials, source code and more!

What can you do with LLMeBench?

Benchmarking suite

Create a suite of tasks and datasets and track a model's progress across all of them from a central place

Exploration

Try a model with different prompts over the same data and see how it responds


Model comparison

Run the same prompt with multiple models (say base and fine-tuned) to track performance and progress

Many More...

The framework is flexible and extensible for specific needs

Features

Popular model providers OpenAI, Petals, FastChat	 Hugging Face Datasets, Inference API, hosted models through FastChat	Caching Minimizes expensive API calls and cache all intermediate results	~300 assets across 12 languages Battle tested with over 300k data samples processed	Automatic Dataset downloading
All standard task types Classification, regression, sequence tagging	Standard evaluation metrics	Flexible learning setups Zero- and Few-shot	Major NLP datasets XNLI, XGLUE, XQuAD, ASAD, Aqmar, SANAD, MADAR, QASR, WikiNews, ANERcorp, ASND, CheckThat!, TyDi QA	

Flexibility and Extensibility

Models Extensible with custom model provider implementations	Datasets Allows implementation of new dataset loaders and customization of data input/output formats	Tasks Supports extending tasks types	Contributions welcome! LLMeBench is actively maintained with many community contributions and over 2k clones. Help us improve benchmarking for all!
--	--	--	---