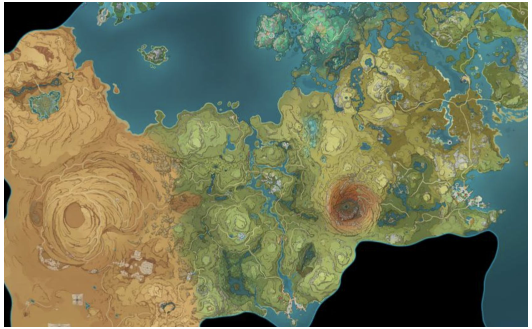
Paimon: your Genshin Impact Chatbot

Team 6 - TaiGen
Jonathan Lin
Han Hung, Chen
Chia Hsin, Wang
YenPing, Chang
Coach - Tanmay Patil

Problem & solution

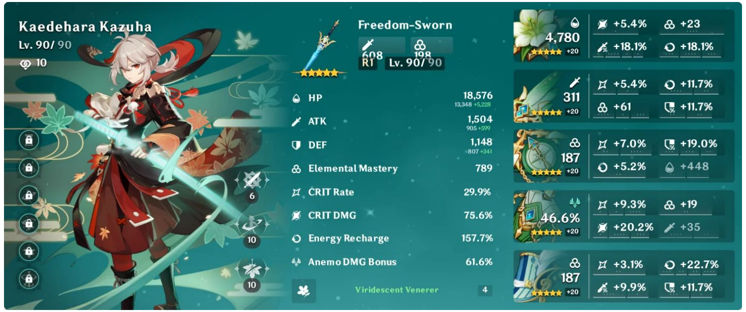
The Challenge: Game Complexity

Vast world map



Abundance of items

Character builds



Kaedehara Kazuha
Lv. 90/ 90
10

Freedom-Sworn
608 198
R1 Lv. 90/ 90

HP	18,576	13,348 +5,228
ATK	1,504	905 +599
DEF	1,148	807 +341
Elemental Mastery	789	
CRIT Rate	29.9%	
CRIT DMG	75.6%	
Energy Recharge	157.7%	
Anemo DMG Bonus	61.6%	

Viridescent Venerer 4

4,780 +20	+5.4%	+23	+18.1%
	+18.1%	+18.1%	

- 4,780 +20: +5.4%, +23, +18.1%, +18.1%
- 311 +20: +5.4%, +11.7%, +61, +11.7%
- 187 +20: +7.0%, +19.0%, +5.2%, +448
- 46.6% +20: +9.3%, +19, +20.2%, +35
- 187 +20: +3.1%, +22.7%, +9.9%, +11.7%

Proposal: AI Chatbot

1

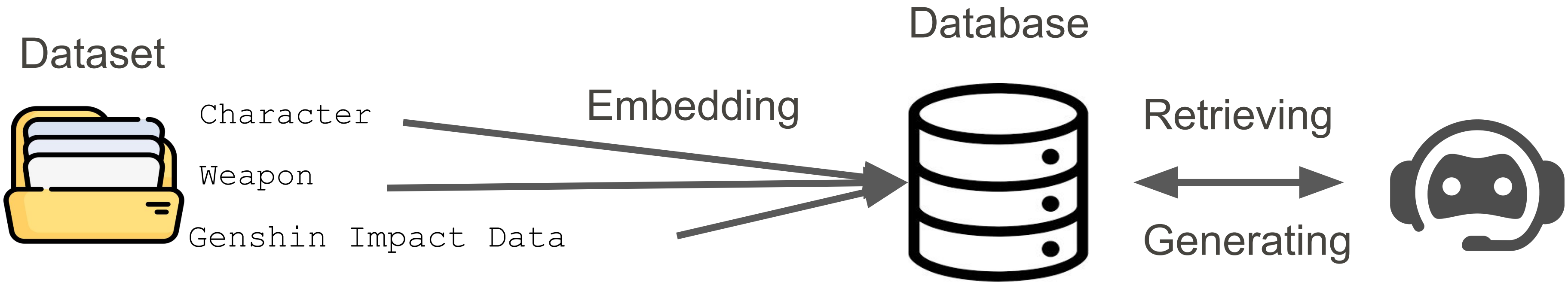
Instant Answers
Get answers to your questions. Paim0n provides quick support.

2

Detailed Information
Access comprehensive game details. Learn about characters, weapons, bosses, wishing, map information, and more.

AI ChatBot Features

RAG



Dataset

Character

Weapon

Genshin Impact Data

Embedding

Database

Retrieving

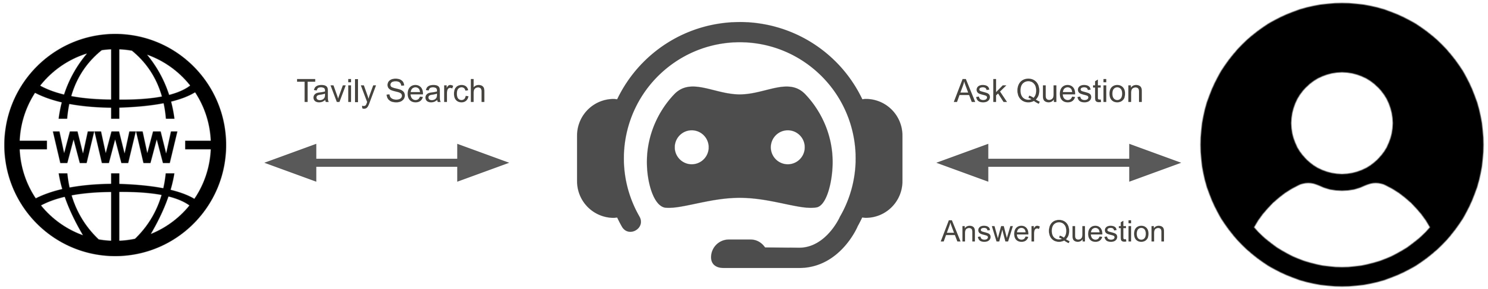
Generating

Tavily Search

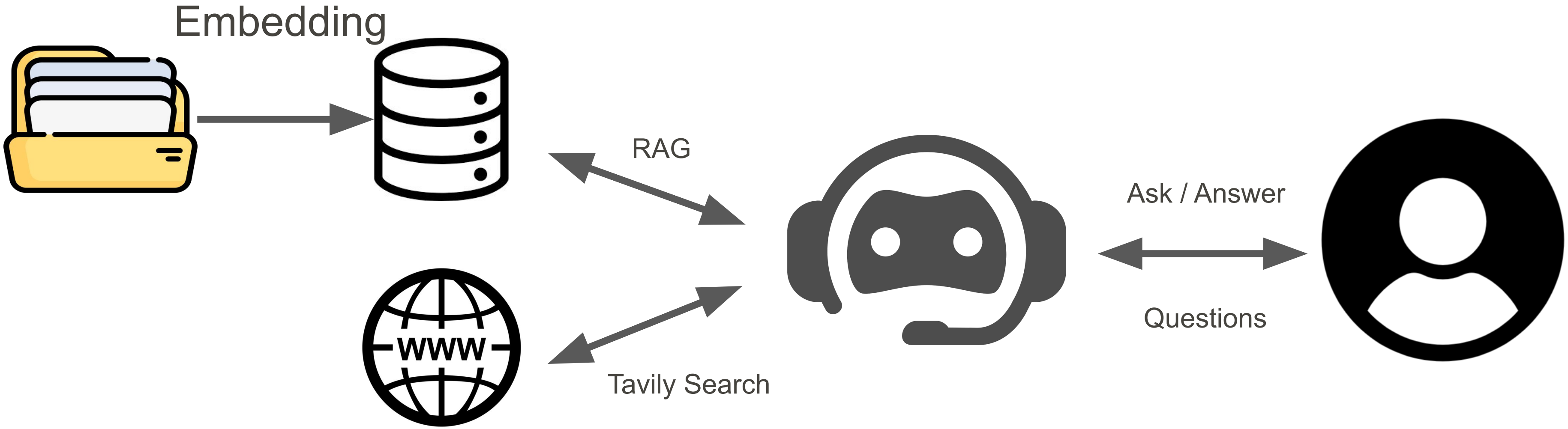
Tavily is a search engine optimized for LLMs.

Tavily focuses on optimizing search for AI developers and AI agents.

Aimed at searching, scraping, filtering and extracting the most relevant information from online sources.



Structure



Live Demo

Example use cases

Example questions:

How do I beat Stormterror? (A boss)

How does the resin system work?

What is a best build for Amber for damage?

When do Furina and traveler talk about Macaroni?

Evaluation

Utilizing Gemini to identify the top 10 most frequently asked questions by Genshin Impact players.

1. Evaluation Method

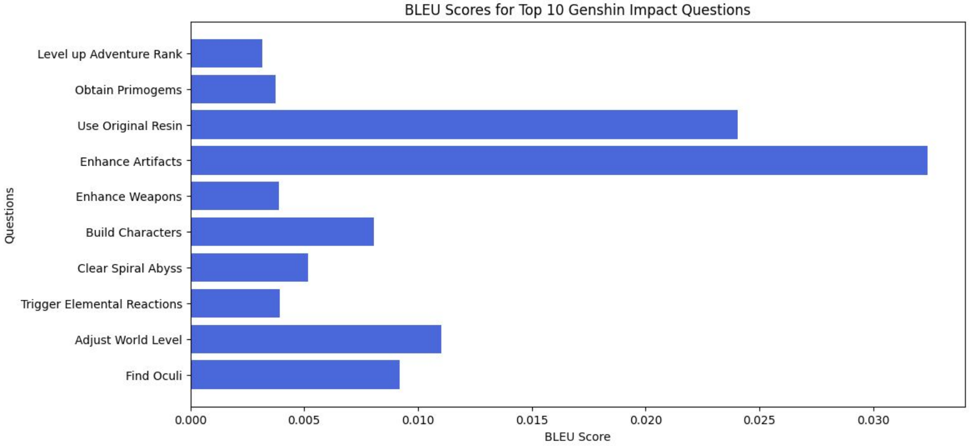
(1) BLEU Score (Bilingual Evaluation Understudy)

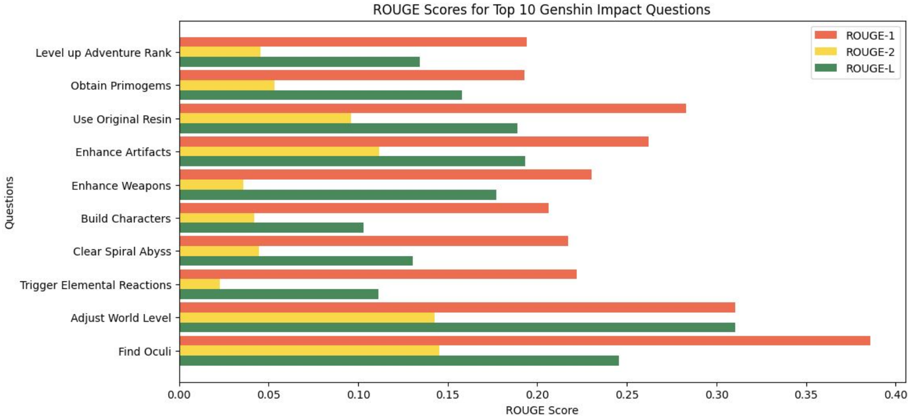
- **Purpose**: Measures the similarity between AI-generated responses and reference answers, primarily used in machine translation and summarization evaluations.
- **Calculation**: Compares n-gram (sequential word segments) matches between AI-generated responses and reference answers, applying smoothing techniques for better handling of short texts.
- **Score Range**: 0 to 1, where a score closer to 1 indicates higher similarity between AI responses and reference answers.

(2) ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)

- **Purpose**: Evaluates recall in AI-generated responses, commonly used in text summarization and generation tasks.
- **Evaluation Metrics**:
  - **ROUGE-1**: Measures unigram (single-word) overlap between AI responses and reference answers.
  - **ROUGE-2**: Measures bigram (two-word phrase) overlap between AI responses and reference answers.
  - **ROUGE-L**: Based on the Longest Common Subsequence (LCS), assessing the syntactic structure and fluency of AI responses.

Evaluation





BLEU Score Analysis

- **Overall low scores** (highest at 0.0324), indicating that AI responses have low lexical similarity to reference answers.
- **Highest BLEU Score:** *"How to choose and enhance Artifacts?"* (BLEU: 0.0324).
- **Lowest BLEU Score:** *"How to level up Adventure Rank quickly?"* (BLEU: 0.0032), suggesting significant wording differences between AI responses and reference answers.
- **Possible Cause:** AI may paraphrase responses differently, failing to align precisely with reference wording and syntax.

ROUGE Score Analysis

- **ROUGE-1 scores are relatively high** (0.2 ~ 0.38), meaning AI responses have **good single-word matching** with reference answers.
- **ROUGE-2 scores are generally low** (mostly below 0.1), showing that AI struggles with **accurate phrase-level matching**, likely due to flexible phrasing.
- **ROUGE-L scores are relatively stable** (0.1 ~ 0.31), indicating **some structural similarity** between AI responses and reference answers.

Conclusion

Low BLEU Scores → AI responses have low lexical similarity to reference answers, indicating possible excessive paraphrasing.

- **Higher ROUGE-1 & ROUGE-L Scores** → AI responses show good single-word and structural similarity, but weak phrase-level matching (ROUGE-2).
- **AI struggles with technical explanations and structured responses**, especially for complex questions, affecting accuracy.
- **Suggested Improvements**:
  1. **Enhance phrase-level matching** (Improve ROUGE-2)
  2. **Reduce excessive paraphrasing** (Increase BLEU scores)
  3. **Incorporate domain-specific data** (Improve AI response accuracy)

➡ **AI needs to refine word alignment and expression to improve response accuracy and readability.**

Any Questions?