Building RAG Q&A Bots for I/O Psychologists

A step-by-step tutorial

SIOP ANNUAL CONFERENCE
DENVER, COLORADO • April 2-5, 2025

AON

SHL.

Introductions

Yuyun Huang

ML / NLP Expert

AON

Karim Badr

Senior Research Scientist

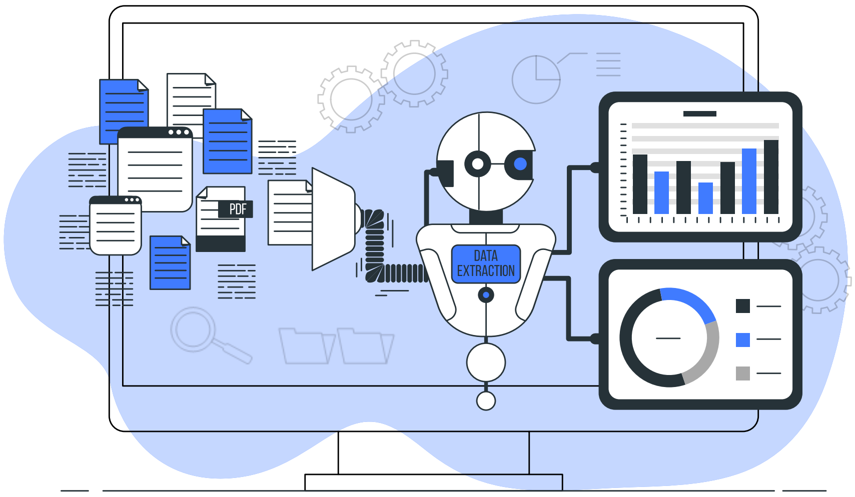
SHL.

Paul Agnello

Senior Associate Consultant

AON

Agenda



Whova

The Basics – What is a RAG Chatbot?

The Technicals – How to Build your own Chatbot?

The Practicals – Hands-On Demos

Chatbots for I/O Psychology

- ***Industry/Practitioner:***
  - Employee support and HR assistance
  - Policy and compliance queries
  - You can upload documents (e.g., policies or employee manuals) and ask questions about them.
- ***Academia/Student/Researcher:***
  - Asking questions about research articles
- ***Industry/Practitioner:***
  - Analyze employee surveys
  - Extract sentiments from comments
  - Analyze data
- ***Other:***
  - Real-time Coaching
  - Simulations of interactions with clients/employees/etc.

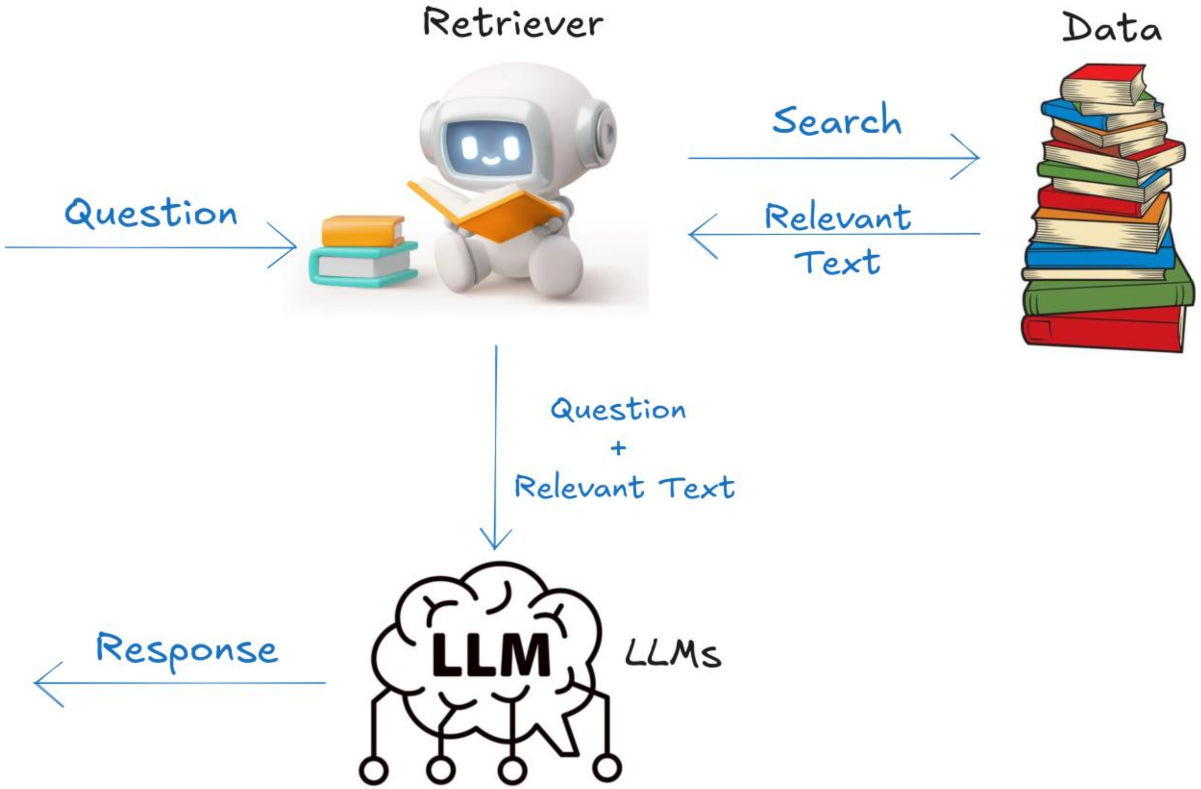
RAG: An Overview

RAG:

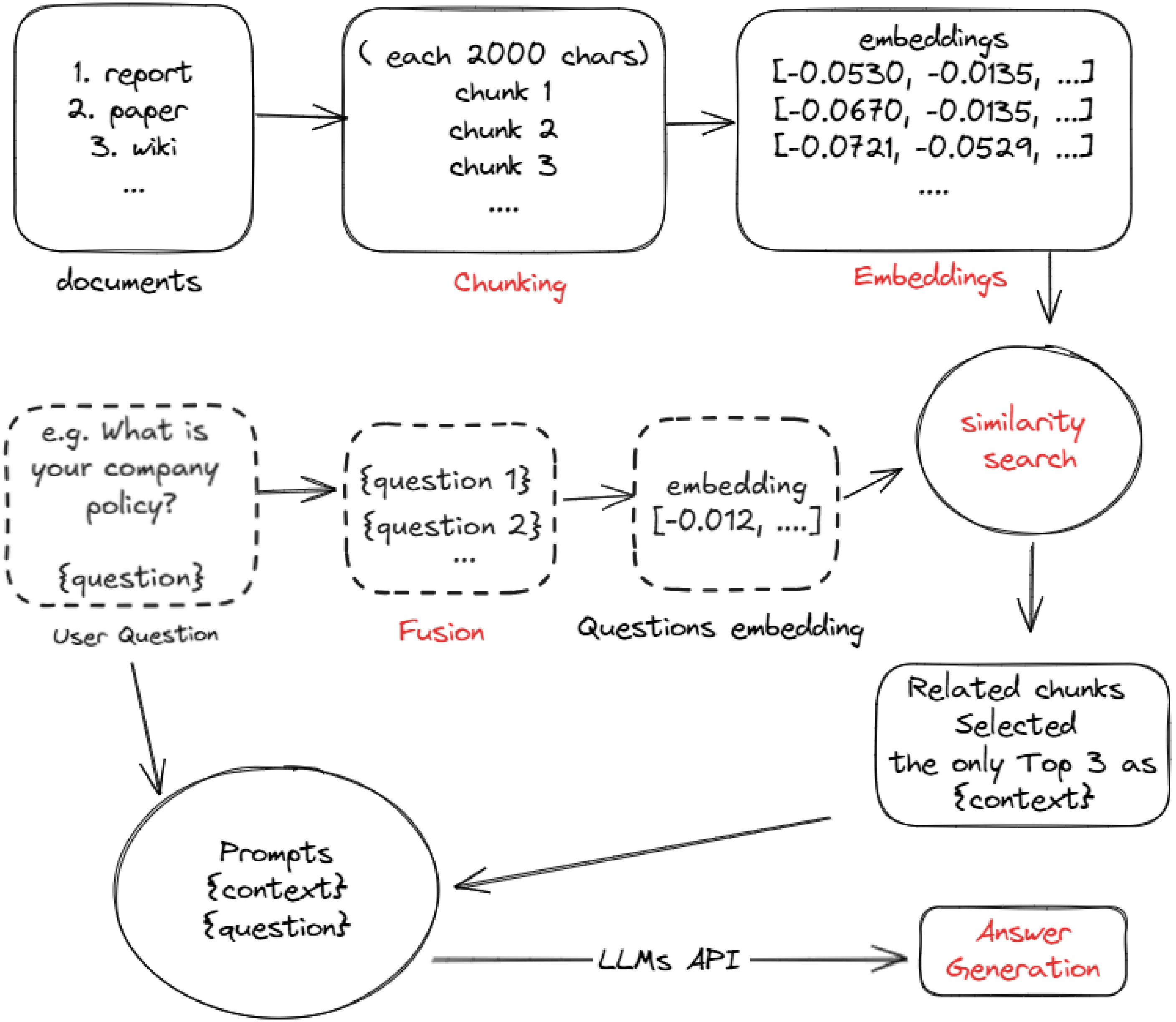
Retrieval-Augmented Generation

It’s a way for AI to give better answers by first searching for useful info in a database (e.g., books, websites) and then using that info to make a clear and smart response.

It’s like an assistant who checks the facts before talking!



RAG:
The Process



1. report
2. paper
3. wiki
...

documents

(each 2000 chars)
chunk 1
chunk 2
chunk 3
....

Chunking

embeddings
[-0.0530, -0.0135, ...]
[-0.0670, -0.0135, ...]
[-0.0721, -0.0529, ...]
....

Embeddings

similarity search

e.g. What is your company policy?

{question}

User Question

{question 1}
{question 2}
...

Fusion

embedding
[-0.012, ...]

Questions embedding

Related chunks Selected the only Top 3 as {context}

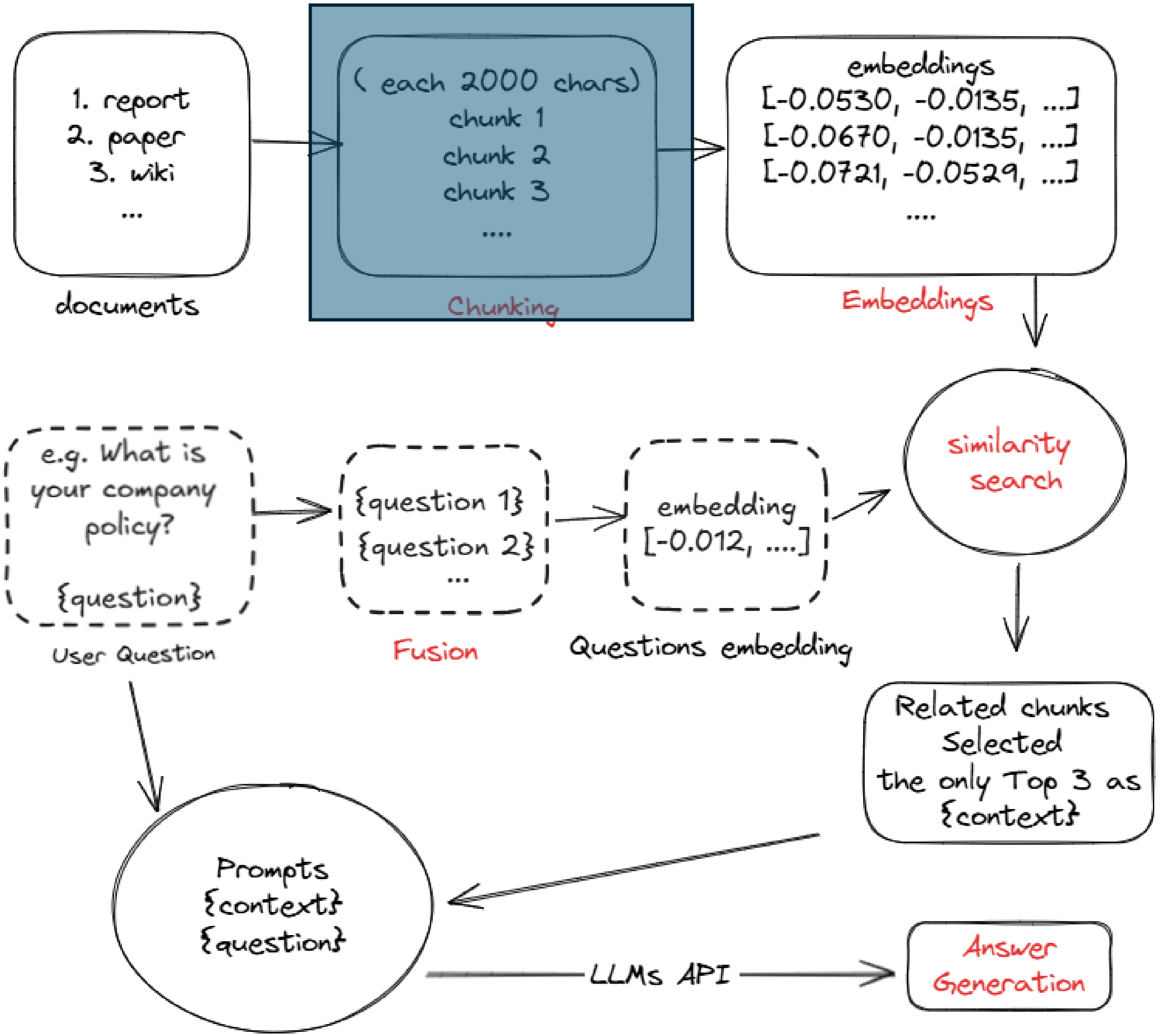
Prompts
{context}
{question}

LLMs API

Answer Generation

RAG:
The Process

Chunking

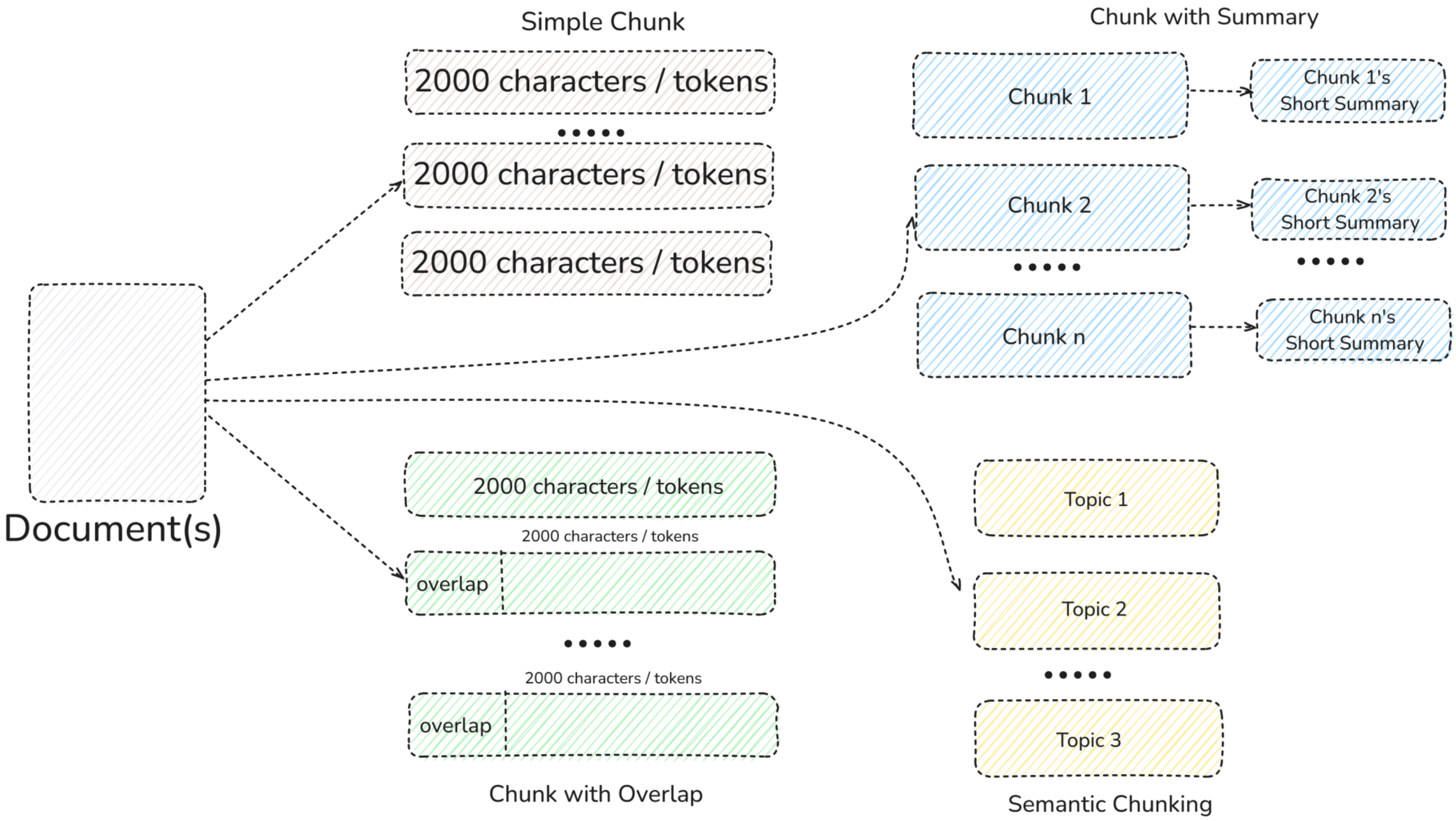


Chunking

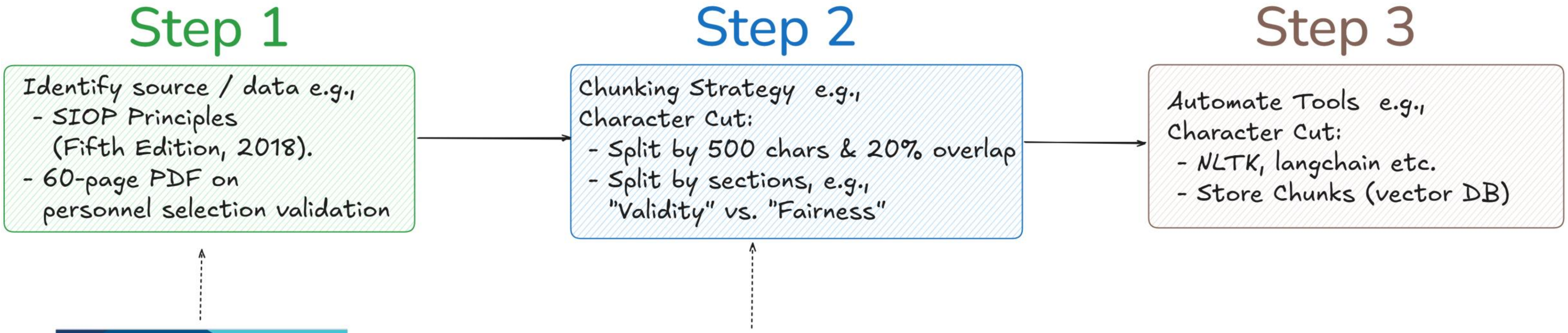
- Definition
  - The process of breaking down large documents or datasets into smaller, manageable pieces (chunks) for the retriever to process.
- Types
  1. **Simple Chunk** (Character / Token cut, Sentence Cut)
  2. **Overlap Chunking**: Add overlapping text between chunks to preserve context (e.g., 20% overlap).
  3. **Chunk with Summary**: Creates chunks and generates a summary for each chunk to preserve context and aid retrieval.
  4. **Semantic Chunking**: Use NLP to split based on meaning (e.g., topic shifts).

Tip: Experiment with chunking strategies based on your data and use case!

Chunking


Simple Chunk
2000 characters / tokens
•••••
2000 characters / tokens
2000 characters / tokens
Chunk with Summary
Chunk 1
Chunk 1's Short Summary
Chunk 2
Chunk 2's Short Summary
•••••
•••••
Chunk n
Chunk n's Short Summary
Document(s)
2000 characters / tokens
2000 characters / tokens
overlap
•••••
2000 characters / tokens
overlap
Chunk with Overlap
Topic 1
Topic 2
•••••
Topic 3
Semantic Chunking

Creating Chunks



Step 1

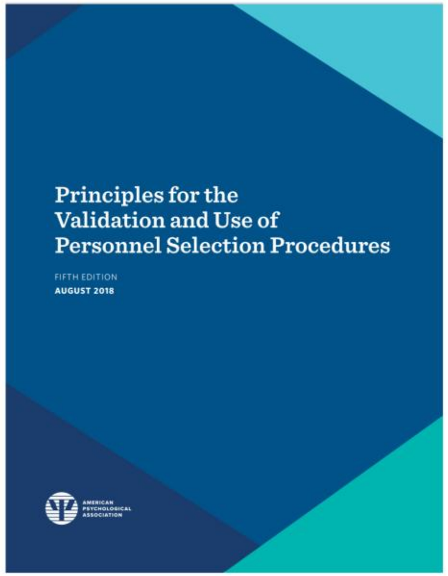
Identify source / data e.g.,
- SIOP Principles (Fifth Edition, 2018).
- 60-page PDF on personnel selection validation

Step 2

Chunking Strategy e.g., Character Cut:
- Split by 500 chars & 20% overlap
- Split by sections, e.g., "Validity" vs. "Fairness"

Step 3

Automate Tools e.g., Character Cut:
- NLTK, langchain etc.
- Store Chunks (vector DB)



Principles for the Validation and Use of Personnel Selection Procedures

FIFTH EDITION

AUGUST 2018

Test:
Can a chunk from

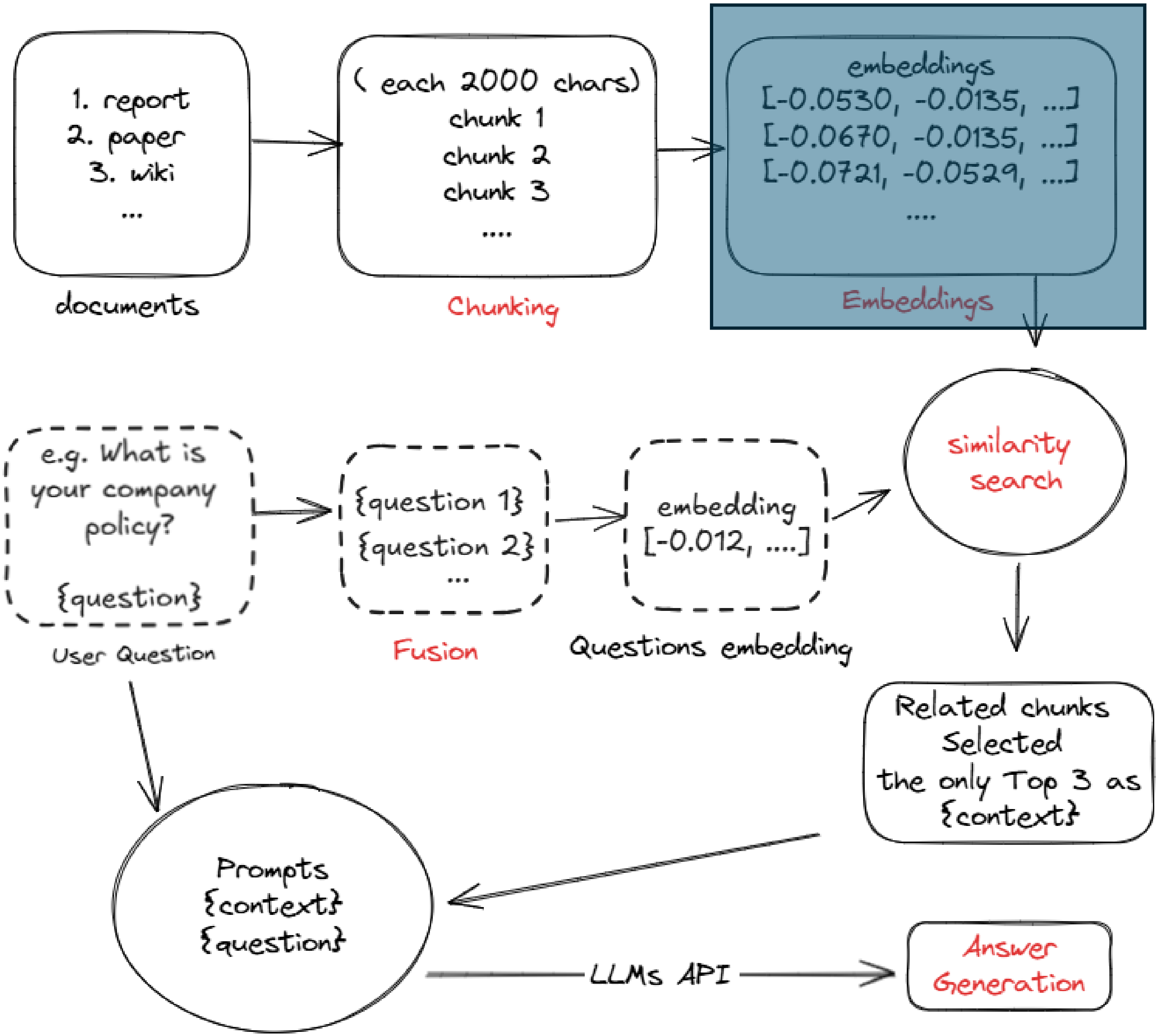
"Fairness and Bias" (p. 27)

answer

"What does SIOP say about predictive bias?

RAG:
The Process

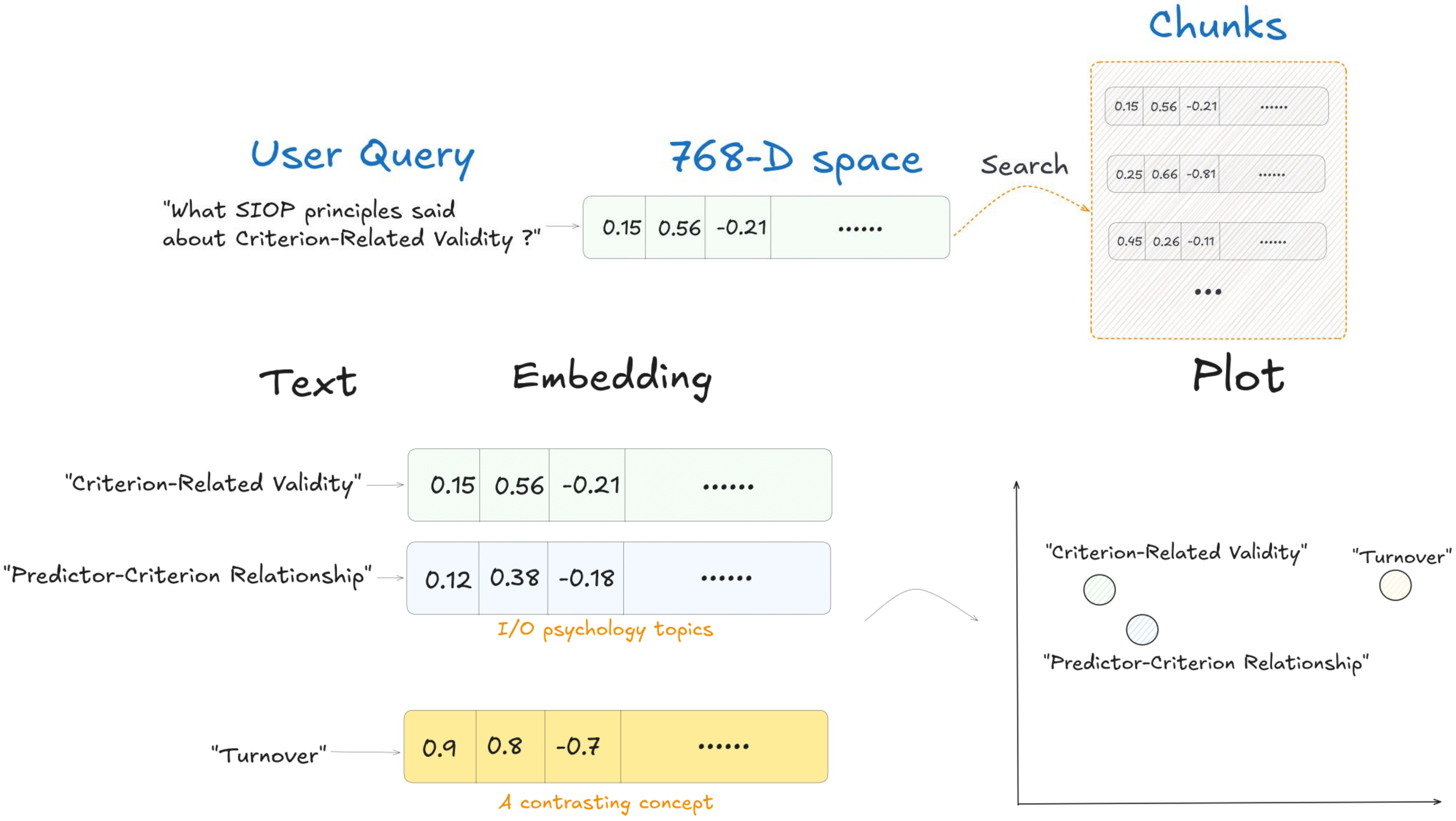
Embeddings



Embeddings

- **Definition**
  - Embeddings are numerical representations (vectors) of text, images, or other data in a high-dimensional space, capturing their meaning or context.
- **Usage in RAG Chatbots**
  - Enables the retriever to understand and compare the similarity between a user’s query and stored chunks.
  - Enables semantic search: e.g., find SIOP sections on “predictive bias” without exact keywords).

Embeddings



User Query

"What SIOP principles said about Criterion-Related Validity ?"

768-D space

0.15	0.56	-0.21	......

Search

Chunks

0.15	0.56	-0.21	......
0.25	0.66	-0.81	......
0.45	0.26	-0.11	......

...

Text	Embedding			
"Criterion-Related Validity"	0.15	0.56	-0.21	......
"Predictor-Criterion Relationship"	0.12	0.38	-0.18	......
"Turnover"	0.9	0.8	-0.7	......

I/O psychology topics

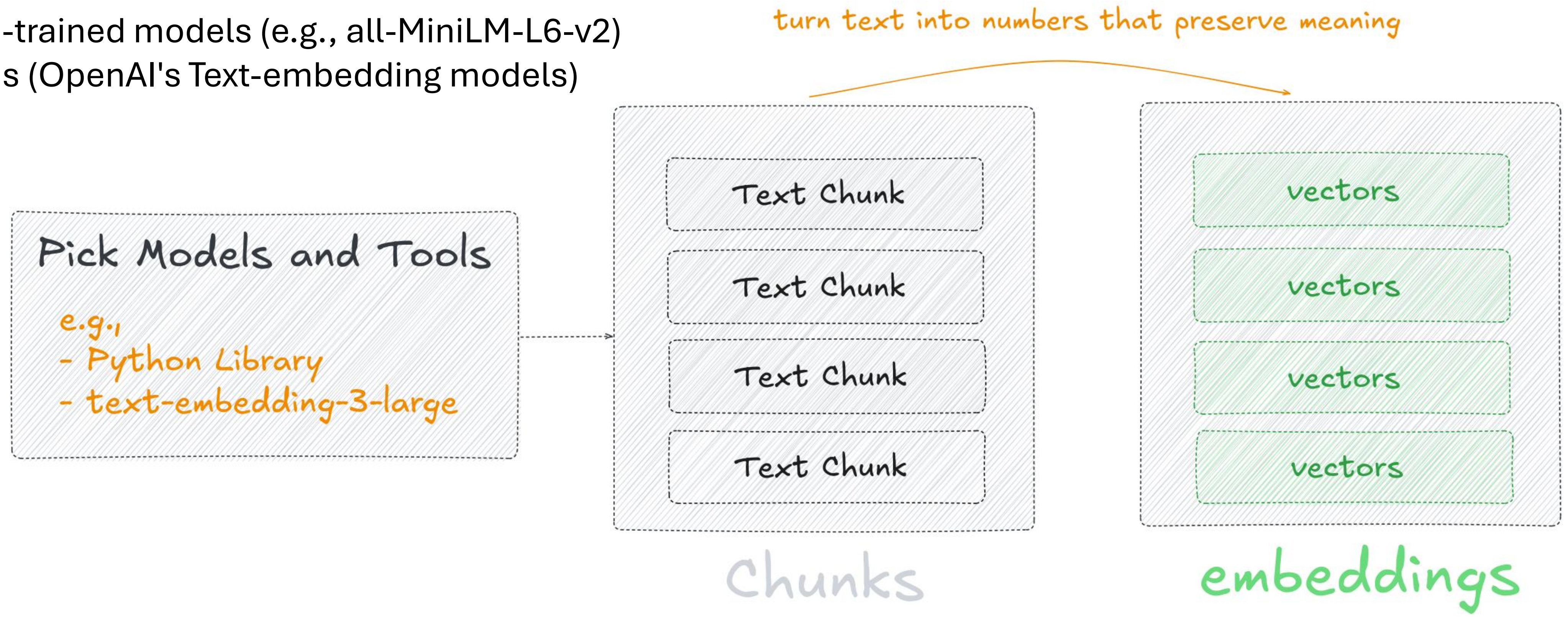
A contrasting concept

Plot

"Criterion-Related Validity"
"Turnover"
"Predictor-Criterion Relationship"

Creating Embeddings

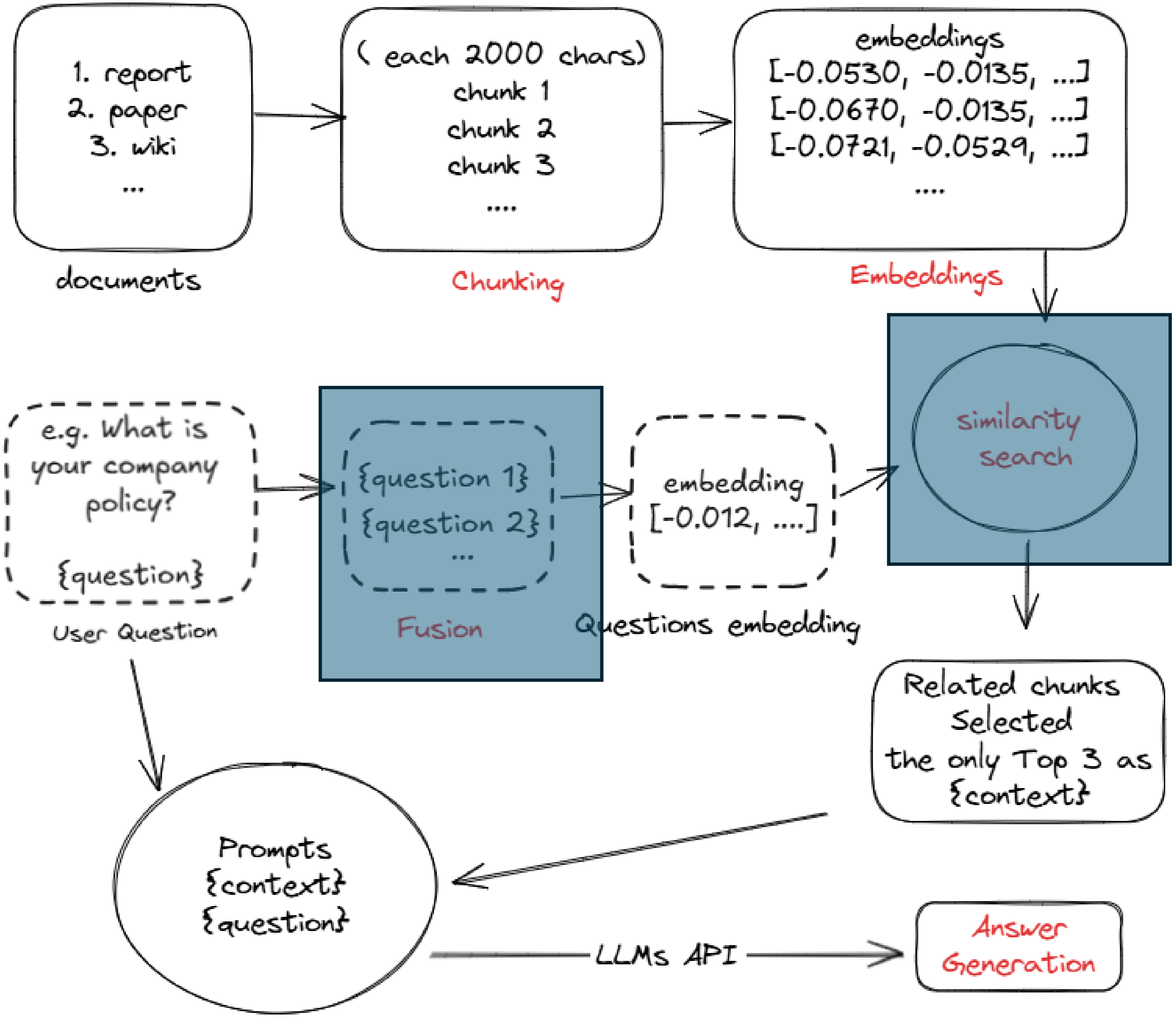
- Pre-trained models (e.g., all-MiniLM-L6-v2)
- APIs (OpenAI's Text-embedding models)



Step 1: Pick a Model **Step 2: Process Chunks** **Step 3: Generate Vectors**

RAG: The Process

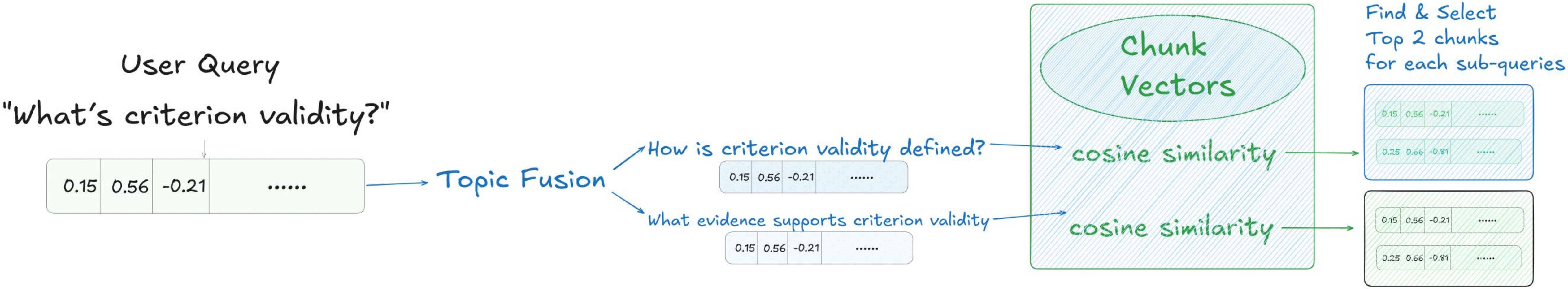
Searching



Searching: Topic Fusion

- **Fuse Topics**: Split query into two richer sub-queries:
  - “What’s criterion validity?” →
    - “How is criterion validity defined?” (targets definitions).
    - “What evidence supports criterion validity?” (targets examples).
- **Embed Queries**: Convert both to vectors
- **Compare Vectors**: Match each to SIOP chunks (e.g., p. 18’s “Evidence for criterion-related validity…”).
- **Rank Results**: Pick top 2 chunks per sub-query (4 total) using cosine similarity.
- **Retrieve**: Return chunks with metadata (e.g., “SIOP Principles,” “p. 18”).

Searching: Topic Fusion

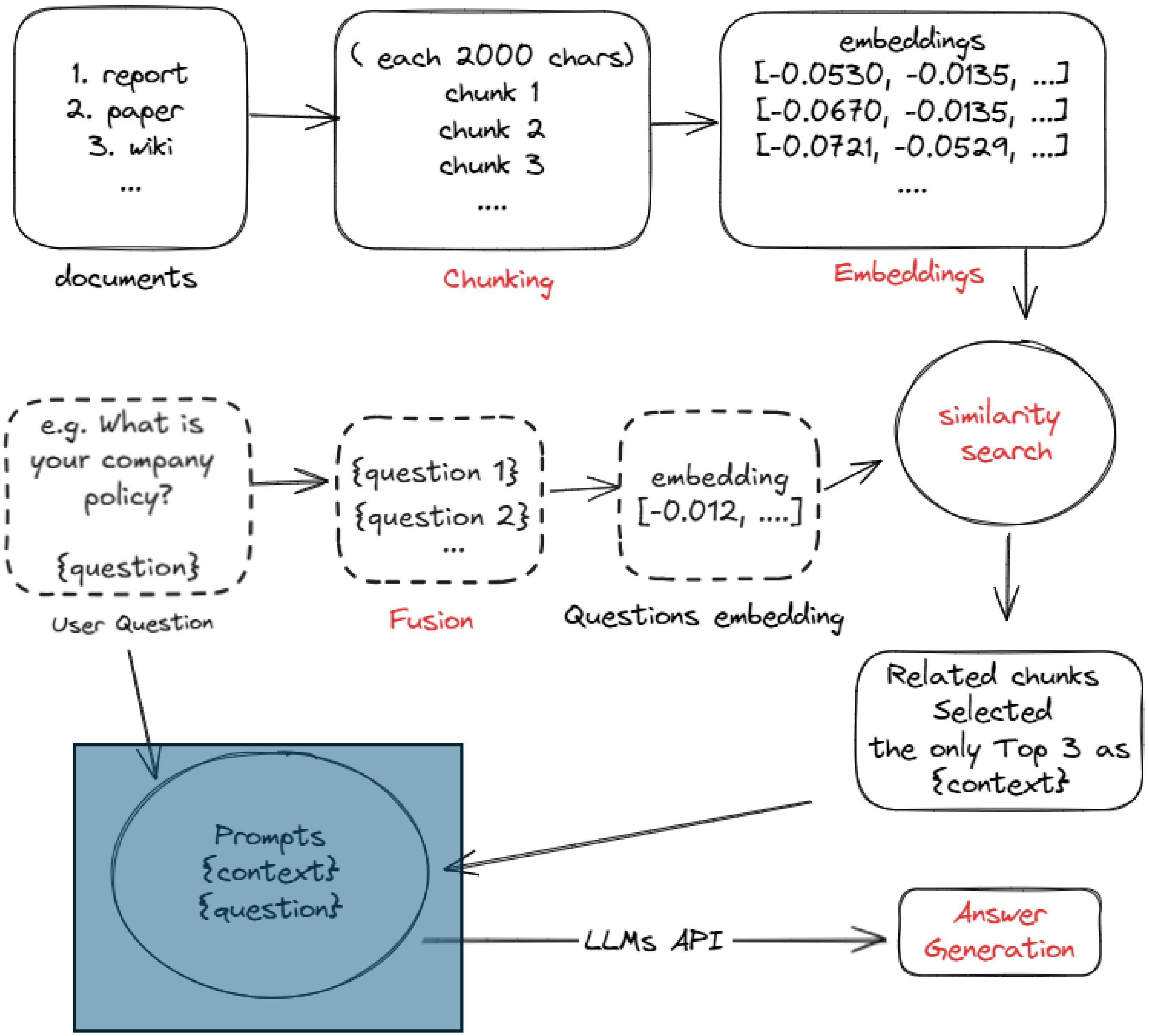


Searching: Cosine Similarity & Vector Search

- **Cosine Similarity**: a measure of how close 2 vectors are
  - 1 = identical
  - 0 = unrelated
- **Example:** Query “What’s criterion validity?” vs. p. 18 chunk “Evidence for criterion-related validity…” → high score (~0.9).
- **Tools**:
  - FAISS (Local) : Fast vector search library
  - Pinecone (Cloud): Scalable vector DB

RAG:
The Process

Prompt Forming



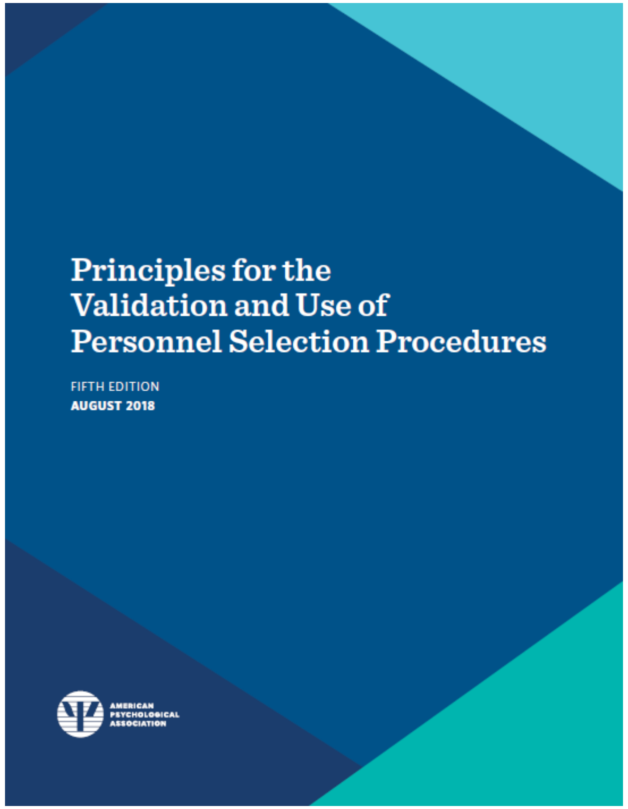
Prompt Forming

- Goal: Combine query + retrieved chunk → generate coherent answer
- Example:

```
Query: {user_query}
Context:
{retrieved_chunk1}
{retrieved_chunk2}
{retrieved_chunk3}
{retrieved_chunk4}
Answer:"
```

Build your own Chatbot – Code

Code demo in Jupyter Notebooks



Principles for the Validation and Use of Personnel Selection Procedures

FIFTH EDITION

AUGUST 2018

AMERICAN PSYCHOLOGICAL ASSOCIATION

Build your own Chatbot: No Code via DIFY

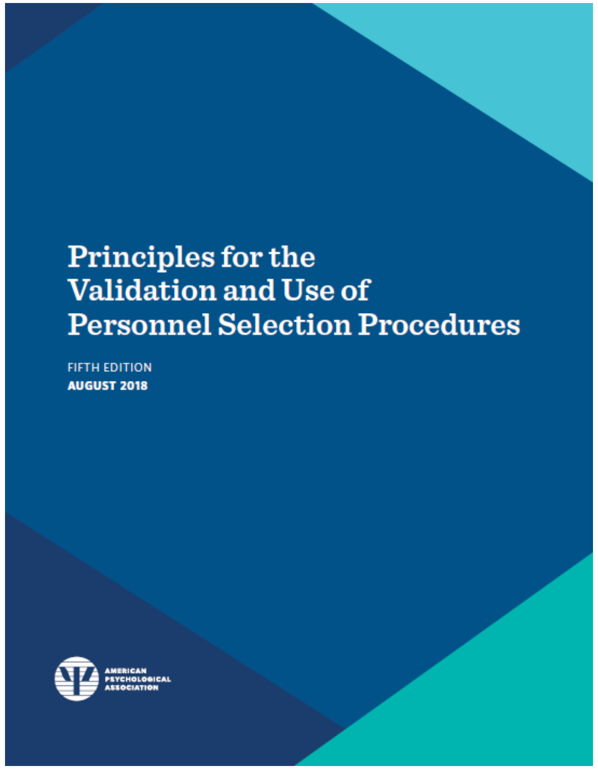
- Many no code tools available
- Example: **DIFY**
  - Online cloud version (free & paid)
  - On-premises (Open-source, install on your server)

Dify

Define & Modify
Do It For You

https://dify.ai/

DIFY Demo

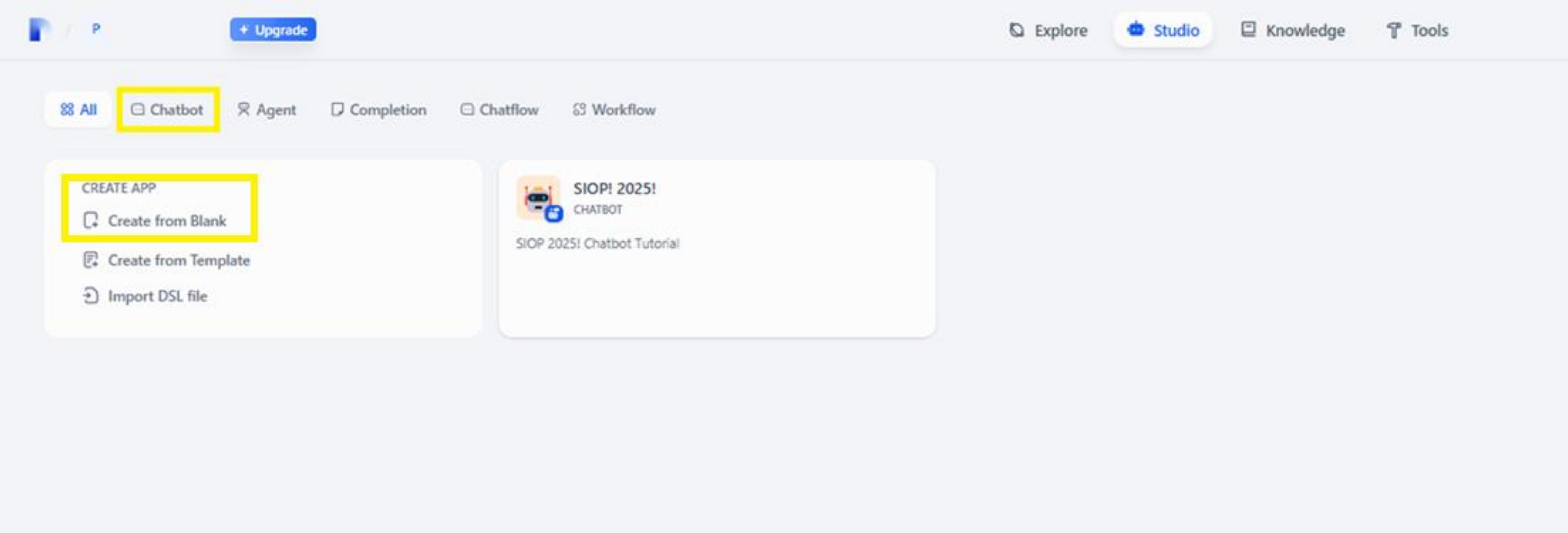


Principles for the
Validation and Use of
Personnel Selection Procedures

FIFTH EDITION
AUGUST 2018

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

DIFY: Landing Page

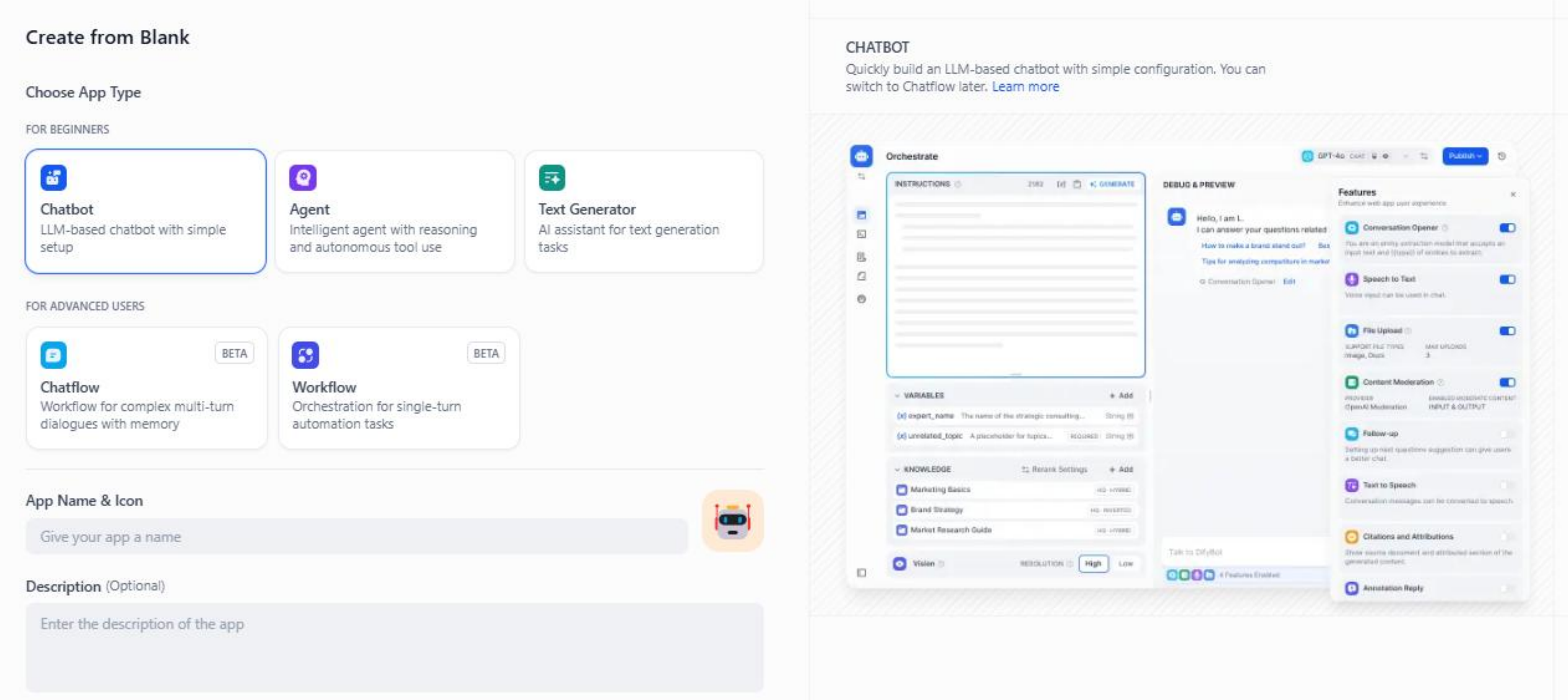


SIOP! 2025!

CHATBOT

SIOP 2025! Chatbot Tutorial

DIFY: Chatbot



Create from Blank

Choose App Type

FOR BEGINNERS

- Chatbot: LLM-based chatbot with simple setup
- Agent: Intelligent agent with reasoning and autonomous tool use
- Text Generator: AI assistant for text generation tasks

FOR ADVANCED USERS

- Chatflow (BETA): Workflow for complex multi-turn dialogues with memory
- Workflow (BETA): Orchestration for single-turn automation tasks

App Name & Icon

Give your app a name

Description (Optional)

Enter the description of the app

CHATBOT

Quickly build an LLM-based chatbot with simple configuration. You can switch to Chatflow later. Learn more

Orchestrate

INSTRUCTIONS 2182 GENERATE

VARIABLES + Add

- {x} expert_name: The name of the strategic consulting... (String)
- {x} unrelated_topic: A placeholder for topics... (REQUIRED, String)

KNOWLEDGE: Rerank Settings + Add

- Marketing Basics (HQ · HYBRID)
- Brand Strategy (HQ · INVERTED)
- Market Research Guide (HQ · HYBRID)

Vision: RESOLUTION High Low

GPT-4o CHAT Publish

DEBUG & PREVIEW

Hello, I am L.
I can answer your questions related

How to make a brand stand out?

Tips for analyzing competitors in market

Conversation Opener Edit

Talk to DifyBot

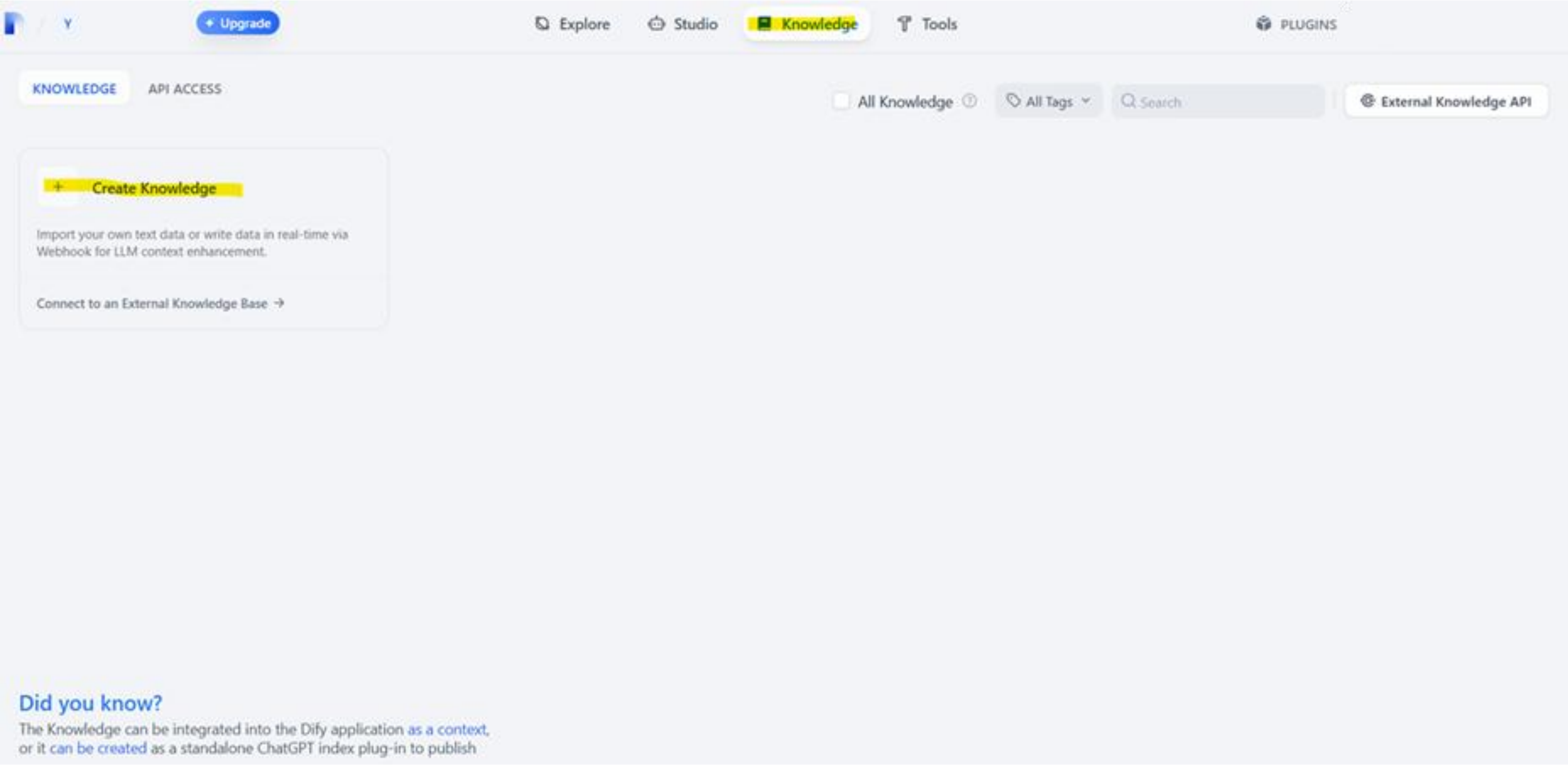
4 Features Enabled

Features

Enhance web app user experience

- Conversation Opener: You are an entity extraction model that accepts an input text and {{type}} of entities to extract.
- Speech to Text: Voice input can be used in chat.
- File Upload: SUPPORT FILE TYPES Image, Docs; MAX UPLOADS 3
- Content Moderation: PROVIDER OpenAI Moderation; ENABLED MODERATE CONTENT INPUT & OUTPUT
- Follow-up: Setting up next questions suggestion can give users a better chat.
- Text to Speech: Conversation messages can be converted to speech.
- Citations and Attributions: Show source document and attributed section of the generated content.
- Annotation Reply

DIFY: Create Knowledge



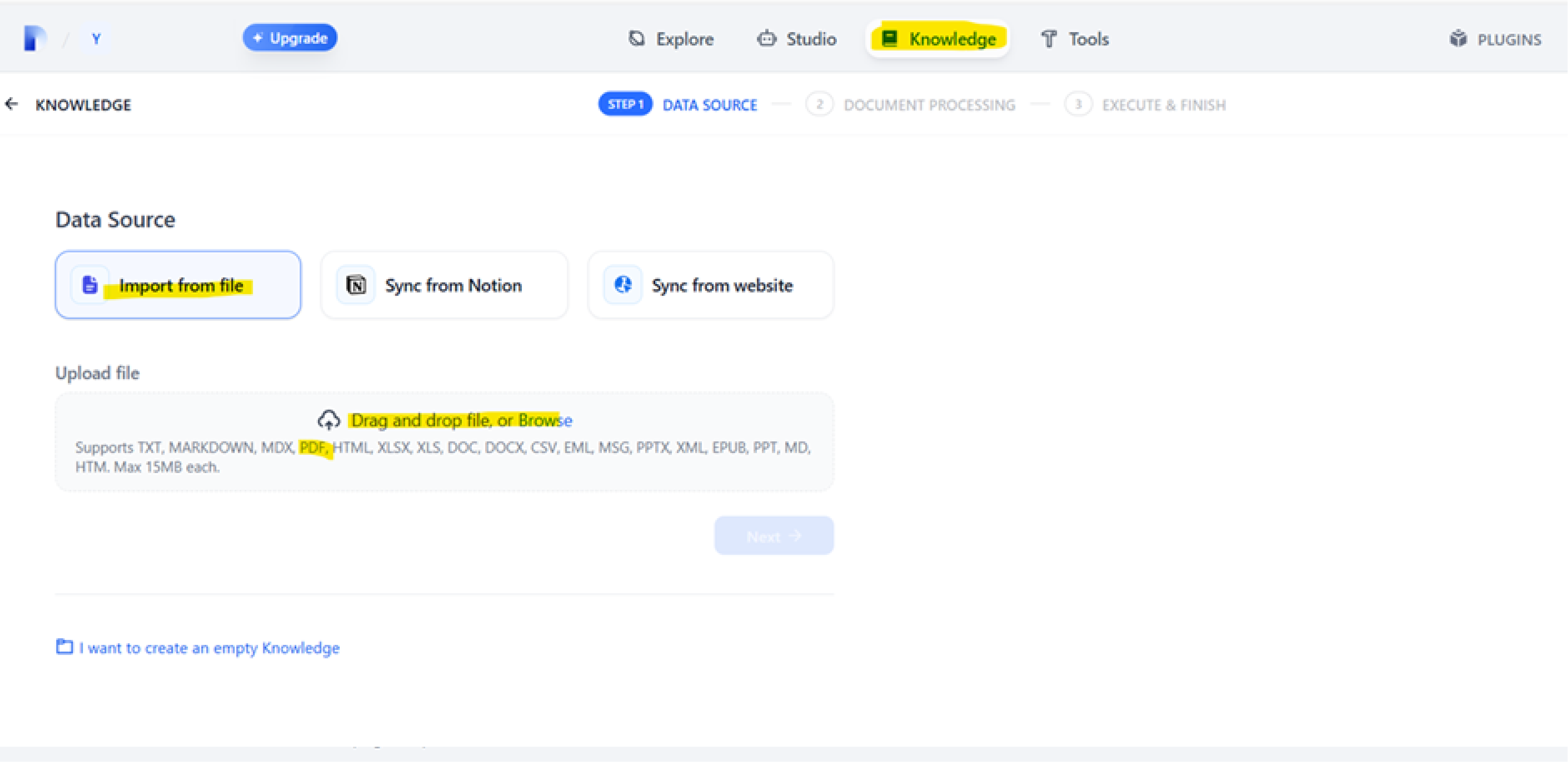
Import your own text data or write data in real-time via Webhook for LLM context enhancement.

Connect to an External Knowledge Base →

Did you know?

The Knowledge can be integrated into the Dify application as a context, or it can be created as a standalone ChatGPT index plug-in to publish

DIFY: Upload Files

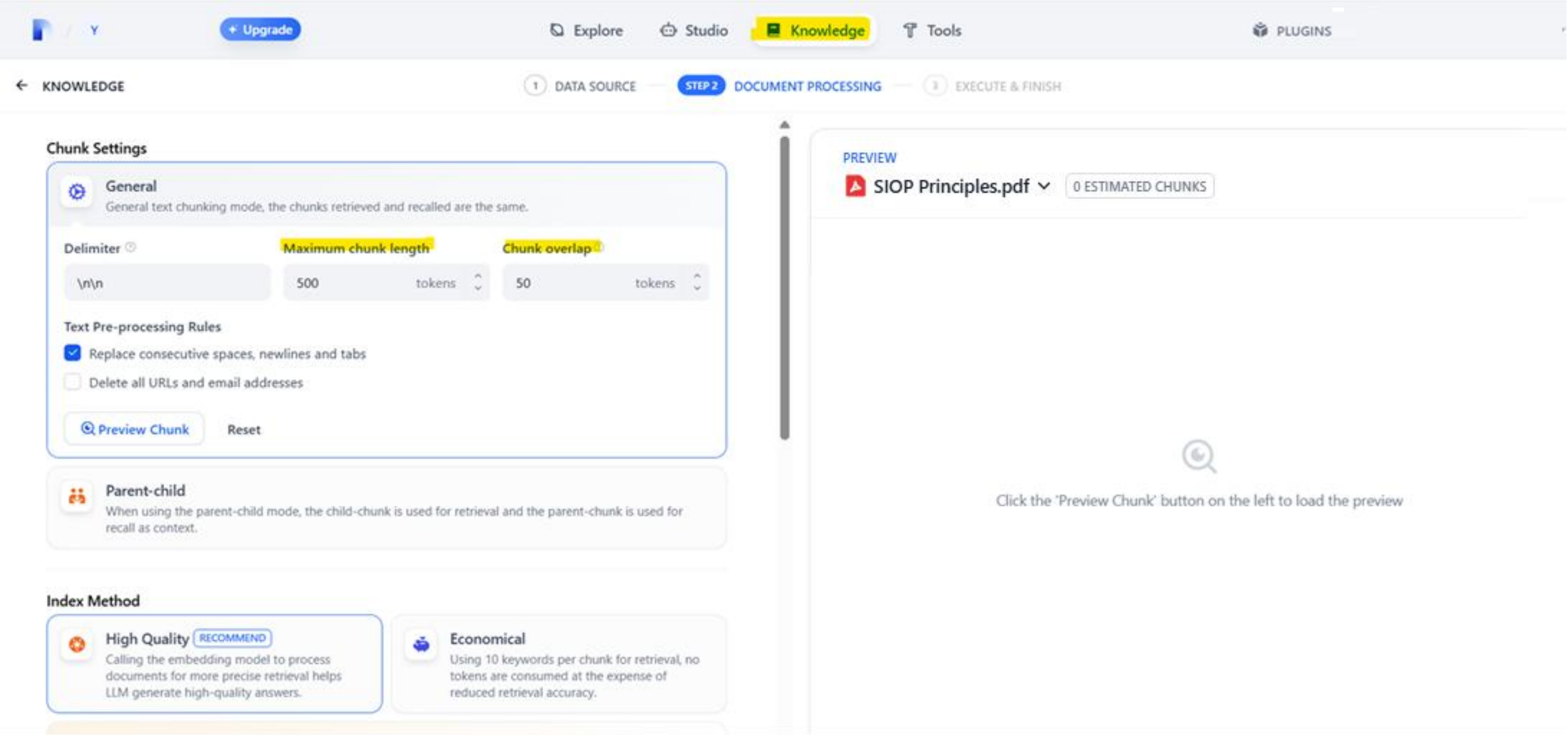


Drag and drop file, or Browse

Supports TXT, MARKDOWN, MDX, PDF, HTML, XLSX, XLS, DOC, DOCX, CSV, EML, MSG, PPTX, XML, EPUB, PPT, MD, HTM. Max 15MB each.

I want to create an empty Knowledge

DIFY: Text Chunking



Chunk Settings

General
General text chunking mode, the chunks retrieved and recalled are the same.

Delimiter: \n\n

Maximum chunk length: 500 tokens

Chunk overlap: 50 tokens

Text Pre-processing Rules

☑ Replace consecutive spaces, newlines and tabs

☐ Delete all URLs and email addresses

Parent-child
When using the parent-child mode, the child-chunk is used for retrieval and the parent-chunk is used for recall as context.

Index Method

High Quality RECOMMEND
Calling the embedding model to process documents for more precise retrieval helps LLM generate high-quality answers.

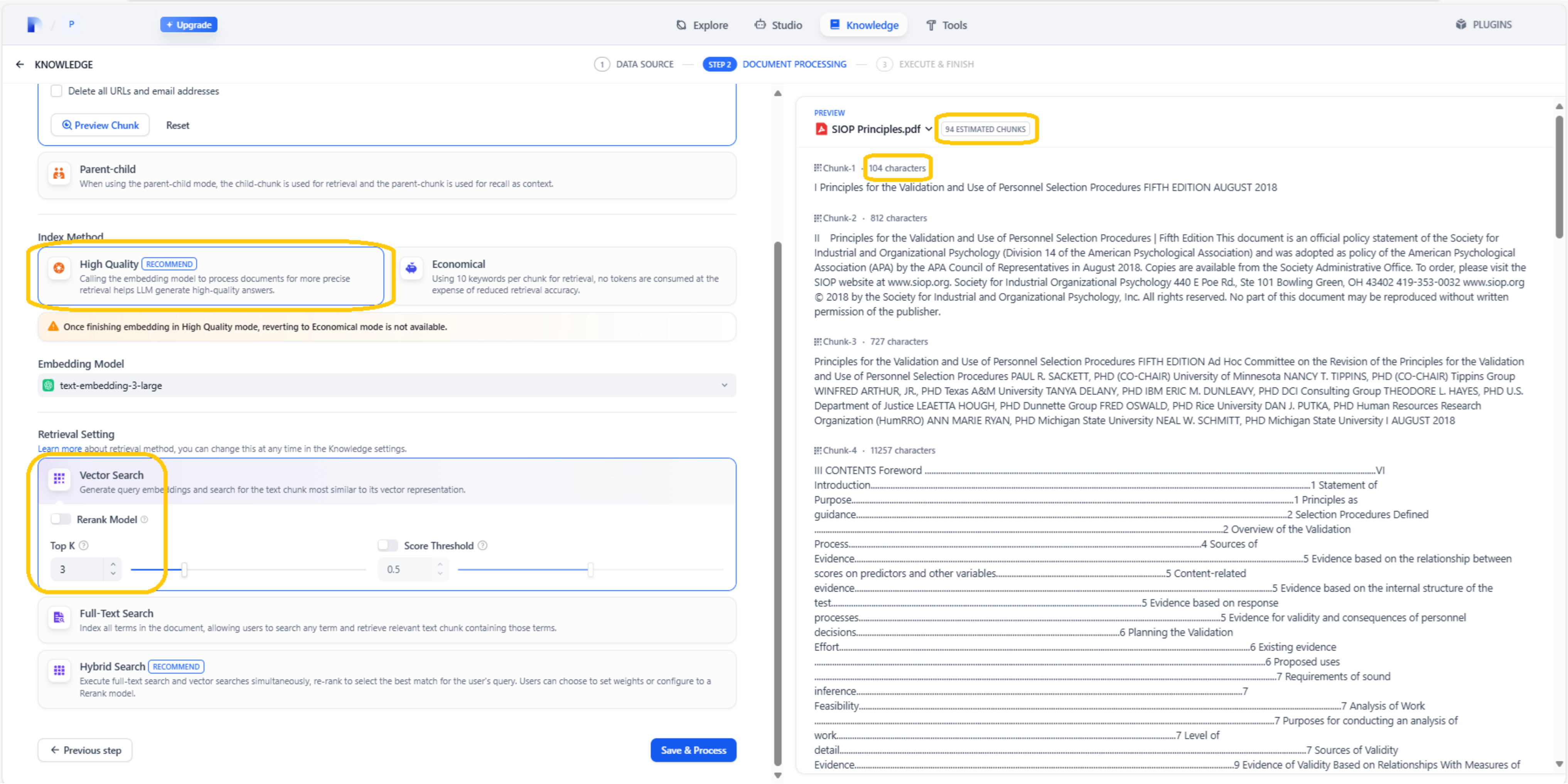
Economical
Using 10 keywords per chunk for retrieval, no tokens are consumed at the expense of reduced retrieval accuracy.

PREVIEW

SIOP Principles.pdf 0 ESTIMATED CHUNKS

Click the 'Preview Chunk' button on the left to load the preview

DIFY: Embedding



□ Delete all URLs and email addresses

Preview Chunk   Reset

Parent-child
When using the parent-child mode, the child-chunk is used for retrieval and the parent-chunk is used for recall as context.

Index Method

High Quality RECOMMEND
Calling the embedding model to process documents for more precise retrieval helps LLM generate high-quality answers.

Economical
Using 10 keywords per chunk for retrieval, no tokens are consumed at the expense of reduced retrieval accuracy.

⚠ Once finishing embedding in High Quality mode, reverting to Economical mode is not available.

Embedding Model
text-embedding-3-large

Retrieval Setting
Learn more about retrieval method, you can change this at any time in the Knowledge settings.

Vector Search
Generate query embeddings and search for the text chunk most similar to its vector representation.

Rerank Model
Top K: 3
Score Threshold: 0.5

Full-Text Search
Index all terms in the document, allowing users to search any term and retrieve relevant text chunk containing those terms.

Hybrid Search RECOMMEND
Execute full-text search and vector searches simultaneously, re-rank to select the best match for the user's query. Users can choose to set weights or configure to a Rerank model.

← Previous step   Save & Process

PREVIEW
SIOP Principles.pdf   94 ESTIMATED CHUNKS

Chunk-1 · 104 characters

I Principles for the Validation and Use of Personnel Selection Procedures FIFTH EDITION AUGUST 2018

Chunk-2 · 812 characters

II Principles for the Validation and Use of Personnel Selection Procedures | Fifth Edition This document is an official policy statement of the Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association) and was adopted as policy of the American Psychological Association (APA) by the APA Council of Representatives in August 2018. Copies are available from the Society Administrative Office. To order, please visit the SIOP website at www.siop.org. Society for Industrial Organizational Psychology 440 E Poe Rd., Ste 101 Bowling Green, OH 43402 419-353-0032 www.siop.org © 2018 by the Society for Industrial and Organizational Psychology, Inc. All rights reserved. No part of this document may be reproduced without written permission of the publisher.

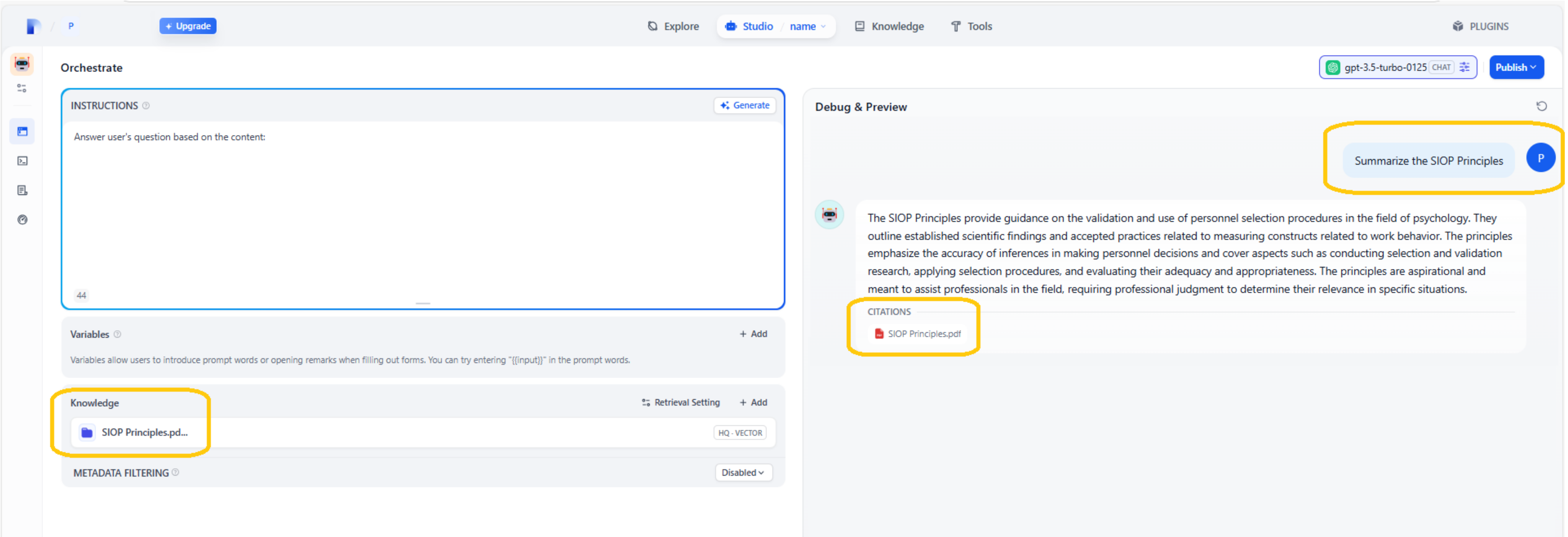
Chunk-3 · 727 characters

Principles for the Validation and Use of Personnel Selection Procedures FIFTH EDITION Ad Hoc Committee on the Revision of the Principles for the Validation and Use of Personnel Selection Procedures PAUL R. SACKETT, PHD (CO-CHAIR) University of Minnesota NANCY T. TIPPINS, PHD (CO-CHAIR) Tippins Group WINFRED ARTHUR, JR., PHD Texas A&M University TANYA DELANY, PHD IBM ERIC M. DUNLEAVY, PHD DCI Consulting Group THEODORE L. HAYES, PHD U.S. Department of Justice LEAETTA HOUGH, PHD Dunnette Group FRED OSWALD, PHD Rice University DAN J. PUTKA, PHD Human Resources Research Organization (HumRRO) ANN MARIE RYAN, PHD Michigan State University NEAL W. SCHMITT, PHD Michigan State University I AUGUST 2018

Chunk-4 · 11257 characters

III CONTENTS Foreword VI
Introduction..........1 Statement of
Purpose..........1 Principles as
guidance..........2 Selection Procedures Defined
..........2 Overview of the Validation
Process..........4 Sources of
Evidence..........5 Evidence based on the relationship between
scores on predictors and other variables..........5 Content-related
evidence..........5 Evidence based on the internal structure of the
test..........5 Evidence based on response
processes..........5 Evidence for validity and consequences of personnel
decisions..........6 Planning the Validation
Effort..........6 Existing evidence
..........6 Proposed uses
..........7 Requirements of sound
inference..........7
Feasibility..........7 Analysis of Work
..........7 Purposes for conducting an analysis of
work..........7 Level of
detail..........7 Sources of Validity
Evidence..........9 Evidence of Validity Based on Relationships With Measures of

DIFY: Debugging & Trialing the Chatbot



INSTRUCTIONS

Answer user's question based on the content:

Variables

Variables allow users to introduce prompt words or opening remarks when filling out forms. You can try entering "{{input}}" in the prompt words.

Knowledge

SIOP Principles.pd...

METADATA FILTERING

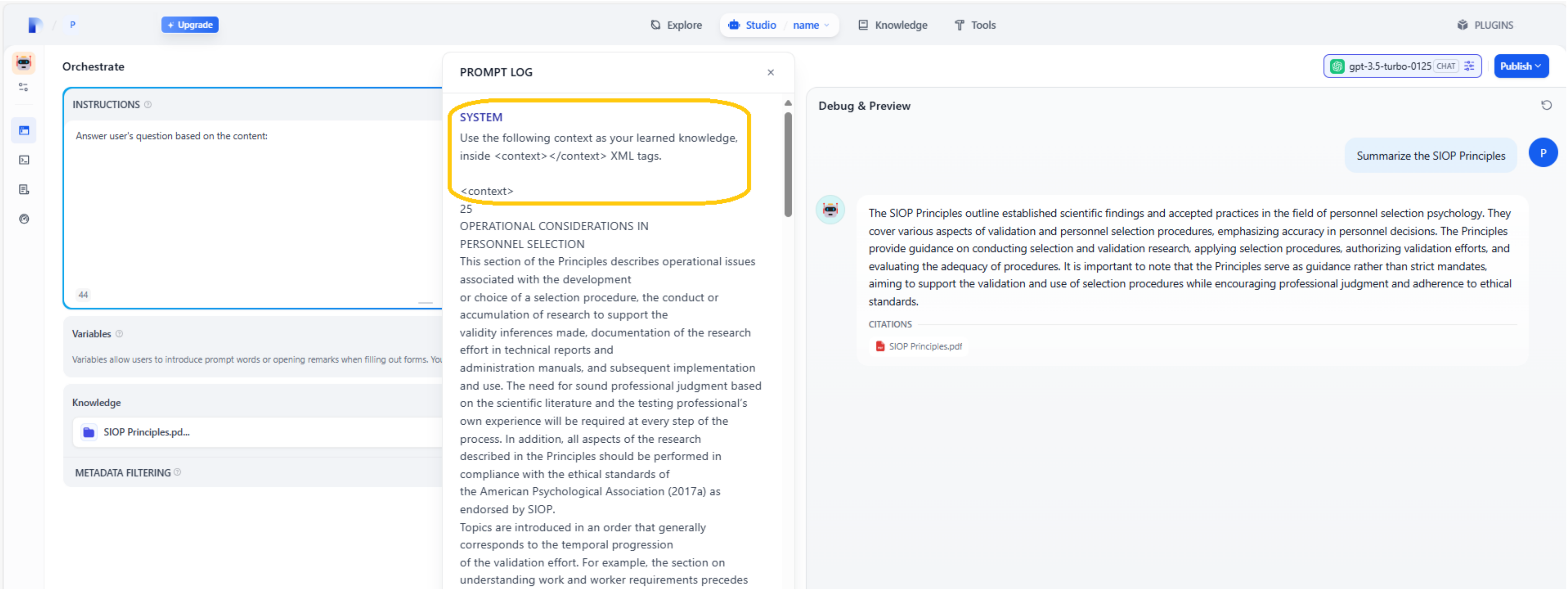
Summarize the SIOP Principles

The SIOP Principles provide guidance on the validation and use of personnel selection procedures in the field of psychology. They outline established scientific findings and accepted practices related to measuring constructs related to work behavior. The principles emphasize the accuracy of inferences in making personnel decisions and cover aspects such as conducting selection and validation research, applying selection procedures, and evaluating their adequacy and appropriateness. The principles are aspirational and meant to assist professionals in the field, requiring professional judgment to determine their relevance in specific situations.

CITATIONS

SIOP Principles.pdf

DIFY: Debugging & Trialing the Chatbot



Answer user's question based on the content:

Variables allow users to introduce prompt words or opening remarks when filling out forms. You

SIOP Principles.pd...

SYSTEM

Use the following context as your learned knowledge, inside <context></context> XML tags.

<context>

25

OPERATIONAL CONSIDERATIONS IN
PERSONNEL SELECTION

This section of the Principles describes operational issues associated with the development
or choice of a selection procedure, the conduct or accumulation of research to support the
validity inferences made, documentation of the research effort in technical reports and
administration manuals, and subsequent implementation and use. The need for sound professional judgment based on the scientific literature and the testing professional's own experience will be required at every step of the process. In addition, all aspects of the research described in the Principles should be performed in compliance with the ethical standards of
the American Psychological Association (2017a) as endorsed by SIOP.

Topics are introduced in an order that generally corresponds to the temporal progression
of the validation effort. For example, the section on understanding work and worker requirements precedes

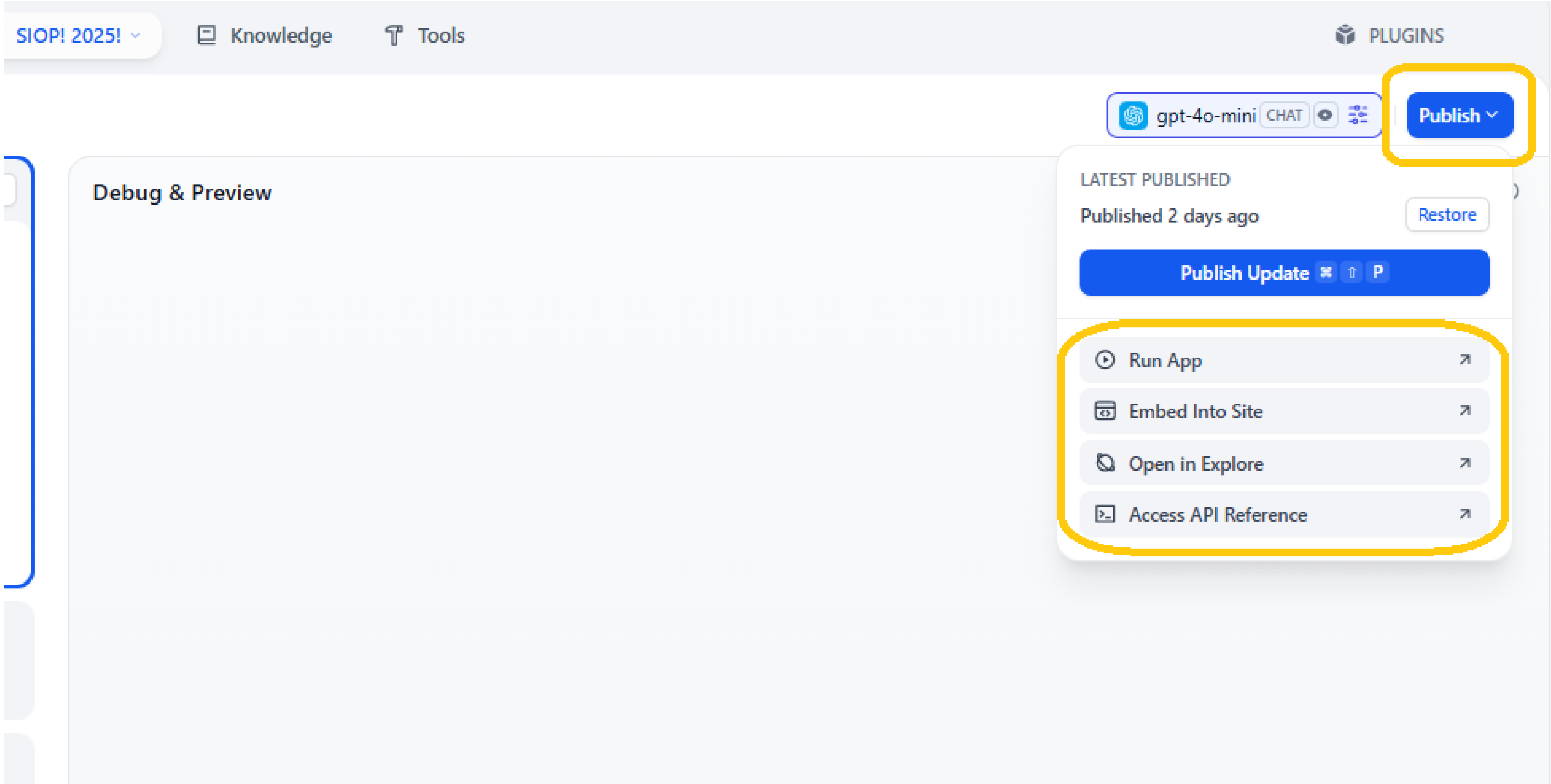
Summarize the SIOP Principles

The SIOP Principles outline established scientific findings and accepted practices in the field of personnel selection psychology. They cover various aspects of validation and personnel selection procedures, emphasizing accuracy in personnel decisions. The Principles provide guidance on conducting selection and validation research, applying selection procedures, authorizing validation efforts, and evaluating the adequacy of procedures. It is important to note that the Principles serve as guidance rather than strict mandates, aiming to support the validation and use of selection procedures while encouraging professional judgment and adherence to ethical standards.

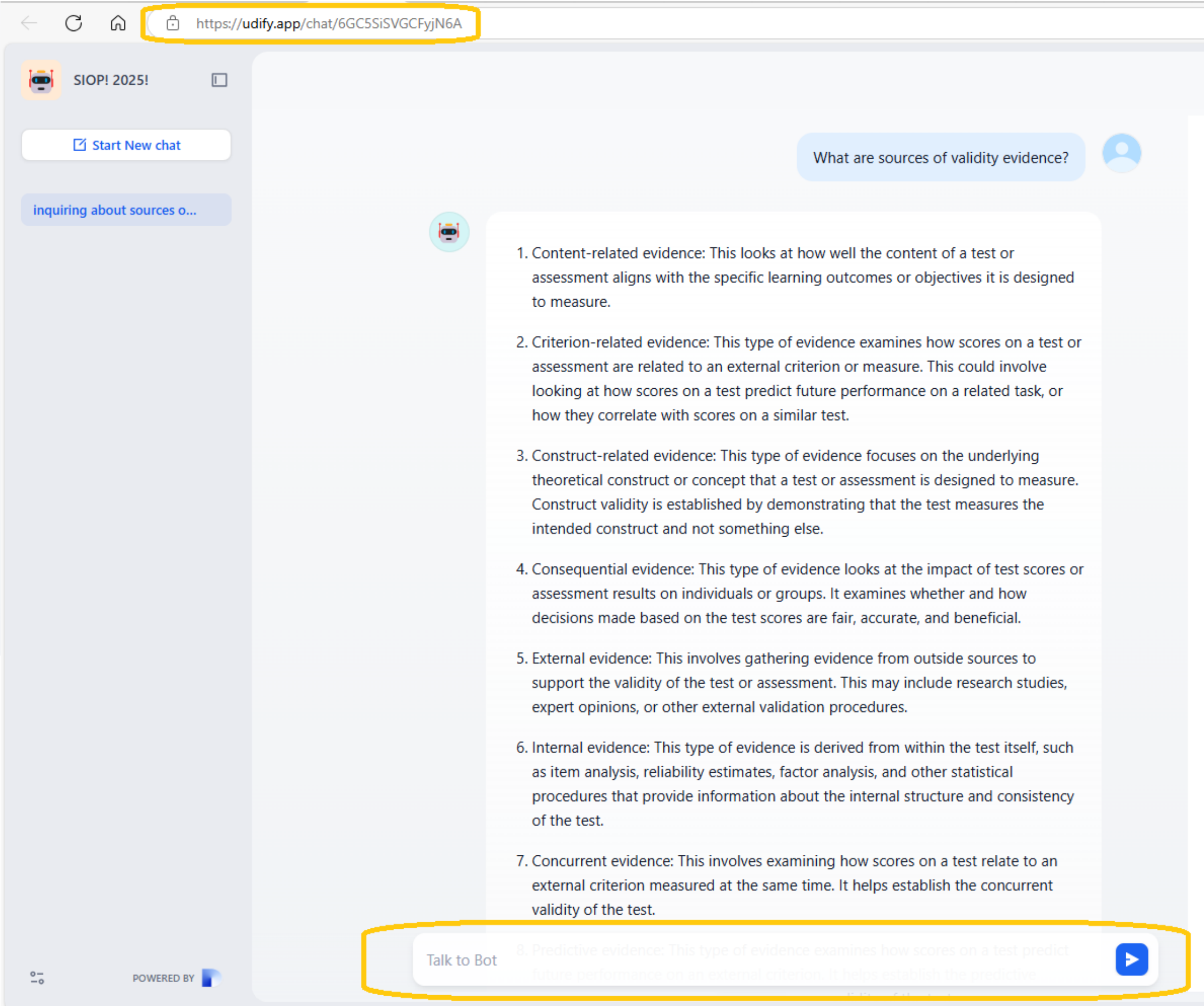
CITATIONS

SIOP Principles.pdf

DIFY: Using the Chatbot



DIFY: Using the Chatbot



https://udify.app/chat/6GC5SiSVGCFyjN6A
SIOP! 2025!
Start New chat
inquiring about sources o...
POWERED BY
Talk to Bot

What are sources of validity evidence?

1. Content-related evidence: This looks at how well the content of a test or assessment aligns with the specific learning outcomes or objectives it is designed to measure.
2. Criterion-related evidence: This type of evidence examines how scores on a test or assessment are related to an external criterion or measure. This could involve looking at how scores on a test predict future performance on a related task, or how they correlate with scores on a similar test.
3. Construct-related evidence: This type of evidence focuses on the underlying theoretical construct or concept that a test or assessment is designed to measure. Construct validity is established by demonstrating that the test measures the intended construct and not something else.
4. Consequential evidence: This type of evidence looks at the impact of test scores or assessment results on individuals or groups. It examines whether and how decisions made based on the test scores are fair, accurate, and beneficial.
5. External evidence: This involves gathering evidence from outside sources to support the validity of the test or assessment. This may include research studies, expert opinions, or other external validation procedures.
6. Internal evidence: This type of evidence is derived from within the test itself, such as item analysis, reliability estimates, factor analysis, and other statistical procedures that provide information about the internal structure and consistency of the test.
7. Concurrent evidence: This involves examining how scores on a test relate to an external criterion measured at the same time. It helps establish the concurrent validity of the test.
8. Predictive evidence: This type of evidence examines how scores on a test predict future performance on an external criterion. It helps establish the predictive

Q&A Resources

- Q&A
- Slides available on **Whova**
- All materials (slides, code, and supplementary resources) found on **GitHub repo** here: https://github.com/karimhbadr1/SIOP_2025_ChatBot_Master_Tutorial