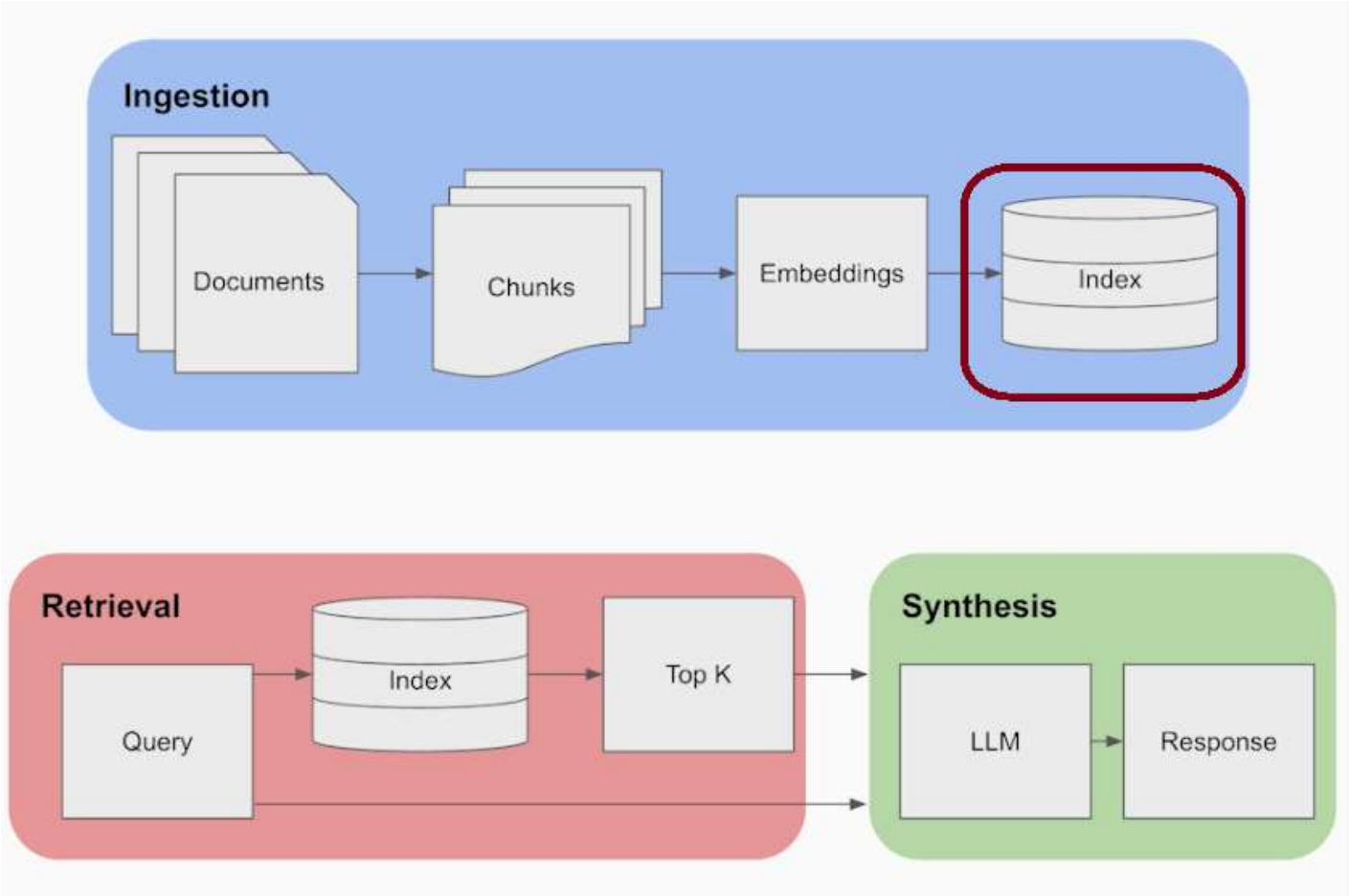
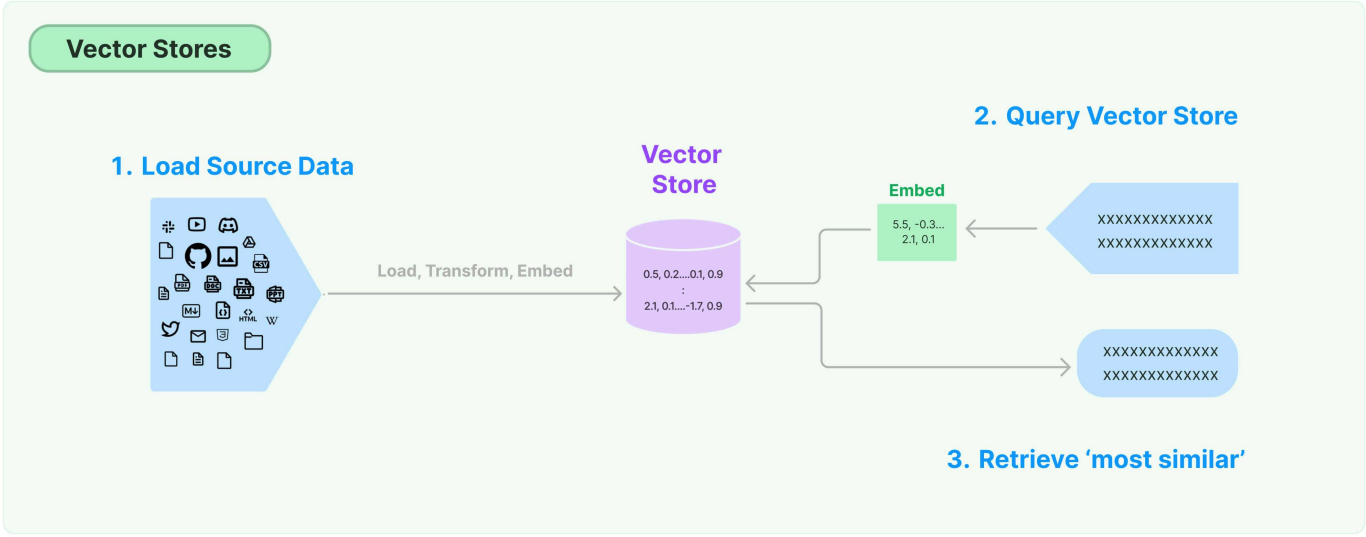
Vector Store

One of the most common ways to store and search over unstructured data is to embed it and store the resulting embedding vectors, and then at query time to embed the unstructured query and retrieve the embedding vectors that are 'most similar' to the embedded query. A vector store takes care of storing embedded data and performing vector search for you.



How Data is Stored in VectorDB




```
1 !pip install qdrant_client langchain_huggingface langchain-community langchain-qdrant py
```

```
Collecting qdrant_client
  Downloading qdrant_client-1.11.1-py3-none-any.whl.metadata (10 kB)
Collecting langchain_huggingface
  Downloading langchain_huggingface-0.0.3-py3-none-any.whl.metadata (1.2 kB)
Collecting langchain-community
  Downloading langchain_community-0.2.16-py3-none-any.whl.metadata (2.7 kB)
Collecting langchain-qdrant
  Downloading langchain_qdrant-0.1.3-py3-none-any.whl.metadata (1.7 kB)
Collecting pypdf
  Downloading pypdf-4.3.1-py3-none-any.whl.metadata (7.4 kB)
Collecting openai
  Downloading openai-1.44.1-py3-none-any.whl.metadata (22 kB)
Collecting langchain
  Downloading langchain-0.2.16-py3-none-any.whl.metadata (7.1 kB)
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: grpcio>=1.41.0 in /usr/local/lib/python3.10/dist-packag
Collecting grpcio-tools>=1.41.0 (from qdrant_client)
  Downloading grpcio_tools-1.66.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_
Collecting httpx>=0.20.0 (from httpx[http2]>=0.20.0->qdrant_client)
  Downloading httpx-0.27.2-py3-none-any.whl.metadata (7.1 kB)
Requirement already satisfied: numpy>=1.21 in /usr/local/lib/python3.10/dist-packages
Collecting portalocker<3.0.0,>=2.7.0 (from qdrant_client)
  Downloading portalocker-2.10.1-py3-none-any.whl.metadata (8.5 kB)
Requirement already satisfied: pydantic>=1.10.8 in /usr/local/lib/python3.10/dist-pacl
Requirement already satisfied: urllib3<3,>=1.26.14 in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: huggingface-hub>=0.23.0 in /usr/local/lib/python3.10/di
Collecting langchain-core<0.3,>=0.1.52 (from langchain_huggingface)
  Downloading langchain_core-0.2.38-py3-none-any.whl.metadata (6.2 kB)
Collecting sentence-transformers>=2.6.0 (from langchain_huggingface)
  Downloading sentence_transformers-3.0.1-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: tokenizers>=0.19.1 in /usr/local/lib/python3.10/dist-pa
Requirement already satisfied: PyYAML>=5.3 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: SQLAlchemy<3,>=1.4 in /usr/local/lib/python3.10/dist-pa
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in /usr/local/lib/python3.10/dist
```

```
Collecting dataclasses-json<0.7,>=0.5.7 (from langchain-community)
  Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Collecting langsmith<0.2.0,>=0.1.0 (from langchain-community)
  Downloading langsmith-0.1.117-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: requests<3,>=2 in /usr/local/lib/python3.10/dist-packag
Collecting tenacity!=8.4.0,<9.0.0,>=8.1.0 (from langchain-community)
  Downloading tenacity-8.5.0-py3-none-any.whl.metadata (1.2 kB)
Requirement already satisfied: typing_extensions>=4.0 in /usr/local/lib/python3.10/dis
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: distro<2,>=1.7.0 in /usr/lib/python3/dist-packages (fro
Collecting jiter<1,>=0.4.0 (from openai)
  Downloading jiter-0.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.m
Requirement already satisfied: sniffio in /usr/local/lib/python3.10/dist-packages (fro
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: async-timeout<5.0.0,>=4.0.0 in /usr/local/lib/python3.1
Collecting langchain-text-splitters<0.3.0,>=0.2.0 (from langchain)
  Downloading langchain_text_splitters-0.2.4-py3-none-any.whl.metadata (2.3 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (fr
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.10/dist-pa
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.10/di
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-pack
```

```python
from qdrant_client import QdrantClient
from langchain_core.documents import Document
from langchain.document_loaders import PyPDFLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain_qdrant import QdrantVectorStore
import openai
import os
```

Embedding Model

Dimensions 384

Max Input Token 512

```python
# Initialize embedding model with BAAI/bge-small-en-v1.5
embed_model = HuggingFaceEmbeddings(model_name='BAAI/bge-small-en-v1.5')

```

```
<ipython-input-3-3d4ca37c1a92>:2: LangChainDeprecationWarning: The class `HuggingFaceEmb
  embed_model = HuggingFaceEmbeddings(model_name='BAAI/bge-small-en-v1.5')
/usr/local/lib/python3.10/dist-packages/sentence_transformers/cross_encoder/CrossEncoder
  from tqdm.autonotebook import tqdm, trange
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public model
  warnings.warn(
```

modules.json: 100% 349/349 [00:00<00:00, 10.5kB/s]

config_sentence_transformers.json: 100% 124/124 [00:00<00:00, 5.18kB/s]

README.md: 100% 94.8k/94.8k [00:00<00:00, 2.54MB/s]

sentence_bert_config.json: 100% 52.0/52.0 [00:00<00:00, 1.62kB/s]

config.json: 100% 743/743 [00:00<00:00, 31.4kB/s]

model.safetensors: 100% 133M/133M [00:01<00:00, 121MB/s]

tokenizer_config.json: 100% 366/366 [00:00<00:00, 22.1kB/s]

vocab.txt: 100% 232k/232k [00:00<00:00, 9.24MB/s]

tokenizer.json: 100% 711k/711k [00:00<00:00, 11.5MB/s]

special_tokens_map.json: 100% 125/125 [00:00<00:00, 5.54kB/s]

1_Pooling/config.json: 100% 190/190 [00:00<00:00, 5.19kB/s]

Loading the Data

```
# Load the PDF document using PyPDFLoader
loaders = PyPDFLoader("/content/National_AI_Policy_Consultation_Draft_1722220582.pdf")

# Extract pages from the loaded PDF
pages = loaders.load()

```

```
pages[15]
```

```
Document(metadata={'source':
'/content/National_AI_Policy_Consultation_Draft_1722220582.pdf', 'page': 15},
page_content=" \n \n1 \n 4 Policy Directives  \nThe policy directives are minimalistic,
focusing  on resolving issues and achieving targets set for stimulating \ngrowth in AI
across the board. Empathizing with  the common  person's journey  for different aspects
\nassociated with their socio -economic development and well -being in the current
technological disruption \nis driven through the following developmental pillars.
\n4.1 1st Pillar: AI Market Enablement  \n4.1.1  National Artificial Intelligence Fund
```

```
(NA IF)  \nGiven the evidence regarding the state of AI in Pakistan, the projected global outlook of AI in terms of its \nuse and market size, the impact of AI on the local ecosystem , and claiming its demographic share through \nresponsible use of data, the Ministry  of IT & Telecom through its underutilized resources and funds aims \nto establish a National AI Fund with following objectives . \nI. In accordance with the stipulations of clauses 33D (II) & (III) of the Telecommunication Re -\norganization Act 1996 (amended 2006),  the Ministry of IT & Telecom, while exercising its right to \nissue policy directives, shall direct the Research & Development Fund (Ignite – Technology Fund) \nto allocate a part (not less than 30%) of its funds to NAIF on a perpetual basis for the research and \ndevelopment of AI and allied technologies. \nII. The Ministry of IT & Telecom shall notify the establish ment  of an autonomous high -tech National \nAI Fund (NAIF) organization within six (6) months from the promulgation of this policy.  \nIII. The NAIF shall undertake  all the responsibilities and implement guidelines as stipulated in this \npolicy or as directed by the Federal Government of Pakistan via the Ministry of IT & Telecom from \ntime to time.  \nIV. The Ministry of IT & Telecom shall allocate a budget through PSDP funds  as Initial Working Capital \nto support the initiative expeditiously in the first two (2) years.  \nV. Once the organization is formed and Funds are allocated and transferred into NAIF from the \nNational ICT R&D Fund , all the ongoing and subsequent programs shall be organized through the \nperpetual Fund.  \nVI. The fund shall be administered through a n independent  Board of Directors (not more than 11 \nmembers) to ensure seamless operations and transparency.  \nVII. The BoD shall comprise members from industry and academia with rele vant techno -commercial \nbackgrounds in high -tech (especially AI & allied technologies development), representatives of \nIgnite R&D Fund BoD , and the government (ex -officio). It shall be chaired by Secretary/Member \nIT. \nVIII. NAIF shall be able to raise funds throug h international grants/aids from bilateral and multilateral \nplatforms, co -invest with local/international hi -tech organizations, provide a bridge between \nglobal VCs/CVCs , and incubate R&D initiatives and startups for early commercialization and \nsustainabil ity. \nIX. The funds allocated and disbursed to NAIF shall not be lapsable upon completion of a financial \nyear . Part of NAIF's fund shall only be reimbursable by Ignite – Technology Fund on a year -on-year \nbasis.  \nX. NAIF administration shall ensure that the funds ar e utilized per the stipulations of this policy and \nwithin the defined mandate of the Research & Development Fund (Ignite – Technology Fund).  ")
```

```
1 len(pages)
```

```
41
```

Splitting the Document into Chuncks

```
1 text_splitter = RecursiveCharacterTextSplitter(
2     chunk_size = 1500,
3     chunk_overlap = 150
4 )
```

Double-click (or enter) to edit

Meta Data preprocessing

```
from langchain.docstore.document import Document

```

```
# Create an empty list to store processed document chunks
doc_list = []

# Iterate over each page in the extracted pages
for page in pages:
    # Split the page content into smaller chunks
    pg_split = text_splitter.split_text(page.page_content)

    # Iterate over each chunk and create Document objects
    for pg_sub_split in pg_split:
        # Metadata for each chunk, including source and page number
        metadata = {"source": "AI policy", "page_no": page.metadata["page"] + 1}

        # Create a Document object with content and metadata
        doc_string = Document(page_content=pg_sub_split, metadata=metadata)

        # Append the Document object to the list
        doc_list.append(doc_string)
```

```
doc_list[10]
```

```
Document(metadata={'source': 'AI policy', 'page_no': 6}, page_content='6 \n 1 Executive Summary  \nPakistan has a unique opportunity to harness digital disruption by educating an eager young population \nthat can potentially propel the n ation onto a growth trajectory to sustain our future national \ncompetitiveness and improve the lives of citizens. Artificial Intelligence (AI) represents the next frontier of \ntechnological opportunities, and it has been widely proven and understood that the  collection, processing, \nuse, and exchange of data through automated/intelligent means would drive the entire society into the \nnext stage of its evolution which is unprecedented and requires a progressive , yet careful approach. So, \nafter a thorough analysi s of the global perspective and based on the evidence collected through more \nextensive consultations with the stakeholders, the Ministry of IT & Telecom has come to a much -desired \nconclusion that it needs to chalk out a developmental roadmap for better, faster and responsible adoption \nof AI in the country . For that, the policy document is put in place to reap long -term and sustainable benefits \nfor its people.  \nThe policy document offers a wide range of developmental initiatives necessary for awareness and \nadoption, reimagining the transparent and fair use of personal data using AI and stimulating innovation \nthrough industry -academia collaborations and investments in AI-led initiatives. The National AI Policy is')
```

```
len(doc_list)
```

```
103
```

Qdrant Vectore Store

Qdrant Credentials

```
qdrant_url = "-"
qdrant_key = "-"
collection_name = "AI_policy_new"
```

```
# Initialize QdrantVectorStore with documents and embedding model
qdrant = QdrantVectorStore.from_documents(
    doc_list,                # List of Document objects to be stored in the vector store
    embed_model,             # Embedding model used to convert documents into vectors
    url=qdrant_url,          # URL for the Qdrant service
    api_key=qdrant_key,      # API key for accessing the Qdrant service
    collection_name=collection_name  # Name of the collection to store the vectors in
)
```

Query Vector Store

Once your vector store has been created and the relevant documents have been added you will most likely wish to query it during the running of your chain or agent.

Query directly

The simplest scenario for using Qdrant vector store is to perform a similarity search. Under the hood, our query will be encoded into vector embeddings and used to find similar documents in Qdrant collection.

```
query = "what is Ai policy for students?"

# Retrieve relevant documents
results = qdrant.similarity_search(query, k=5)
```

```
results[3]
```

```
Document(metadata={'source': 'AI policy', 'page_no': 6, '_id': 'd022222e-eae7-4d5c-87a3-459669e64dcd', '_collection_name': 'AI_policy_new'}, page_content='through industry -academia collaborations and investments in AI-led initiatives. The National AI Policy is \ncrafted to focus on the equitable distribution o f opportunity and its responsible use , having the following \ndefining attributes.  \n• Evidence -Based and Target Oriented  \n• User -Centric and Forward -Looking  \n• Objective and Overarching \nThe AI policy further aims to augment AI and allied technologies through balanced demand and supply -\nside interventions , as briefly described below.   \n• Market
```

```
Enablement - Establishment  of research & innovation cent ers in AI for developing,
test -\nbedding, deploying, and scaling AI solutions.  This includes learning how to
improve govern ance  \nand manage the impact of AI.  \n• Progressive  and Trusted
Environment – Responsible  use of AI to generate economic gains and \nimprove lives. In
addition,  AI will raise the Government’s capability to deliver anticipatory and
\npersonali zed services.  \n• Enabling AI through Awareness and Readiness  - Pakistan
shall increase awareness and \nunderstand ing of AI technologies and the ir benefits ;
our workforce will  be equ ipped with the \nnecessary competencies to participate in
the AI economy.  \n• Transformation & Evolution - Transformation of sectors and
industries towards effective use of \nAI, facilitated by national IT boards through
creating awareness and offering training pr ograms \nthrough sectoral cooperation.')
```

```
results[0].page_content
```

```
'4.2.3  Algorithms, Data Science  & AI in Basic Education  \nI. Where the policy docume
nt emphasizes the funda mental  understanding and awareness of \npersonal data protecti
on and AI, it also aims to stimulate  an incremental impact of AI on society \nright fr
om the grassroots . Therefore, the policy has given equal importance to teach ing algor
ithms, \ndata science , and AI in basic STEM education . In this regard, it stipulates
that  CoE-AI shall  hire a \nlocal/international consultant with expertise in high -tec
```

Pinecone Vector Store

```
%pip install -qU langchain-pinecone pinecone-notebooks
```

```
                                        244.8/244.8 kB 6.1 MB/s eta 0:00:00
                                        117.6/117.6 kB 6.1 MB/s eta 0:00:00
```

Pinecone credentials

```
PINECONE_API_KEY="2d256283-3bbc-4669-93b3-b824eacebfde"
index_name="demo-vectorstore"
```

Data Upsertion in Pinecone

```
from langchain_pinecone import PineconeVectorStore as lang_pinecone
import os
os.environ["PINECONE_API_KEY"] = PINECONE_API_KEY
```

```
# Convert documents into vectors using LangPinecone
vector = lang_pinecone.from_documents(
    doc_list,               # List of Document objects to be converted into vectors
    embed_model,            # Embedding model used for generating vector representatior
```

```
    index_name=index_name    # Name of the Pinecone index where vectors will be stored
)
```

Query vectorstore

```
# Define a query to search for relevant information
query = "What is AI policy for students?"

# Perform similarity search to find the top 5 most relevant results
pinecone_results = vector.similarity_search(query, k=5)

```

```
pinecone_results
```

[Document(metadata={'page_no': 25.0, 'source': 'AI policy'}, page_content='4.2.3 Algorithms, Data Science & AI in Basic Education \nI. Where the policy document emphasizes the funda mental understanding and awareness of \npersonal data protection and AI, it also aims to stimulate an incremental impact of AI on society \nright from the grassroots . Therefore, the policy has given equal importance to teach ing algorithms, \ndata science , and AI in basic STEM education . In this regard, it stipulates that. CoE-AI shall hire a \nlocal/international consultant with expertise in high -tech curriculum development to develop a \nNational Curriculum in Algorithms, Data Sciences, AI , and Allied Technologies from the sixth to the \ntwelfth standard.'),
 Document(metadata={'page_no': 25.0, 'source': 'AI policy'}, page_content='4.2.3 Algorithms, Data Science & AI in Basic Education \nI. Where the policy document emphasizes the funda mental understanding and awareness of \npersonal data protection and AI, it also aims to stimulate an incremental impact of AI on society \nright from the grassroots . Therefore, the policy has given equal importance to teach ing algorithms, \ndata science , and AI in basic STEM education . In this regard, it stipulates that. CoE-AI shall hire a \nlocal/international consultant with expertise in high -tech curriculum development to develop a \nNational Curriculum in Algorithms, Data Sciences, AI , and Allied Technologies from the sixth to the \ntwelfth standard.'),
 Document(metadata={'page_no': 27.0, 'source': 'AI policy'}, page_content='policies. \nIX. Develop regulation policies and standards for data -sharing among countries and lead multilateral \ndiplomatic efforts to arra nge such agreements. \nX. Encourage local businesses to embrace new AI solutions and provide them with a platform for \ntechnical support and some incentives and regulations. Moreover, it should catalyze the creation \nof new businesses based on AI technology throu gh start -up funds and incubation centers . \nXI. Formulate policies to develop and maintain highly resilient cutting -edge computing, storage, and \nconnectivity infrastructure. \nXII. Participate in international efforts to bring standardization in all aspects of AI, e.g. , data formats, \nnetwork and systems architecture, data , application integration protocols, requirements on test \ncases, and services. \nXIII. Develop a data -sharing framework and use AI algorithms consistent with social, cultural, and \nreligious norms and internatio nal guidelines. \nXIV. A governance mechanism that will facilitate fairness, data privacy, ethical values control, and \nalgorithmic accountability will be implemented to support reliability in AI studies.'),
 Document(metadata={'page_no': 27.0, 'source': 'AI policy'}, page_content='policies. \nIX. Develop regulation policies and standards for data -sharing among countries and lead multilateral \ndiplomatic efforts to arra nge such agreements. \nX. Encourage

```
local businesses to embrace new AI solutions and provide them with a platform for
\ntechnical support and some incentives and regulations. Moreover, it should catalyze
the creation \nof new businesses based on AI technology throu gh start -up funds and
incubation centers . \nXI. Formulate policies to develop and maintain highly resilient
cutting -edge computing, storage, and \nconnectivity infrastructure.  \nXII.
Participate in international efforts to bring standardization in all aspects of AI,
e.g. , data formats, \nnetwork and systems architecture, data , application integration
protocols, requirements on test \ncases, and services.  \nXIII. Develop a data -sharing
framework  and use AI algorithms consistent with social, cultural, and \nreligious
norms and internatio nal guidelines.  \nXIV. A governance mechanism that will
facilitate fairness, data privacy, ethical values control, and \nalgorithmic
accountability will be implemented to support reliability in AI studies.'),
 Document(metadata={'page_no': 1.0, 'source': 'AI policy'}, page_content='i\n \n \n \n
\n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n
Draft\nNational\n \nArtificial Intelligence Policy\n \n \n \n \n \n \n \n \n \n \n \n
\n \n \n \n \n \n \n \n \n \n \n \n \nGovernment of Pakistan\nMinistry of Information
Technology & Telecommunication\nhttps://moitt.gov.pk')]
```

```
pinecone_results[3].page_content
```

```
'policies.  \nIX. Develop regulation policies and standards for data -sharing among cou
ntries and lead multilateral \ndiplomatic efforts to arra nge such agreements.  \nX. En
courage local businesses to embrace new AI solutions and provide them with a platform f
or \ntechnical support and some incentives and regulations. Moreover, it should catalyz
e the creation \nof new businesses based on AI technology throu gh start -up funds and
```