GenAI: Small Language Models

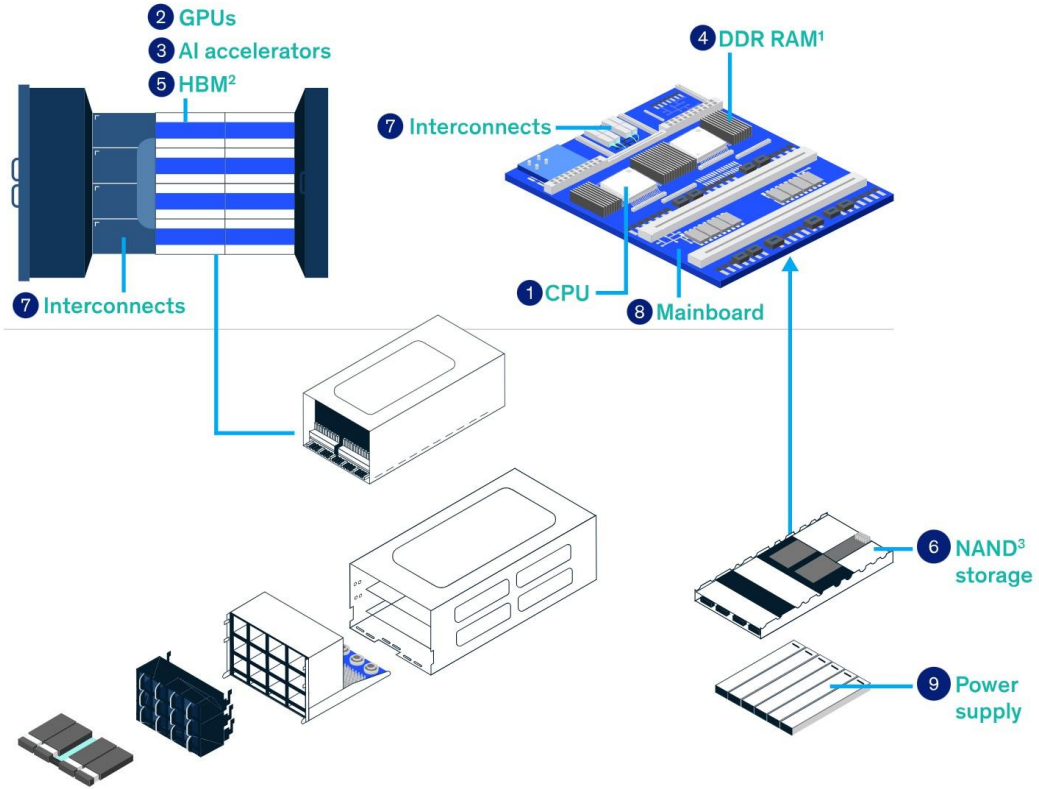
Improved Privacy, Domain Specific and Lower Computational Requirements


Finder File Edit View Go Window Help
Mon 9 Sep 9:41 AM
Type to Siri
Here's how to merge PDFs together into one PDF by dragging pages in Preview:
1. Open the PDFs you want to combine in Preview.
2. In each open PDF, click View in the menu bar, then click Thumbnails to display the page thumbnails in the sidebar.
3. Drag the thumbnails you want to add from the sidebar of one PDF to the sidebar of the other PDF.
Preview User Guide
Mon 9 Tiburon
9:41
Directions to Home
Play Road Trip Classics
Share my location
Ask Siri...
Call
Play
Set
Mon 9 Tiburon
9:41

Components of AI Server

"AI is a complex pipeline of interlocked technologies, including networking, data storage, memory, accelerators, models, tools, and algorithms to name a few. "

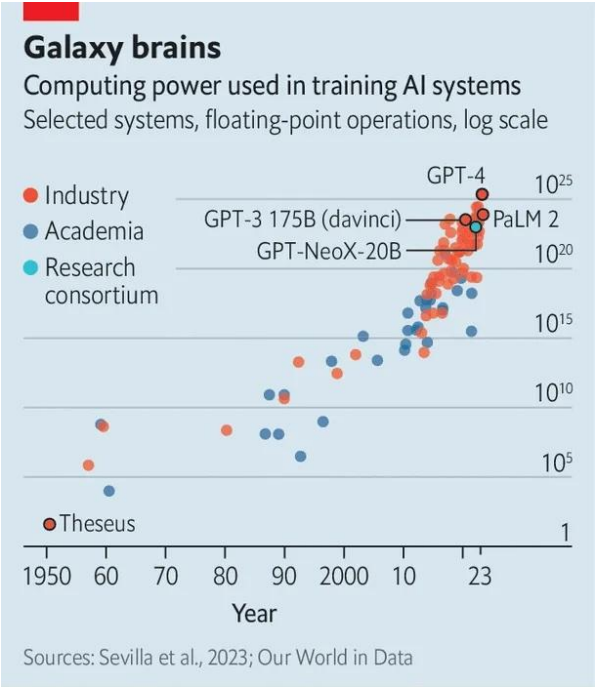
Source: The Flash Storage-AI connection, explained. Pure Storage Blog.



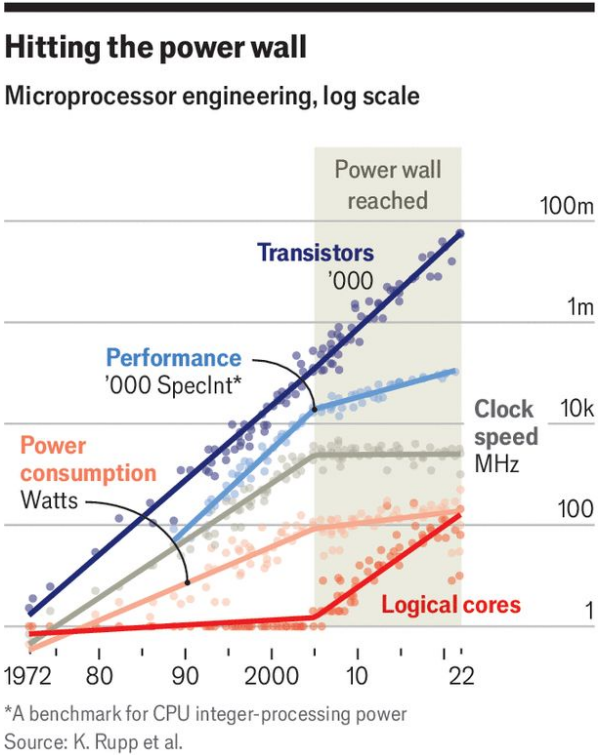
McKinsey & Company.

Compute Demand

The **“Power Wall”** refers to the difficulty of scaling the performance of computing chips and systems at historical levels.

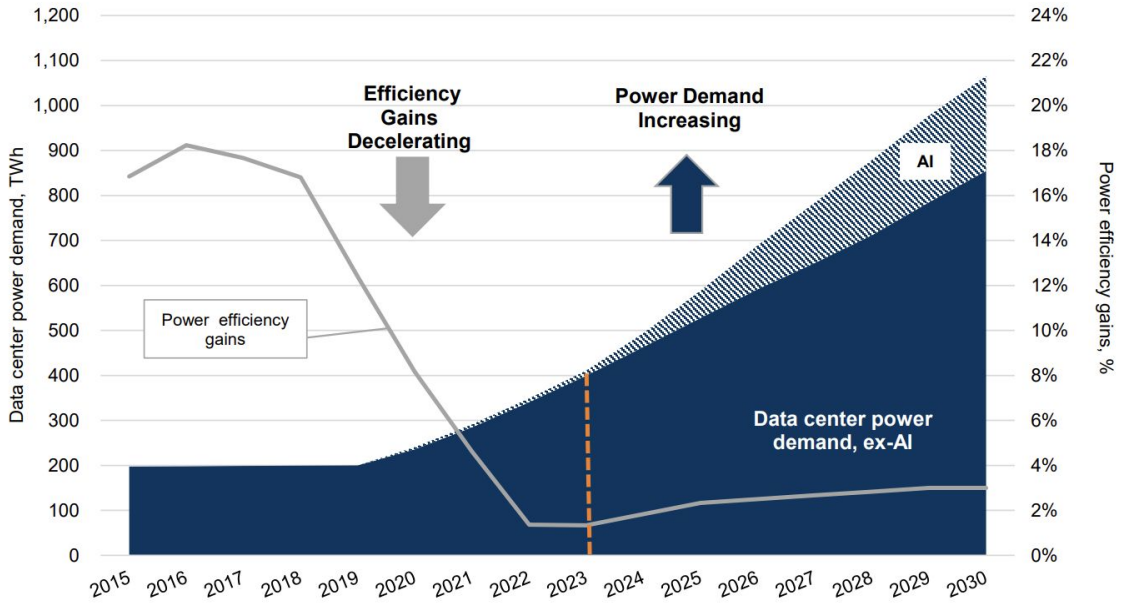


The Economist



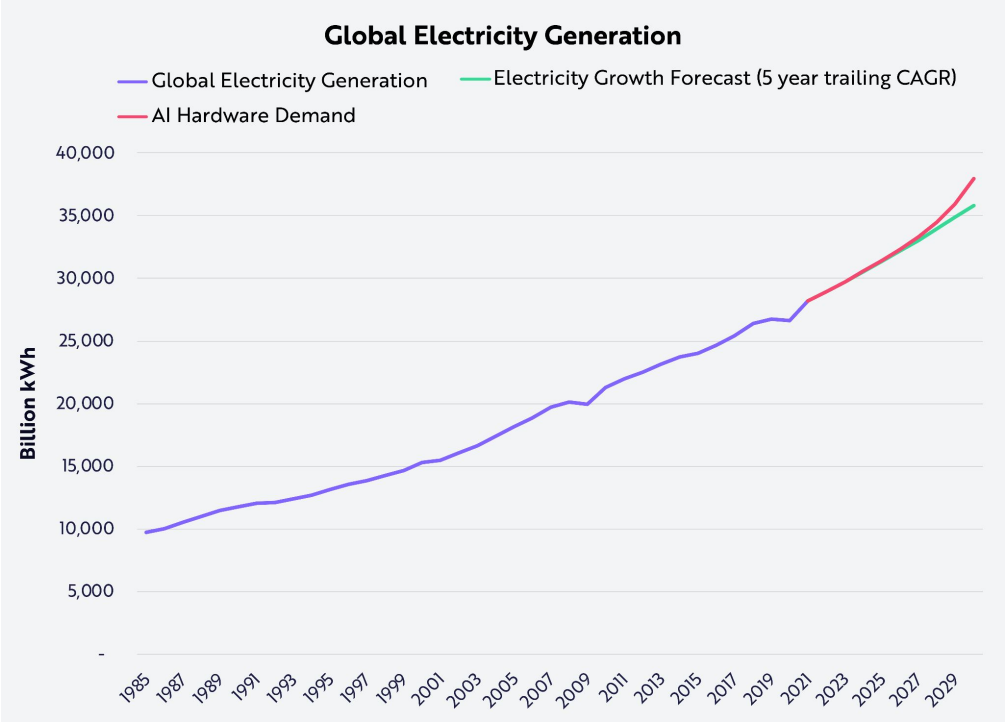
Decreasing Data Centers Efficiency

Exhibit 1: After being flattish for 2015-19, we see power demand from data centers more than tripling in 2030 vs. 2020, with an upside case more than double the base case depending in part on product efficiencies and AI demand
Data center electricity consumption, TWh (LHS) and 3-year rolling average power efficiency gains yoy, % (RHS)



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Global Investment Research

Power Demand For AI Hardware



Ark Invest

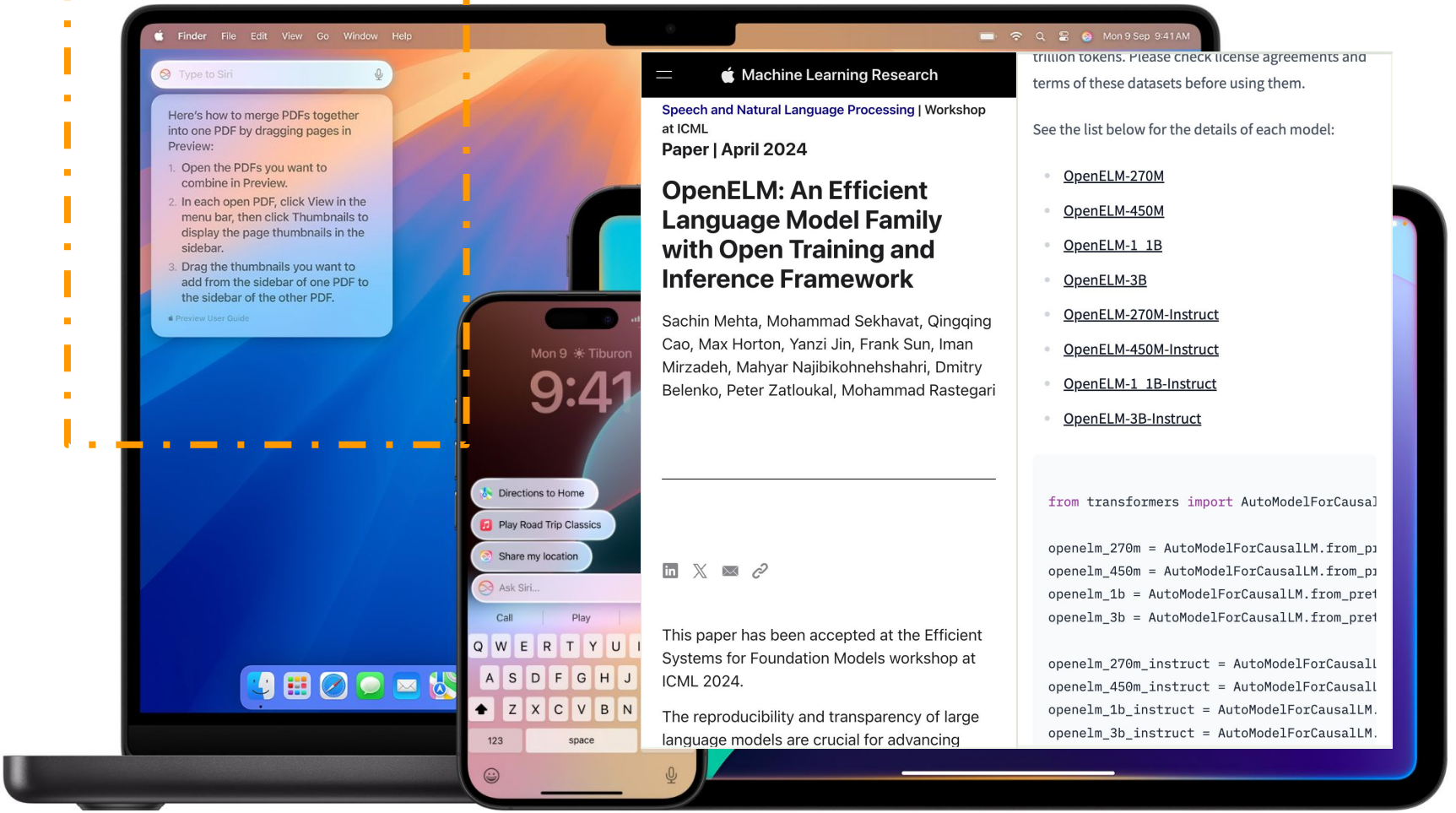
What is SLM?

Small language models (SLMs) are artificial intelligence (AI) models capable of processing, understanding and generating natural language content. As their name implies, SLMs are smaller in scale and scope than large language models (LLMs).

LLMs serve as the base for SLMs. Model compression techniques (Pruning,Quantization,Low-rank factorization,Knowledge distillation) are applied to build a leaner model from a larger one. Compressing a model entails reducing its size while still retaining as much of its accuracy as possible.

Examples of SMLs: DistilBERT,Gemma,GPT-4o mini, Granite,Llama, Mistral,Phi

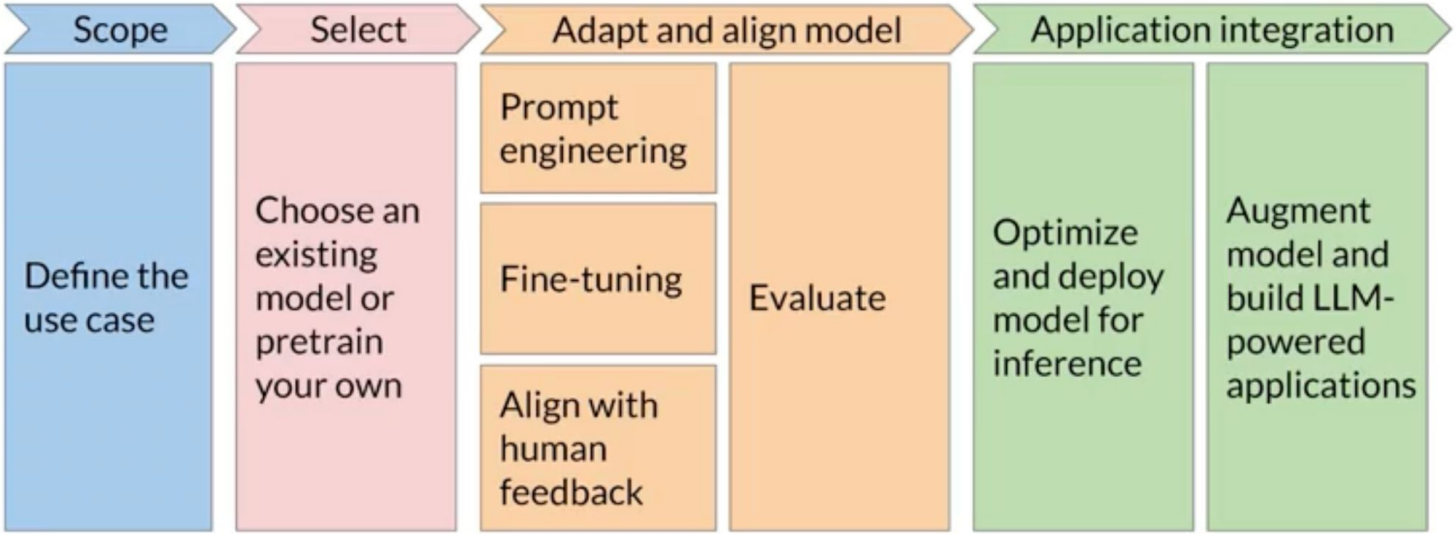
Source: What are small language models?- IBM Blog


Finder File Edit View Go Window Help
Mon 9 Sep 9:41 AM
Type to Siri
Here's how to merge PDFs together into one PDF by dragging pages in Preview:
1. Open the PDFs you want to combine in Preview.
2. In each open PDF, click View in the menu bar, then click Thumbnails to display the page thumbnails in the sidebar.
3. Drag the thumbnails you want to add from the sidebar of one PDF to the sidebar of the other PDF.
Preview User Guide
Mon 9 Tiburon
9:41
Directions to Home
Play Road Trip Classics
Share my location
Ask Siri...
Call
Play
Machine Learning Research
Speech and Natural Language Processing | Workshop at ICML
Paper | April 2024
OpenELM: An Efficient Language Model Family with Open Training and Inference Framework
Sachin Mehta, Mohammad Sekhavat, Qingqing Cao, Max Horton, Yanzi Jin, Frank Sun, Iman Mirzadeh, Mahyar Najibikohnehshahri, Dmitry Belenko, Peter Zatloukal, Mohammad Rastegari
This paper has been accepted at the Efficient Systems for Foundation Models workshop at ICML 2024.
The reproducibility and transparency of large language models are crucial for advancing
trillion tokens. Please check license agreements and terms of these datasets before using them.
See the list below for the details of each model:
OpenELM-270M
OpenELM-450M
OpenELM-1_1B
OpenELM-3B
OpenELM-270M-Instruct
OpenELM-450M-Instruct
OpenELM-1_1B-Instruct
OpenELM-3B-Instruct
from transformers import AutoModelForCausal
openelm_270m = AutoModelForCausalLM.from_p
openelm_450m = AutoModelForCausalLM.from_p
openelm_1b = AutoModelForCausalLM.from_pret
openelm_3b = AutoModelForCausalLM.from_pret
openelm_270m_instruct = AutoModelForCausall
openelm_450m_instruct = AutoModelForCausall
openelm_1b_instruct = AutoModelForCausalLM.
openelm_3b_instruct = AutoModelForCausalLM.

Will knowledge distillation, combined with quantization, become the most prevalent and effective method for creating high-performing SLMs, thereby accounting for most of commercial applications within the next three years?

- ➔ SLM Saves Cost
- ➔ SLM is Private
- ➔ Yes some SLMs can run on CPUs
- ➔ Can be made more effective with Tools
- ➔ SLM != LLM (Tool Use, but not Agents*)
- ➔ Protype your ideas on SLM, deploy on LLM

Get started: Lifecycle



Get started: Use Cases

There are many applications of generative AI across modalities.

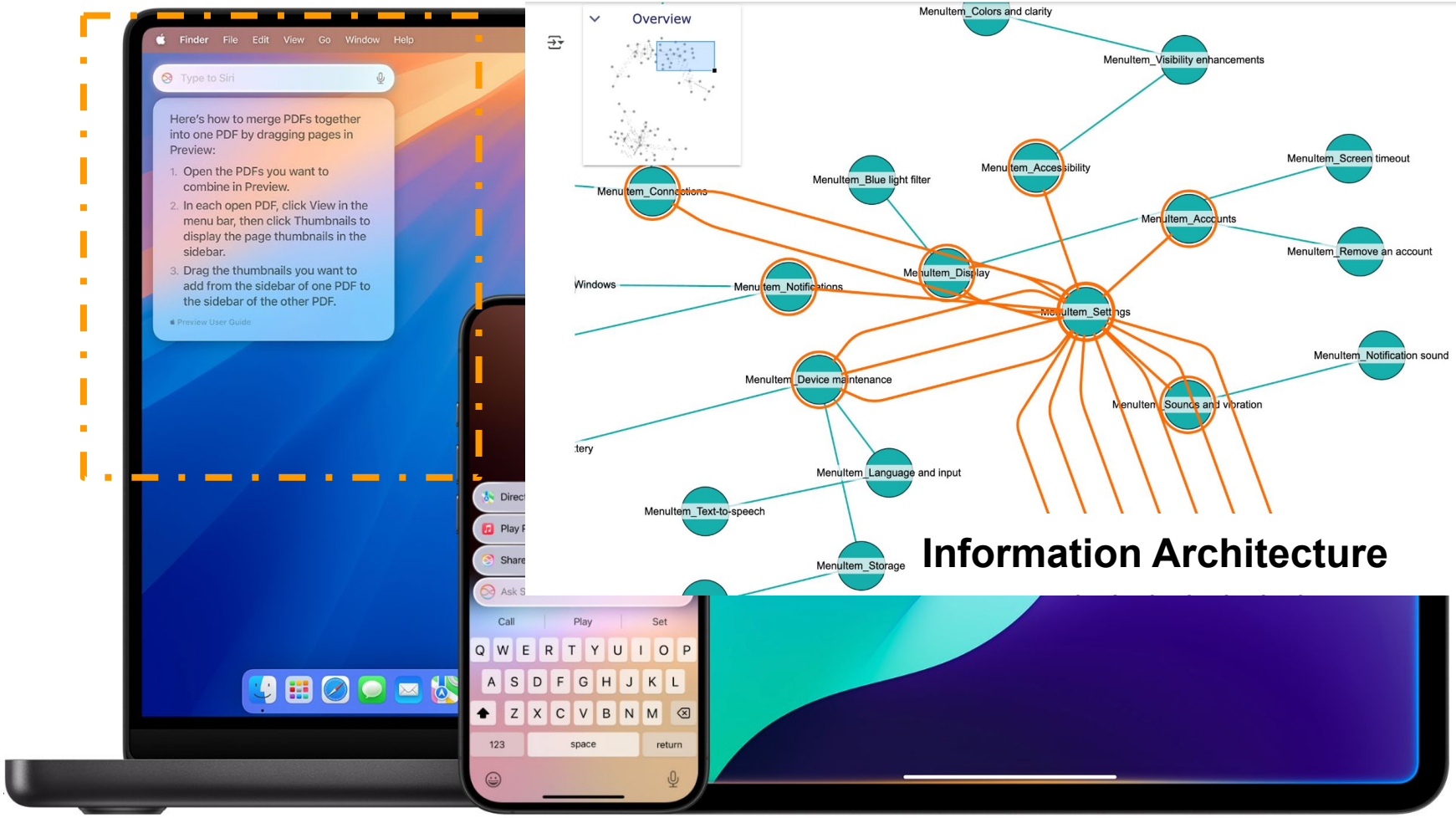
Generative AI use cases, nonexhaustive

Modality	Application	Example use cases
Text	Content writing	Marketing: creating personalized emails and posts Talent: drafting interview questions, job descriptions
	Chatbots or assistants	Customer service: using chatbots to boost conversion on websites
	Search	Making more natural web search Corporate knowledge: enhancing internal search tools
	Analysis and synthesis	Sales: analyzing customer interactions to extract insights Risk and legal: summarizing regulatory documents
Code	Code generation	IT: accelerating application development and quality with automatic code recommendations
	Application prototype and design	IT: quickly generating user interface designs
	Data set generation	Generating synthetic data sets to improve AI models' quality
Image	Stock image generator	Marketing and sales: generating unique media
	Image editor	Marketing and sales: personalizing content quickly
Audio	Text to voice generation	Trainings: creating educational voiceover
	Sound creation	Entertainment: making custom sounds without copyright violations
	Audio editing	Entertainment: editing podcast in post without having to rerecord
3-D or other	3-D object generation	Video games: writing scenes, characters Digital representation: creating interior-design mockups and virtual staging for architecture design
	Product design and discovery	Manufacturing: optimizing material design Drug discovery: accelerating R&D process
Video	Video creation	Entertainment: generating short-form videos for TikTok Training or learning: creating video lessons or corporate presentations using AI avatars
	Video editing	Entertainment: shortening videos for social media E-commerce: adding personalization to generic videos Entertainment: removing background images and background noise in post
	Voice translation and adjustments	Video dubbing: translating into new languages using AI-generated or original-speaker voices Live translation: for corporate meetings, video conferencing Voice cloning: replicating actor voice or changing for studio effect such as aging
	Face swaps and adjustments	Virtual effects: enabling rapid high-end aging; de-aging; cosmetic, wig, and prosthetic fixes Lip syncing or "visual" dubbing in postproduction: editing footage to achieve release in multiple ratings or languages Face swapping and deep-fake visual effects Video conferencing: real-time gaze correction

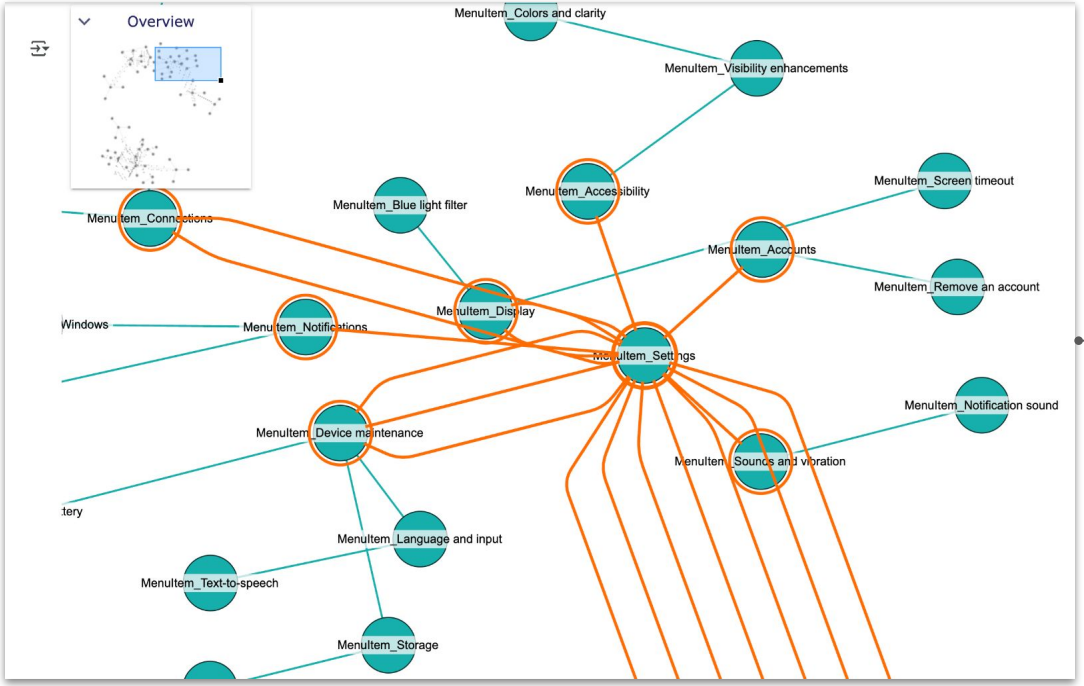
McKinsey & Company.

Graph RAG

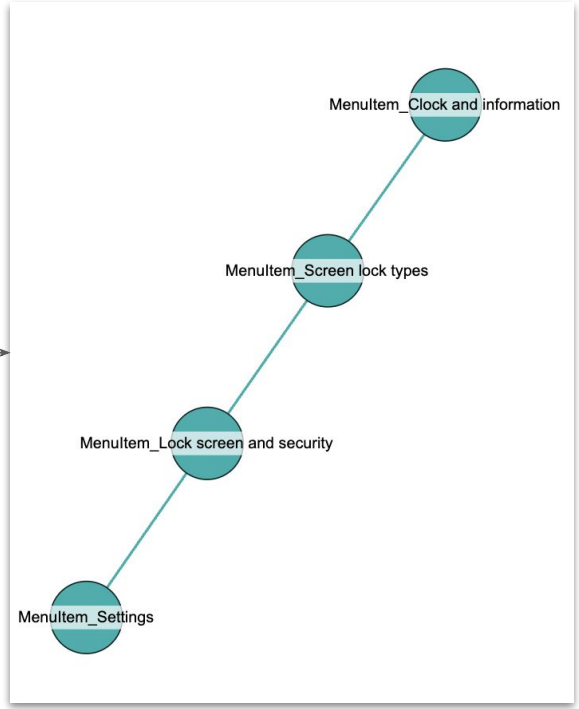
Modality	Application	Example Use Case
Text	Chatbot Assistant	Customer Service: Using ChatBot to Answer Device related Queries using information available in User Manual


Finder File Edit View Go Window Help
Type to Siri
Here's how to merge PDFs together into one PDF by dragging pages in Preview:
1. Open the PDFs you want to combine in Preview.
2. In each open PDF, click View in the menu bar, then click Thumbnails to display the page thumbnails in the sidebar.
3. Drag the thumbnails you want to add from the sidebar of one PDF to the sidebar of the other PDF.
Preview User Guide
Overview
MenuItem_Colors and clarity
MenuItem_Visibility enhancements
MenuItem_Accessibility
MenuItem_Blue light filter
MenuItem_Connections
MenuItem_Screen timeout
MenuItem_Accounts
MenuItem_Remove an account
MenuItem_Display
Windows
MenuItem_Notifications
MenuItem_Settings
MenuItem_Notification sound
MenuItem_Device maintenance
MenuItem_Sounds and vibration
MenuItem_Language and input
MenuItem_Text-to-speech
MenuItem_Storage
Information Architecture
Call Play Set
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123 space return

Information Architecture

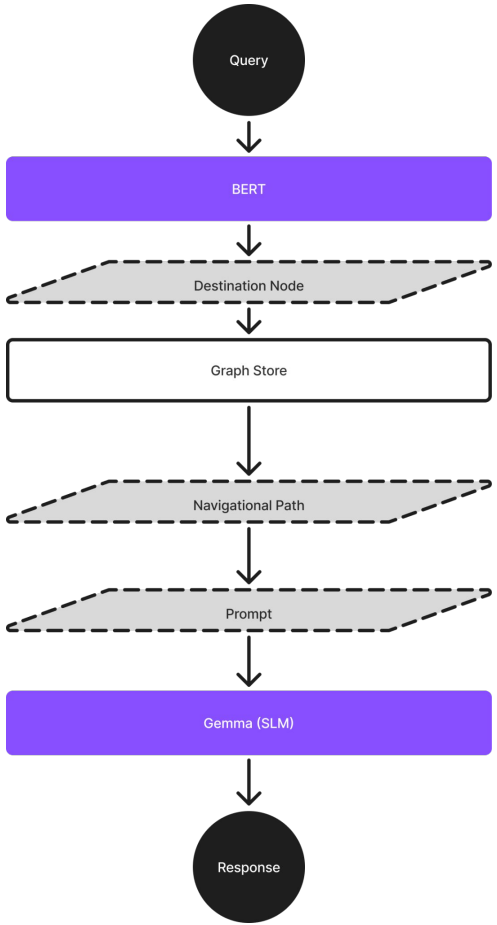


Navigational Path

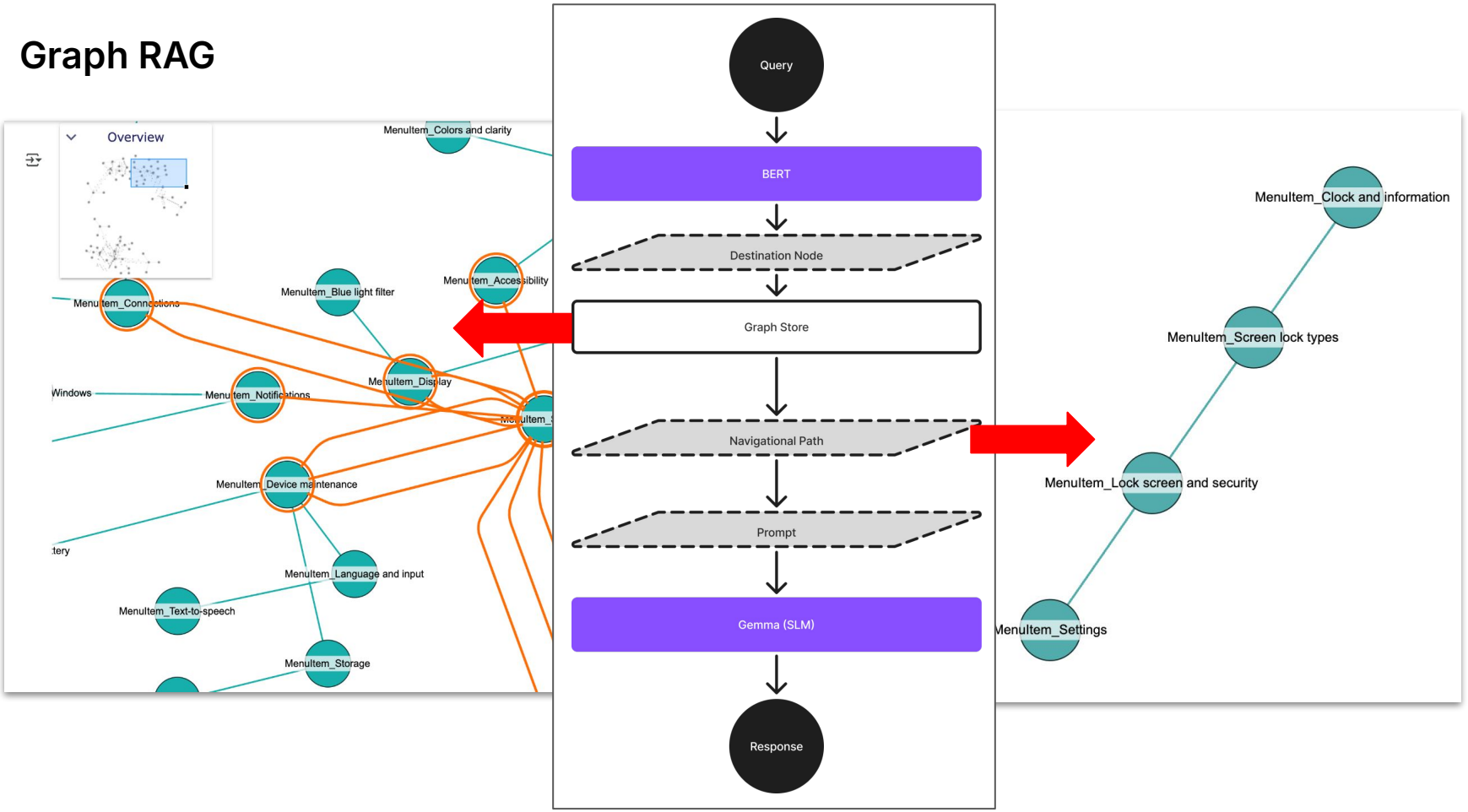


Graph RAG

Modality	Application	Example Use Case
Text	Chatbot Assistant	Customer Service: Using ChatBot to Answer Device related Queries using information available in User Manual



Graph RAG


Overview
MenuItem_Colors and clarity
MenuItem_Blue light filter
MenuItem_Accessibility
MenuItem_Connections
MenuItem_Display
Windows
MenuItem_Notifications
MenuItem_Device maintenance
MenuItem_Language and input
MenuItem_Text-to-speech
MenuItem_Storage
Query
BERT
Destination Node
Graph Store
Navigational Path
Prompt
Gemma (SLM)
Response
MenuItem_Clock and information
MenuItem_Screen lock types
MenuItem_Lock screen and security
MenuItem_Settings

README Apache-2.0 license

Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework

This repo has the code for the paper **"Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework"** accepted at **EMNLP 2021 Findings**. The blog on this paper can be found here, the poster here, and a corresponding presentation here.

Required dependencies -

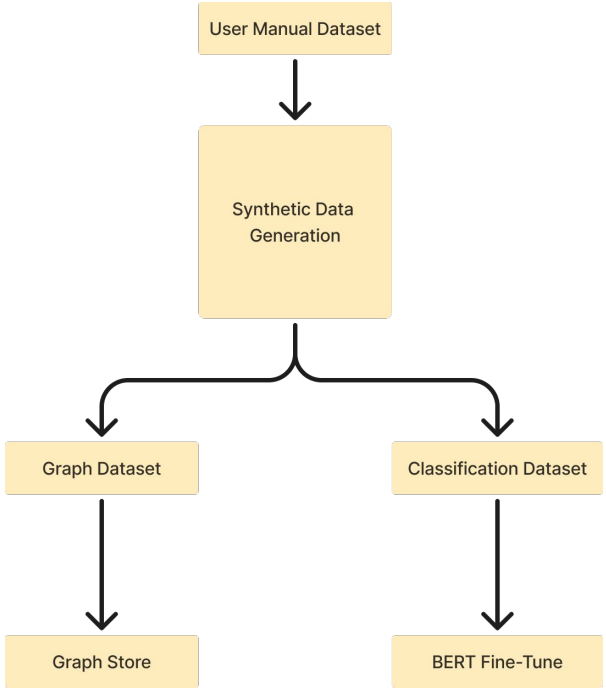
Please run `pip install -r requirements.txt` (`python3` required)

E-Manual pre-training corpus

Go to this link. A RoBERTa BASE Model pre-trained on the corpus can be found here, and a BERT BASE UNCASED Model pre-trained on the same here.

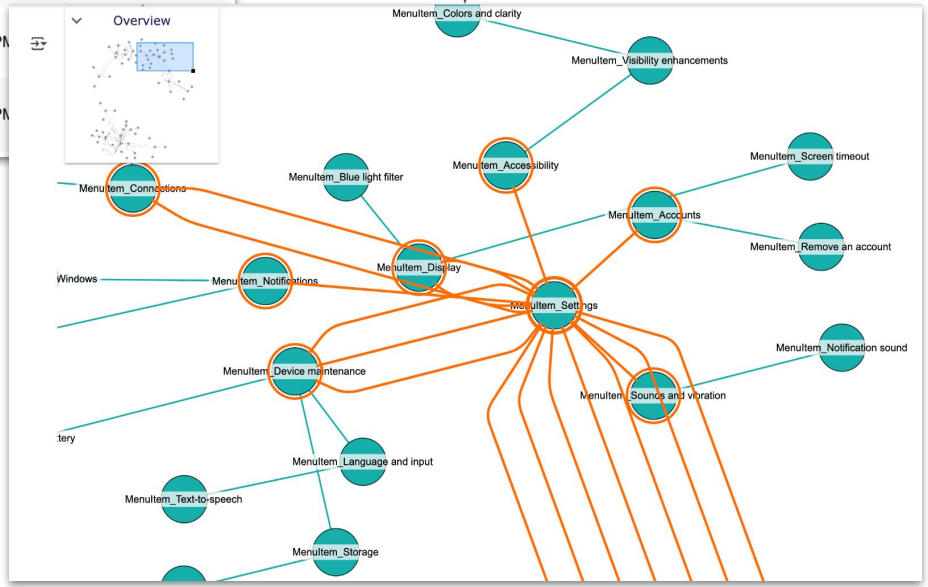
Codes

- Annotated Data and Amazon User Forum Data Samples are present in `data` (See README)
- Data Analysis is done in `data_analysis` (See README)



```
s10_file = 's10_50_questions.csv'
s10 = pd.read_csv(s10_file)
s10.head()
```

	question_id	question	section_path	section_link	answer
0	266	What are the advanced recording options availa...	Apps>Samsung apps>Camera>Camera settings	http://downloadcenter.samsung.com/content/PM/2...	Use the icons on the main camera screen and th...
1	37	How can I turn on the GPS ?	Settings>Lock screen and security>Location	http://downloadcenter.samsung.com/content/PM/2...	Location services use a combination of GPS, mo...
2	575	How to apply custom themes for AOD?	Getting started>Start using your device>Always...	http://downloadcenter.samsung.com/content/PM/2...	Apply custom themes for Always On Display. Fro...
3	518	How can I show notifications on always display?	Settings>Lock screen and security>Screen lock ...	http://downloadcenter.samsung.com/content/PM	
4	481	How can I use time zone converter?	Apps>Samsung apps>Clock>Time zone converter	http://downloadcenter.samsung.com/content/PM	



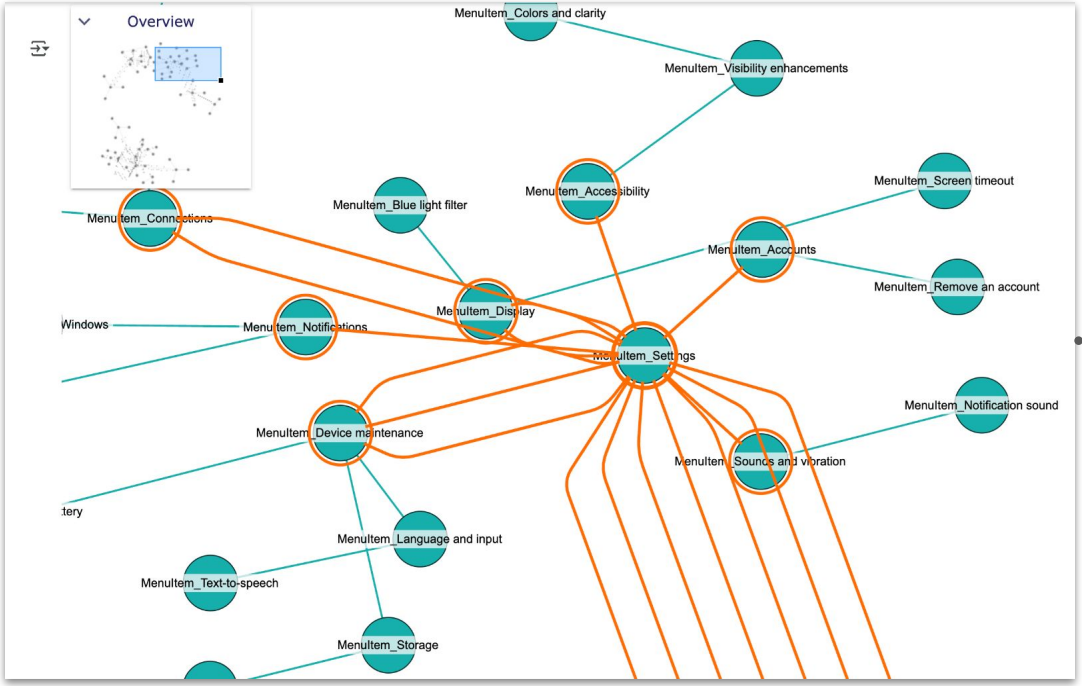

MenuItem_Screen lock types
MenuItem_Lock screen and security
Properties
{
"_id": {
"offset": 11,
"table": 0
},
"_label": "MenuItem",
"name": "Screen lock types",
"description": "Screen lock types refer to the methods you…",
"MenuItem": true,
"label": "MenuItem_Screen lock types"
}

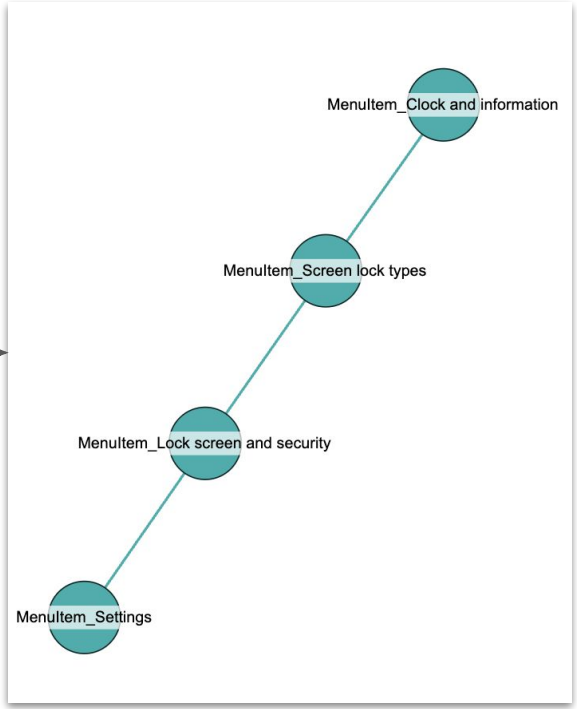
Design Prompt for generating descriptions of each menu

```
[ ] def create_prompt(menu_item,text):
      return f"""
        You are a Content Writer, with responsibility to write User Manual:
        Here is an arbitary TEXT on one of the menu items -'{menu_item}'.
        ------TEXT------
        {text}
        -------TEXT-------
         Without any navigational information; Describe in 4 short sentences \"What is '{menu_item}'\"
      """
```

	name	description
0	Apps	"Apps" is a menu providing access to your devi...
1	Samsung apps	Samsung apps is a collection of pre-installed ...
2	Camera	The Camera feature lets you take photos and vi...
3	Camera settings	Camera settings allow you to customize your ca...
4	Settings	Settings is your device's central control pane...
...	...	...

Information Architecture → BERT → Navigational Path





Resources

1. What are Small Language Models (SLM)? | IBM
2. Quantizing ONNX Models using Intel® Neural Compressor
3. Generative AI with Large Language Models | Coursera
4. Improving Recommendation Systems & Search in the Age of LLMs
5. ONNX Runtime | Getting-started
6. SeedLM: Compressing LLM Weights into Seeds of Pseudo-Random Generators - Apple Machine Learning Research

Thank You

Open for Questions