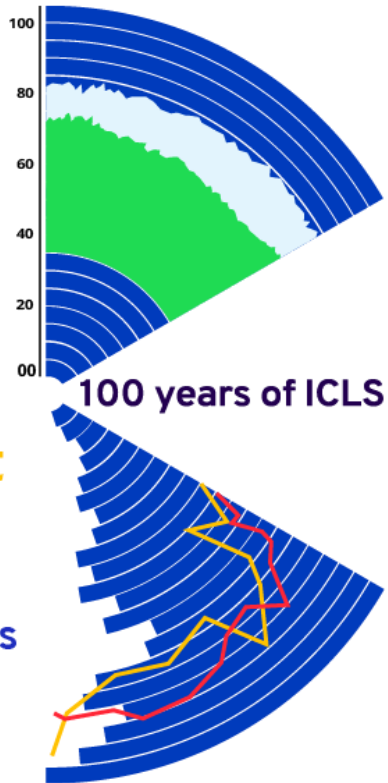




International  
Conference of  
Labour Statisticians  
11-20 October 2023



**Updating occupational classifications  
using machine learning:**

**Pilot test with Azure AI, ChatGPT, and  
acquiring new data sources (Ongoing)**

Shutong DING  
Knowledge Management Officer, ILO Statistics

## ► From the PoC to production

### Sustainability, flexibility & scalability

- Data storage
- Different languages for global reach
- Seamless integration with data, AI tools and services
- Easy tuning & sustainable model
- Minimized human intervention -> automation
- Feedback loop -> continuous learning
- New data sources
- Cost-effective

### Need solutions for

- Infrastructure
- Tools
- More data

## Cloud based platform – Azure

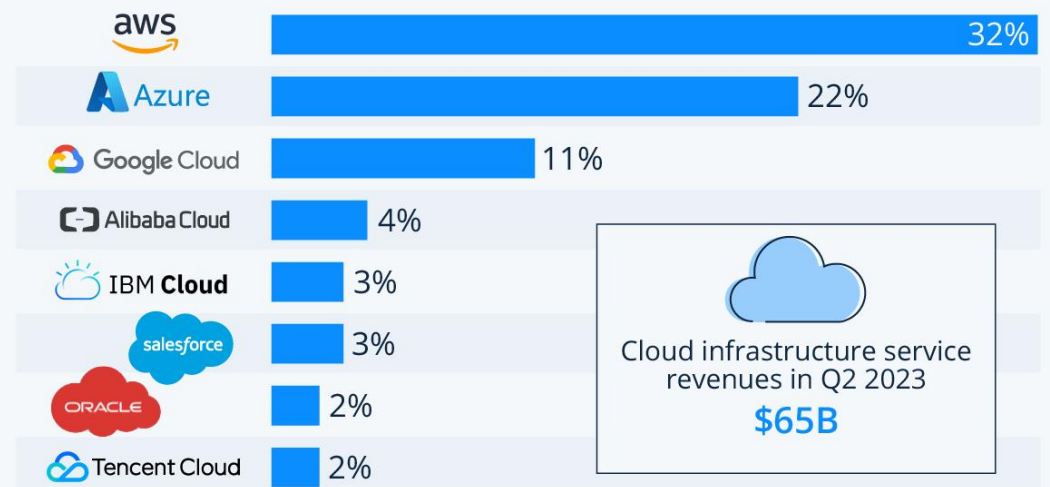
### Why Azure?

- ▶ Leading cloud service provider
- ▶ Natural integration with MS applications (.NET)
- ▶ Latest partnership with OpenAI

### Relevant Azure services

- ▶ Azure Blob Storage
- ▶ **Azure ML Workspace/Studio**
- ▶ Azure Translator
- ▶ **Azure OpenAI - ChatGPT**

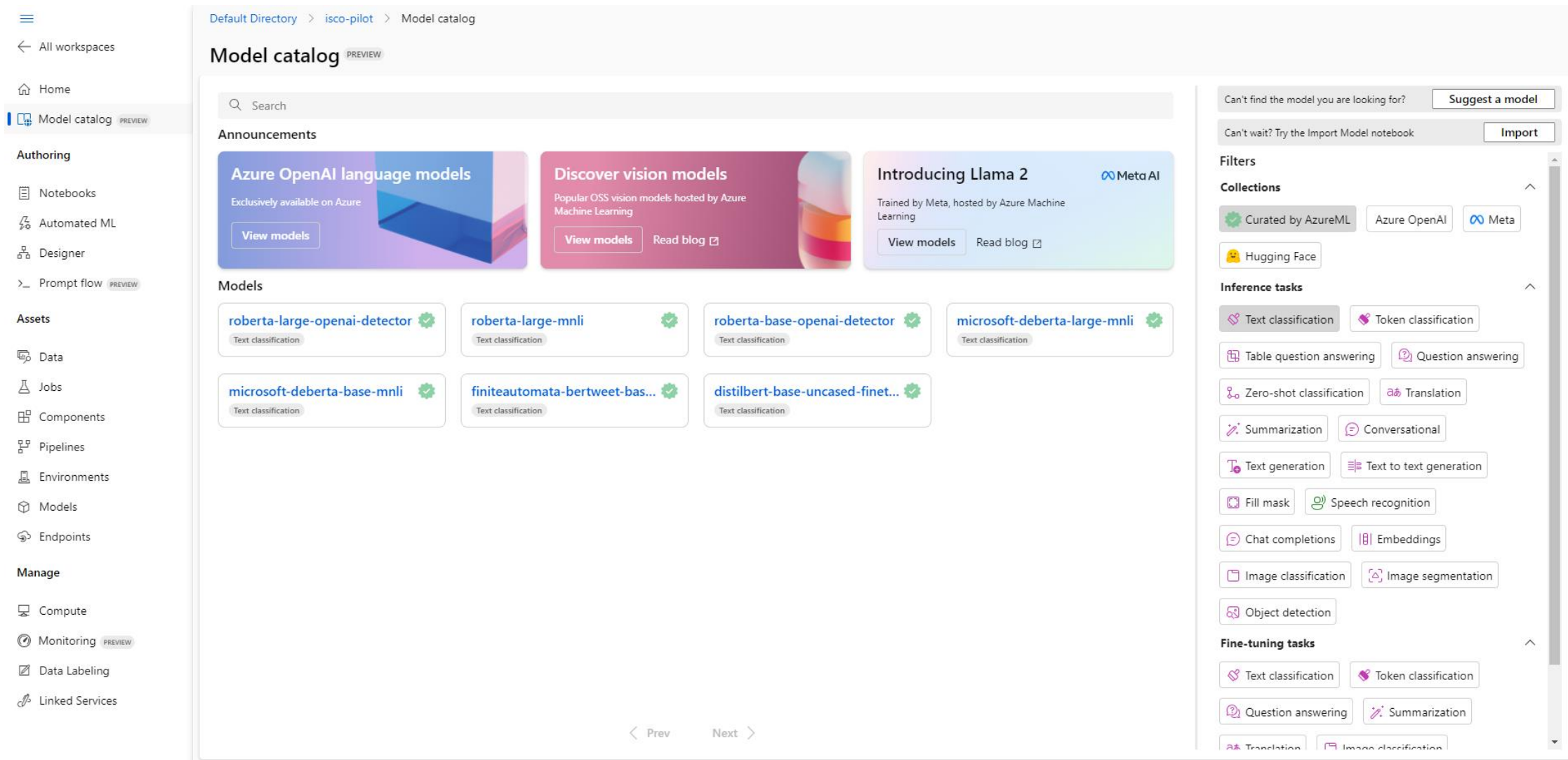
Worldwide market share of leading cloud infrastructure service providers in Q2 2023\*



\* Includes platform as a service (PaaS) and infrastructure as a service (IaaS) as well as hosted private cloud services

Source: Synergy Research Group

# A snapshot of Azure ML Workspace



The screenshot displays the Azure ML Workspace interface, specifically the 'Model catalog' section. The left sidebar contains navigation links for 'All workspaces', 'Home', 'Model catalog' (selected), 'Authoring', 'Notebooks', 'Automated ML', 'Designer', 'Prompt flow', 'Assets', 'Data', 'Jobs', 'Components', 'Pipelines', 'Environments', 'Models', 'Endpoints', 'Manage', 'Compute', 'Monitoring', 'Data Labeling', and 'Linked Services'. The main content area is titled 'Model catalog' and includes a search bar, 'Announcements' (featuring 'Azure OpenAI language models', 'Discover vision models', and 'Introducing Llama 2'), and a 'Models' section listing various text classification models like 'roberta-large-openai-detector', 'roberta-large-mnli', 'roberta-base-openai-detector', 'microsoft-deberta-large-mnli', 'microsoft-deberta-base-mnli', 'finiteautomata-bertweet-bas...', and 'distilbert-base-uncased-finet...'. The right sidebar shows filters for 'Collections' (Curated by AzureML, Azure OpenAI, Meta, Hugging Face) and 'Inference tasks' (Text classification, Token classification, Table question answering, Question answering, Zero-shot classification, Translation, Summarization, Conversational, Text generation, Text to text generation, Fill mask, Speech recognition, Chat completions, Embeddings, Image classification, Image segmentation, Object detection). A 'Fine-tuning tasks' section is also visible at the bottom right. Navigation buttons 'Prev' and 'Next' are at the bottom center.

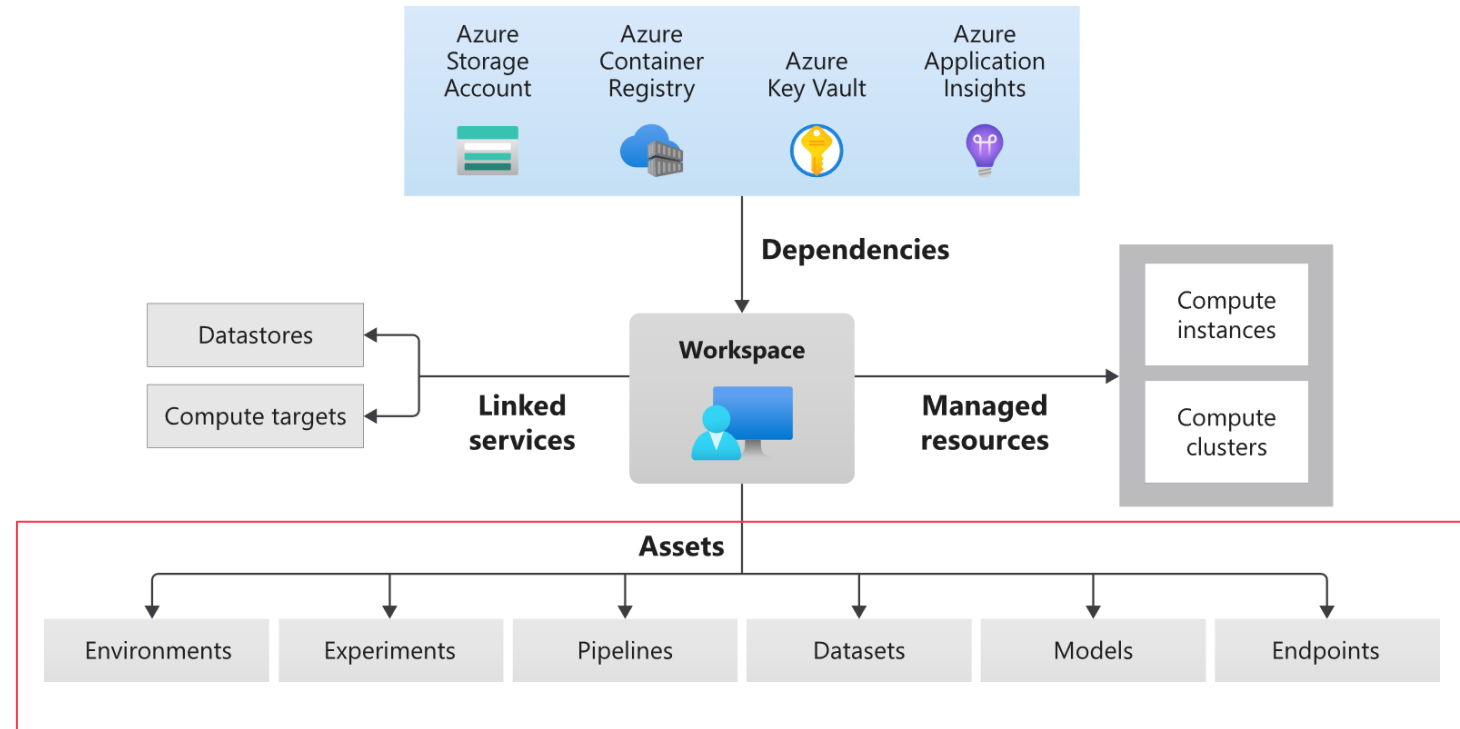
## Azure ML Workspace

### Assets

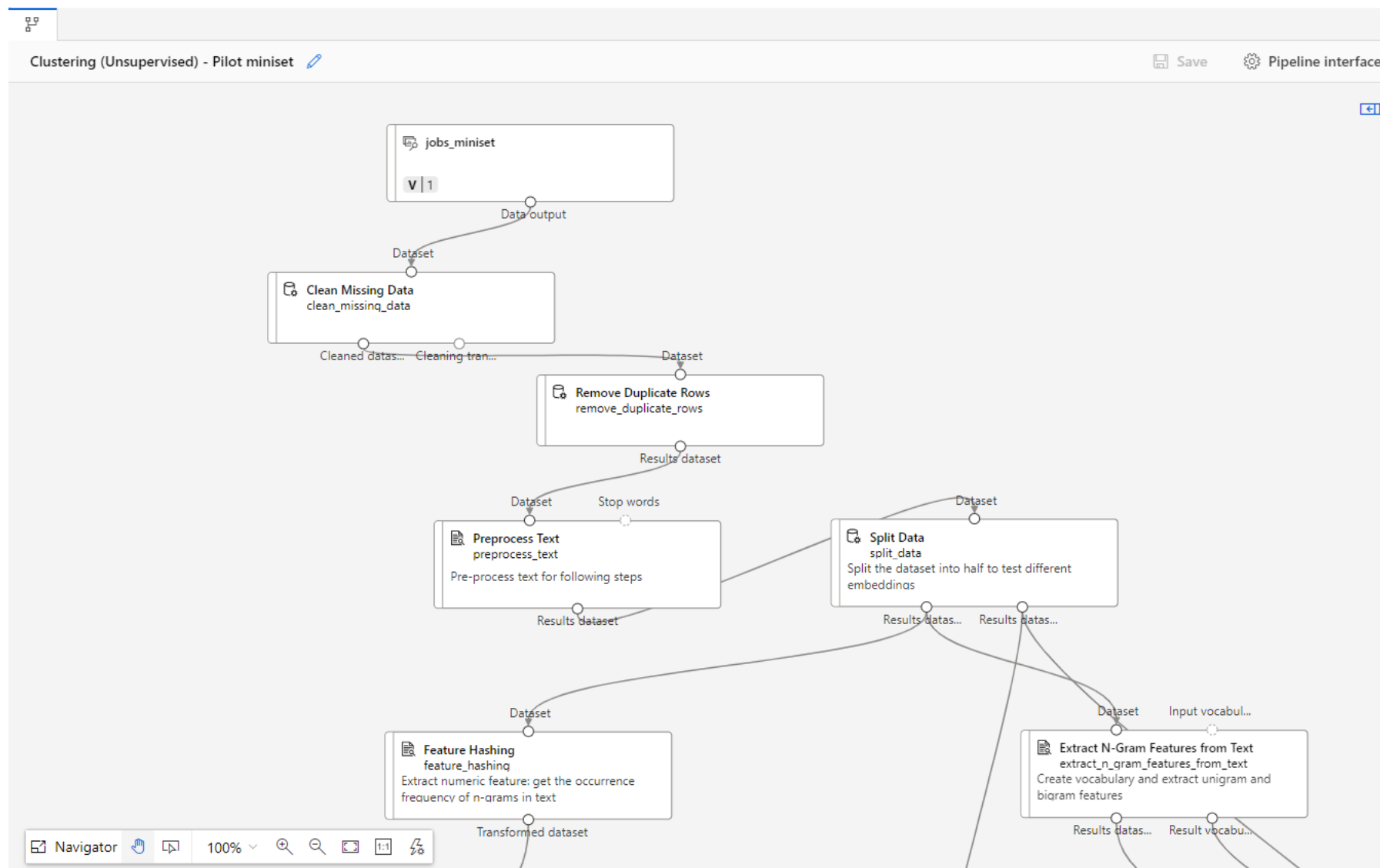
- ▶ Data
- ▶ Components
- ▶ Pipelines
- ▶ Models
- ▶ Jobs

### Manage

- ▶ Compute
- ▶ Monitoring



# Example of an Azure AI pipeline



## ChatGPT



### State-of-the-art AI-powered human-like text generator

- ▶ Information retrieval
- ▶ Dynamic content creation
- ▶ Programming aid
- ▶ Translations

### Based on Large Language Models (LLM)

- ▶ gpt-4/gpt-4-32k
- ▶ gpt-35-turbo/gpt-35-turbo-16k
- ▶ text-embedding-ada-002

## Test ChatGPT Usage 1: Direct detection

### Specification

- **Data:** 1120 rows, job\_title, job\_desc
- **Method:** API
- **Model:** gpt-35-turbo & gpt-4;  
ChatCompletion
- **Randomness:** top\_p = 0
- **Detection:** “not elsewhere” founded in  
any top 5 matches

### Pre-cleaning

- Lower case
- Stop words removed

	Unnamed: 0.1	Unnamed: 0	source	company	job_title	job_desc	clean_MJT	clean_MJD	pred_label
0	0	0	Idealist-Intl	STARS (St. Andrew's Refugee Services)	Resettlement Legal Officer	Founded in 1979, STARS is a refugee service pr...	resettlement legal officer	founded 1979 stars refugee service provider c...	3411
1	1	1	Idealist-Intl	The Cooper Union	ASSISTANT DIRECTOR OF ANNUAL GIVING	JOB SUMMARY: The Office of Alumni Affairs and ...	assistant director of annual giving	job summary the office alumni affairs develop...	1120
2	2	2	Idealist-Intl	Life Span	OPERATIONS ASSISTANT	Life Span empowers survivors of domestic and s...	operations assistant	life span empowers survivors domestic sexual v...	4110
3	3	3	Idealist-Intl	Life Span	Resource Advocate	Life Span is seeking to fill our community Res...	resource advocate	life span seeking fill community resource advo...	2423
4	4	4	Idealist-Intl	NaN	Area Director, Washington State	POSITION: Area Director, Washington State REPO...	area director washington state	position area director washington state repo...	1120

```
query = (f"Find top {num_matches} matches of ISCO-08
4-digit code for title '{job_title}' "
        f"with description '{job_desc}'? "
        "return only ISCO-08 4-digit code and
its job category")
```



## ► Test ChatGPT Usage 1: Direct detection

### Examples of detection

#### ► Senior Ethereum Blockchain Analyst

1. Code: 2133 - "ICT Systems Analysts"
2. Code: 2512 - "Software Developers"
3. Code: 2511 - "Systems Analysts"
4. Code: 1330 - "Information and Communications Technology Services Managers"
5. Code: 2529 - "Software and Applications Developers and Analysts **Not Elsewhere** Classified"

#### ► Drone Operator

1. Code: 3155 - "Ship and aircraft controllers and technicians"
2. Code: 2144 - "Mechanical engineers"
3. Code: 2149 - "Engineering professionals **not elsewhere** classified"
4. Code: 3114 - "Electronics engineering technicians"
5. Code: 2166 - "Graphic and multimedia designers"

#### ► Service Delivery Lead

1. 2512 - ICT Operations and Service Managers
2. 2511 - Systems Administrators
3. 2513 - Software and Applications Managers
4. 2147 - Computer Engineers (except Software Engineers and Designers)
5. 2133 - Electrical and Electronics Engineers

## ► Test ChatGPT Usage 1: Direct detection

### Results and lessons learned

- Promising performance
- Easy to further extract information from job descriptions
- gpt-4 is more robust and consistent

### However,

- Outdated: Doesn't reflect the latest ISCO index
- Lacks a feedback mechanism
- Slow response:
  - **gpt-3.5**: Average 3.6s/query - **gpt-4**: Average 5.8s/query
- Expansive: \$42k for 1 million queries (gpt-4)
- Restrictive rate limits:
  - **tokens-per-minute**: 10k - **requests-per-minute**: 200

## ► Test ChatGPT Usage 2: Post-processing

### Summarization of clusters

- Normalization of titles
- Job descriptions insights
  - Skill requirements
  - Core tasks
  - Educational Prerequisites

4	agile coach scrum, scrum master, apmc scrum master, associate scrum master, digital scrum master, dynamics scrum master, fednow scrum master, guidewire scrum master, intermediate scrum master, kanban program scrum master, kanban scrum master, lead scrum master agile, lead scrum master agility, associate scrum master, ms dynamics scrum master, required scrum master, scrum master, scrum master active, scrum master adas systems, scrum master associate, scrum master digital, scrum master digital agile government, scrum master functional, scrum master functional engineer, scrum master iii, scrum master job, scrum master joiners, scrum master lead developer, scrum master nityo, scrum master pan, scrum master product, scrum master rpa, scrum master telecom, senior scrum contract, senior scrum master digital, senior scrum master, technical scrum master agile coach
---	--



Agile and Scrum Specialists

## ► Test ChatGPT Usage 2: Post-analysis

### Normalized title of clusters (part)

Cluster No. 2: Supply Chain Planners

Cluster No. 4: Agile and Scrum Specialists

Cluster No. 5: Salesforce Engineers

Cluster No. 6: AI and ML Specialists

Cluster No. 7: Language and Recruitment Specialists

Cluster No. 9: UX/UI Designers

Cluster No. 12: HR Recruiters

Cluster No. 13: Operational Risk Managers

Cluster No. 14: Talent Acquisition Specialists

Cluster No. 16: Procurement Associates

Cluster No. 47: Cloud and Network Security Engineers

Cluster No. 48: Embedded and IC Design Engineers

Cluster No. 49: Blockchain and Software Developers

Cluster No. 50: Risk Management Specialists

Cluster No. 51: Healthcare Professionals

Cluster No. 53: Automation Testers

Cluster No. 54: Quality Assurance Managers

Cluster No. 55: Financial and Fraud Risk Managers

Cluster No. 56: Hospitality Workers

Cluster No. 57: Big Data Developers

Cluster No. 59: Reconciliation Officers

Cluster No. 60: Java Developers and Operation VPs

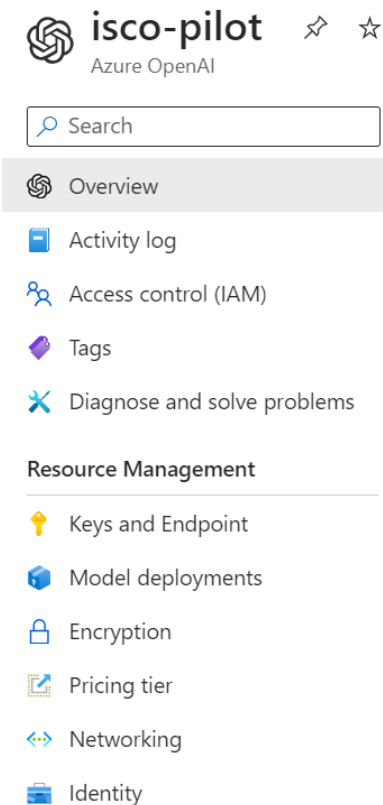
## ► Potential usage of ChatGPT

### Lessons learned so far

- Direction detection is costly and inefficient
- Best for its information retrieval
- It can be used as specific components

### Further experiments

- Embeddings
- Azure **OpenAI** (ongoing)
  - Customized model with our own data (i.e., ISCO index)
  - Feedback loop



## ► Acquiring new data sources

### Job vacancies

- Web scraping
- Partnership with job portals

### Beyond (mapping)

- Updated indexes from member states (SSOC, CNOC, ...)
- Other expert data collection

### Open invitation

- Seeking more data partnerships!

#### Buscojobs (Data Partner)

- Provides ILO with free data access
- Based in Uruguay
- Currently covers 34 markets globally
- Both job posts and CVs
- English, Spanish, Portuguese, Italian, and Malay

## ► Summary – way forward

- **Azure**

- Break down the PoC's process into modular **components** for reusability
- Construct **pipelines** by assembling these components and register the defined **models**
- Initiate **jobs** to run models on fresh datasets
- **Monitor** the process, export and evaluate the results
- Implement **Azure Translator** to accommodate other languages

- **ChatGPT**

- Develop customized GPT model in **Azure OpenAI**
- Define post-process components for normalization and abstraction
- Conduct tests on embeddings

- **Data**

- Identify and integrate other data sources
- Enhance **collaborations** and promote data sharing

► **Thank you!**

[ding@ilo.org](mailto:ding@ilo.org)

