Benchmarking embedding models

Introduction

Benchmarking embedding models

- Embedding models **convert text** (or other modalities) **to a dense vector**.
- With embedding models, you can **build RAG and recommendation systems**.
- But **how do you choose an embedding model**?

Hi Cassiopeia! Have you talked with Andromeda?

↓

Embedding model

↓

[0.69, 0.42, ...]

Benchmarking embedding models

- How do you choose an embedding model?
- You look at **benchmarks**!
- MTEB ranks models based on their performance on **different benchmarks**.

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …	Classification	Clustering
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	70.58	61.69	80.89	74.00	57.65
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33
5	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77
6	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60
7	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75
8	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17
9	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84
10	GritLM-7B	99%	13813	7B	4096	32768	60.92	53.74	70.53	61.83	49.75
11	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84

The MTEB leaderboard on 20/09/2025.

Benchmarking embedding models

- Should we **trust** these public benchmarks?
- You can also create **your own benchmark** with your private data.
- The dataset must be **clean, diverse, and in your language** (or multilingual).

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …	Classification	Clustering
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	70.58	61.69	80.89	74.00	57.65
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33
5	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77
6	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60
7	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75
8	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17
9	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84
10	GritLM-7B	99%	13813	7B	4096	32768	60.92	53.74	70.53	61.83	49.75
11	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84

The MTEB leaderboard on 20/09/2025.

Benchmarking embedding models

In this course, you will:

- Create a **golden dataset**.
- Run **open source** and **proprietary models**.
- Compute **metrics** to grade the models.
- Perform **statistical tests** to prove if a model is better than another.
- **Automate** some steps in the pipeline.
- **Generate tables** to compare the models.

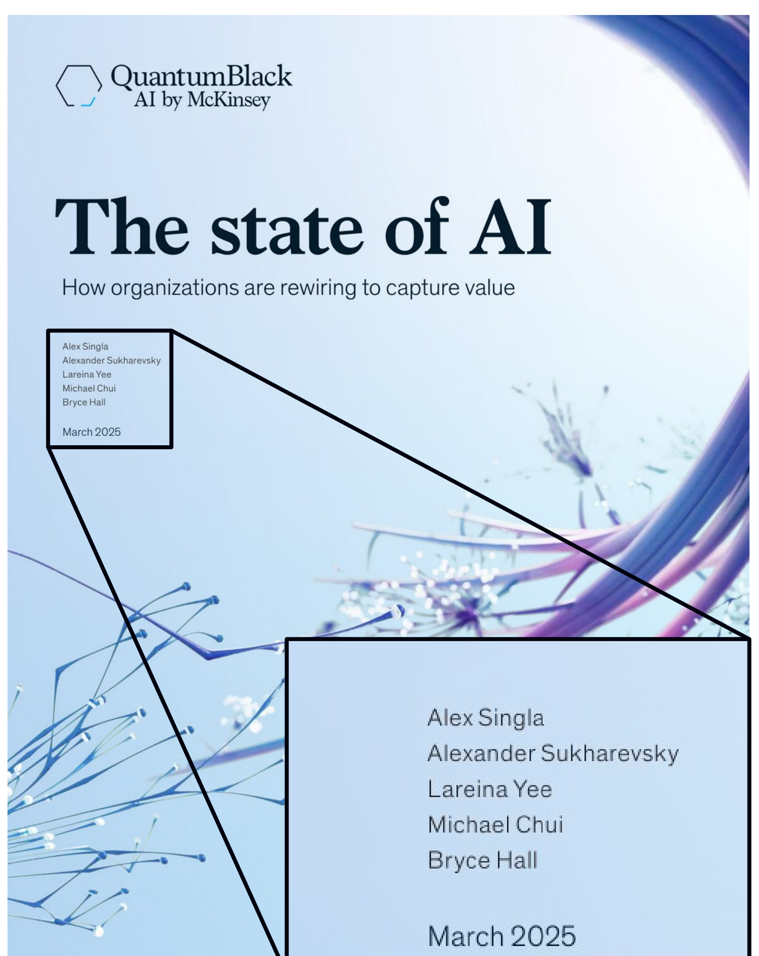
#	Model	mrr	recall@1	recall@5	ndcg@5
a	OpenAI 3-Small	0.780	0.683	0.901	0.804
b	OpenAI 3-Large	0.778	0.681	0.901	0.802
c	Google Gemini-001	0.810 ab	0.725 b	0.927 ab	0.834 ab

Comparing Gemini embedding to OpenAI's embedding models.

Extract text from PDF files

Extract text from PDF files

- Extracting text from PDF files is challenging.
- PDF files can be scanned, have complex layouts, and contain images, tables, etc
- In Python, we can use libraries like PyMuPDF, PyPDF2, and pdfplumber.
- These libraries are ineffective if you want to preserve the structure of the document.



PyMuPDF

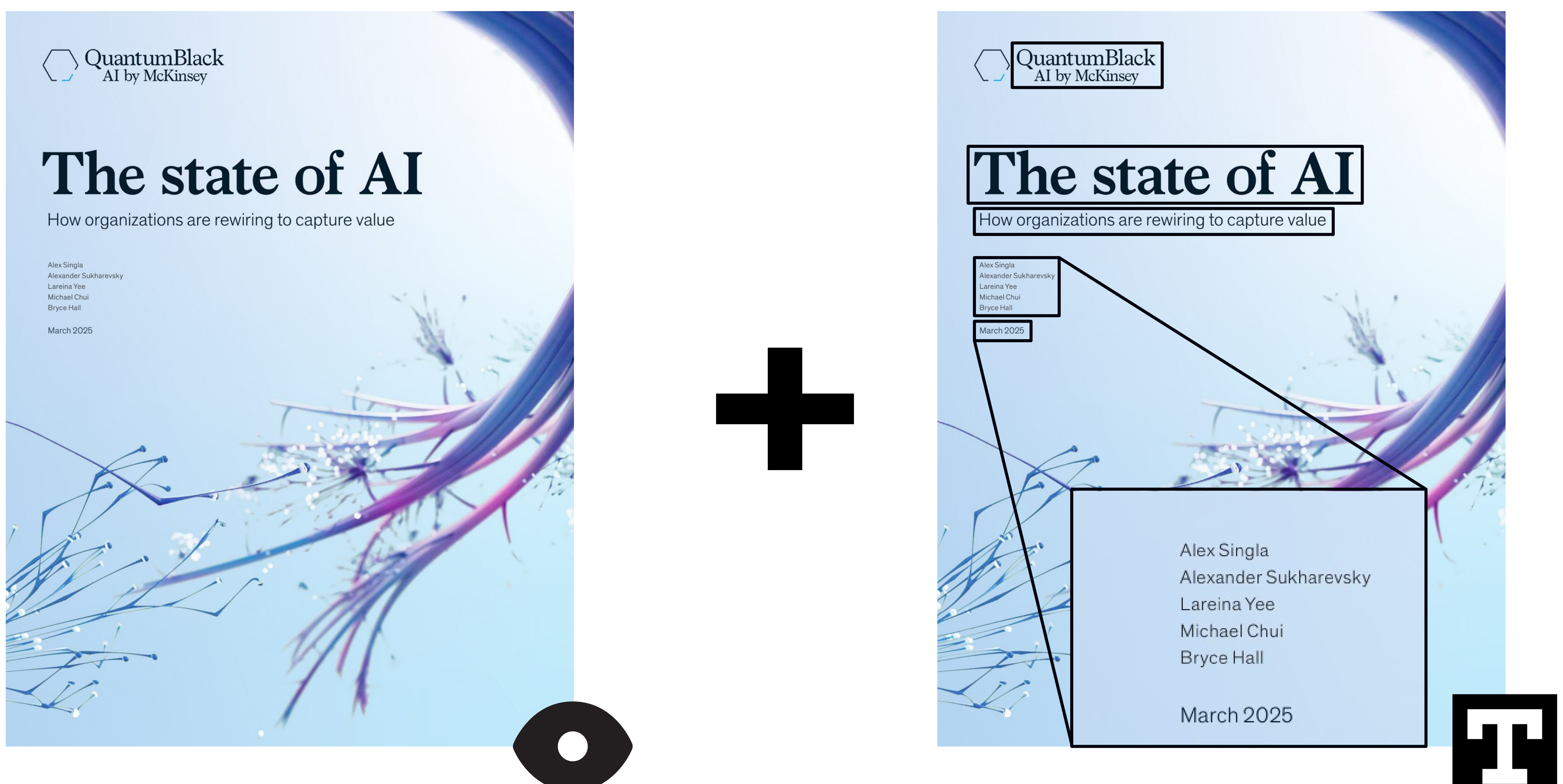
The state of AI
March 2025
Alex Singla
Alexander Sukharevsky
Lareina Yee
Michael Chui
Bryce Hall
How organizations are rewiring to capture value

Extract text from PDF files

- We can use vision language (VL) models to parse PDF documents and images.
- VL models can see images

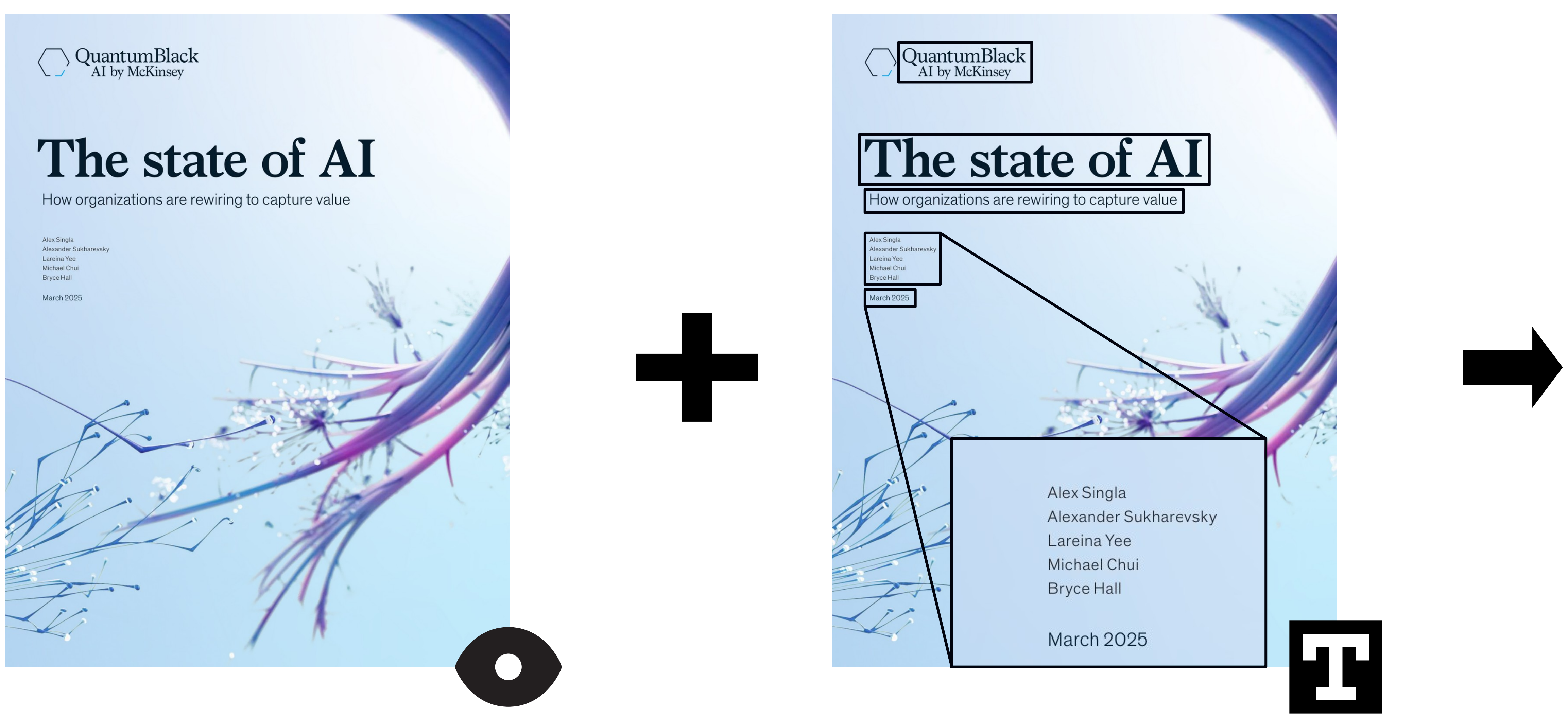
Extract text from PDF files

- We can use vision language (VL) models to parse PDF documents and images.
- VL models can see images, read and understand text.



Extract text from PDF files

- We can use vision language (VL) models to parse PDF documents and images.
- VL models can see images, read and understand text.
- The vision and language parts work together to parse the image effectively.



QuantumBlack

AI by McKinsey

The state of AI

How organizations are rewiring to capture value

Alex Singla

Alexander Sukharevsky

Lareina Yee

Michael Chui

Bryce Hall

March 2025

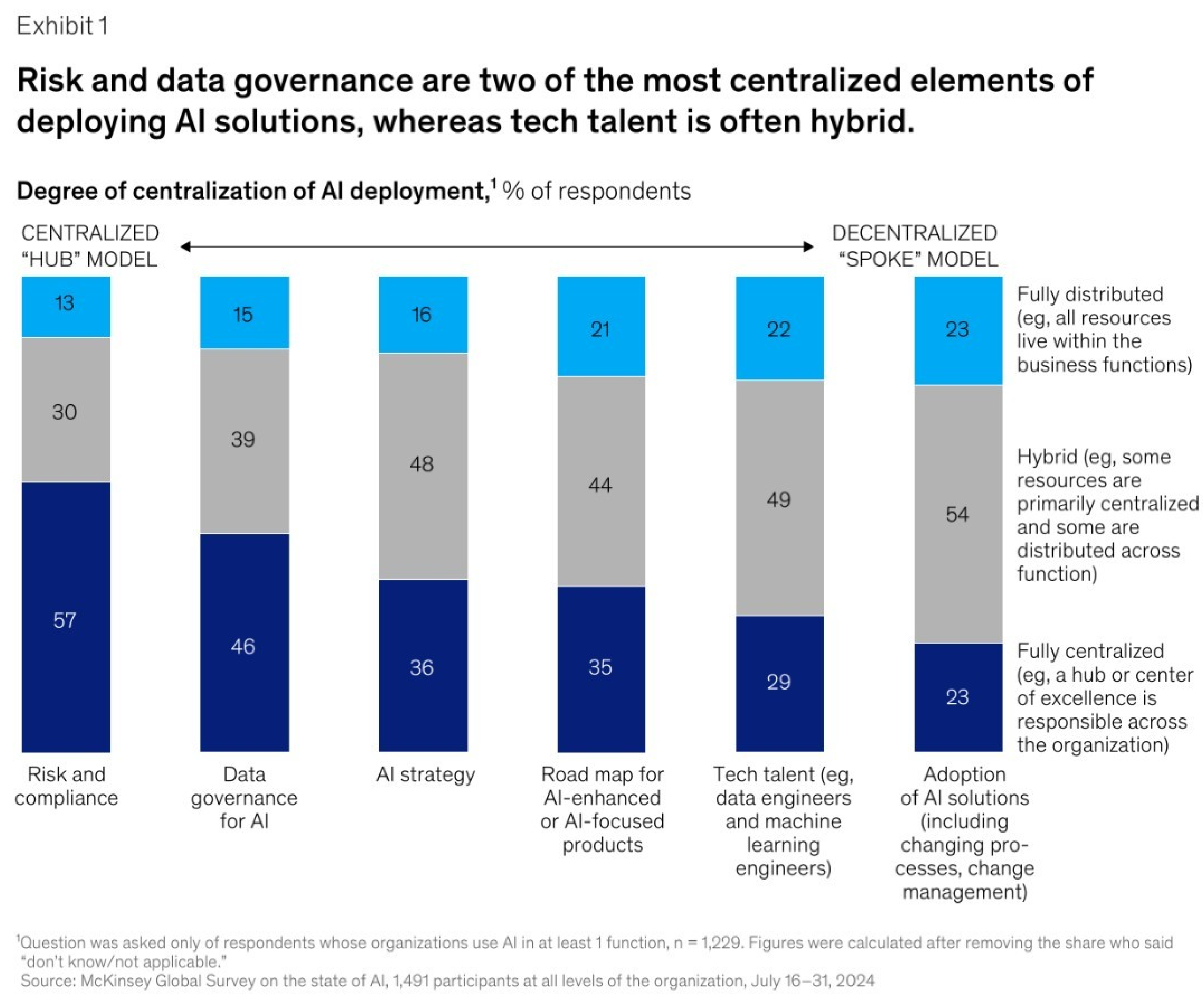
Extract text from PDF files

- What is the trade-off?
- Let’s compare both methods side-by-side.

	Python libraries	**VL models**
Cost	**Free**	**Free** / Paid
Scanned input	No	**Yes**
Resources	**Low**	High / **Low**
Preserve structure	No	**Yes**
Handle complex layouts	No	**Yes**
Understand the content	No	**Yes**
Speed	**Fast**	Slow

Extract text from PDF files

Side-by-side comparison: Test N°1



Extract text from PDF files

Side-by-side comparison: Test N°1

PyMuPDF

Exhibit 1

Degree of centralization of AI deployment,[1] % of respondents

McKinsey & Company

[1]Question was asked only of respondents whose organizations use AI in at least 1 function, n = 1,229. Figures were calculated after removing the share who said

“don’t know/not applicable.”

Source: McKinsey Global Survey on the state of AI, 1,491 participants at all levels of the organization, July 16–31, 2024

Risk and data governance are two of the most centralized elements of

deploying AI solutions, whereas tech talent is often hybrid.

57

46

36

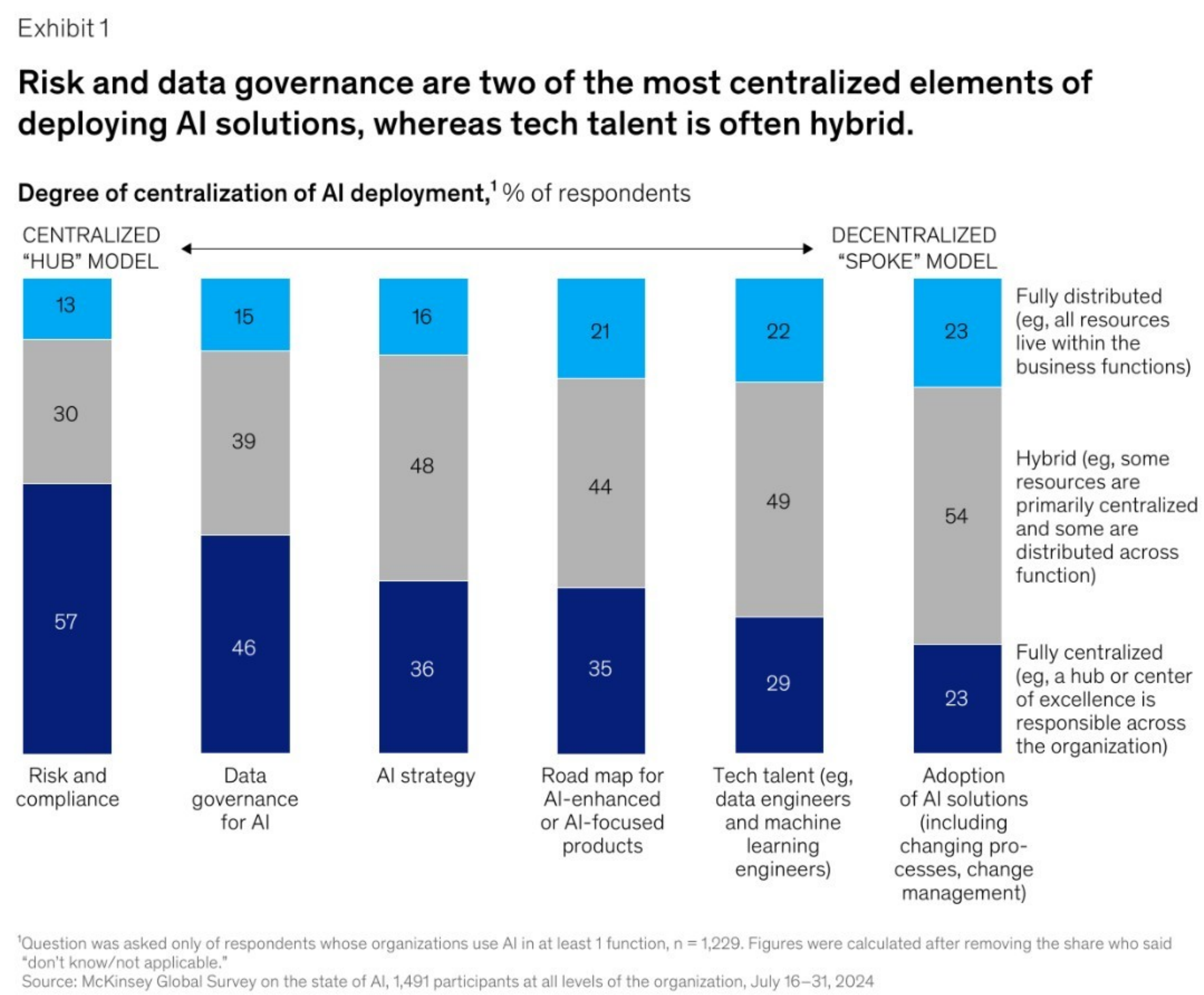
35

29

23

30

39



Extract text from PDF files

Side-by-side comparison: Test N°1

PyMuPDF

Exhibit 1

Degree of centralization of AI deployment,¹ % of respondents

McKinsey & Company

¹Question was asked only of respondents whose organizations use AI in at least 1 function, n = 1,229. Figures were calculated after removing the share who said

“don’t know/not applicable.”

Source: McKinsey Global Survey on the state of AI, 1,491 participants at all levels of the organization, July 16–31, 2024

Risk and data governance are two of the most centralized elements of

deploying AI solutions, whereas tech talent is often hybrid.

57

46

36

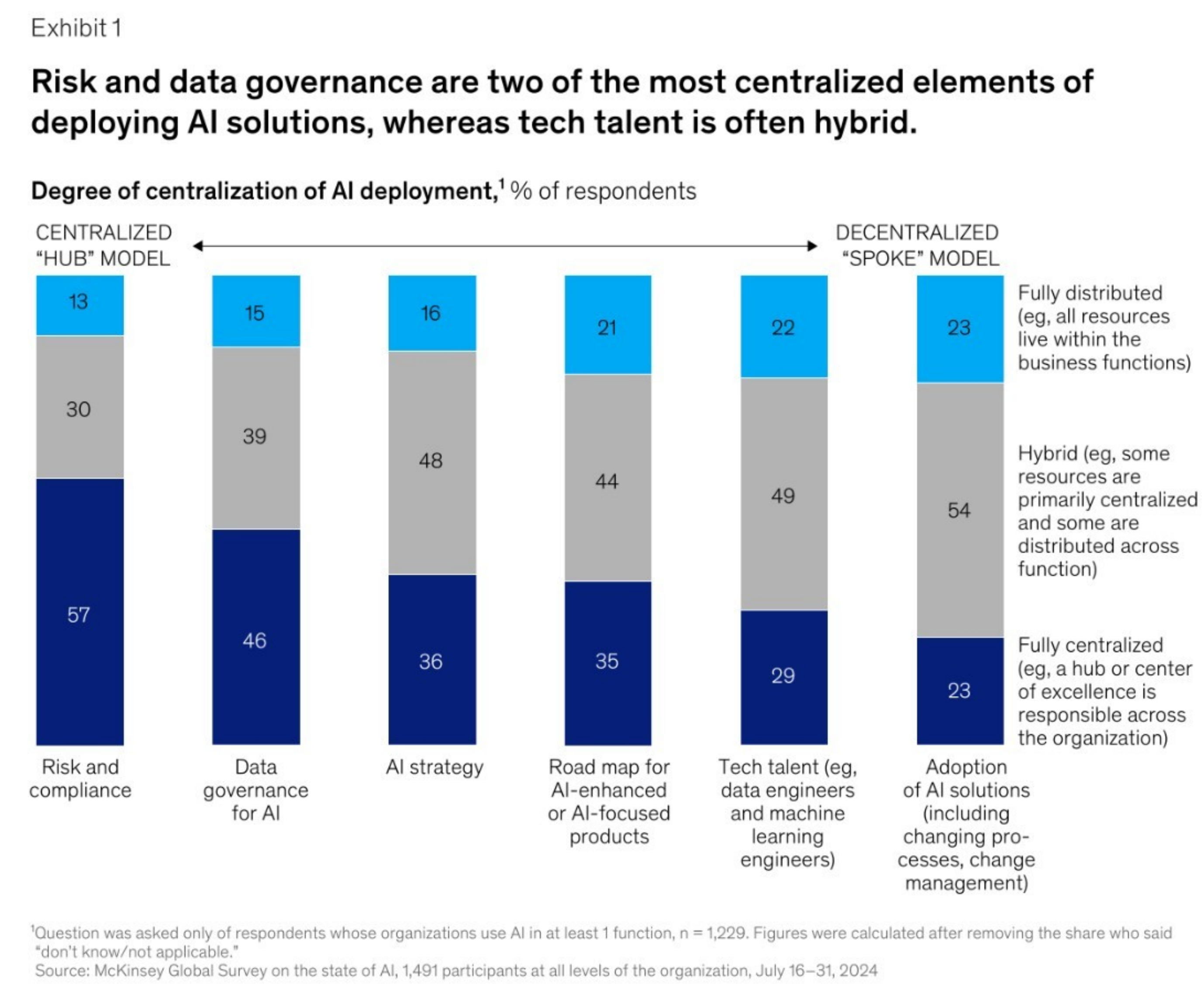
35

29

23

30

39



Extract text from PDF files

Side-by-side comparison: Test N°1

PyMuPDF

Exhibit 1

Degree of centralization of AI deployment,[1] % of respondents

McKinsey & Company

[1]Question was asked only of respondents whose organizations use AI in at least 1 function, n = 1,229. Figures were calculated after removing the share who said

“don’t know/not applicable.”

Source: McKinsey Global Survey on the state of AI, 1,491 participants at all levels of the organization, July 16–31, 2024

Risk and data governance are two of the most centralized elements of

deploying AI solutions, whereas tech talent is often hybrid.

57

46

36

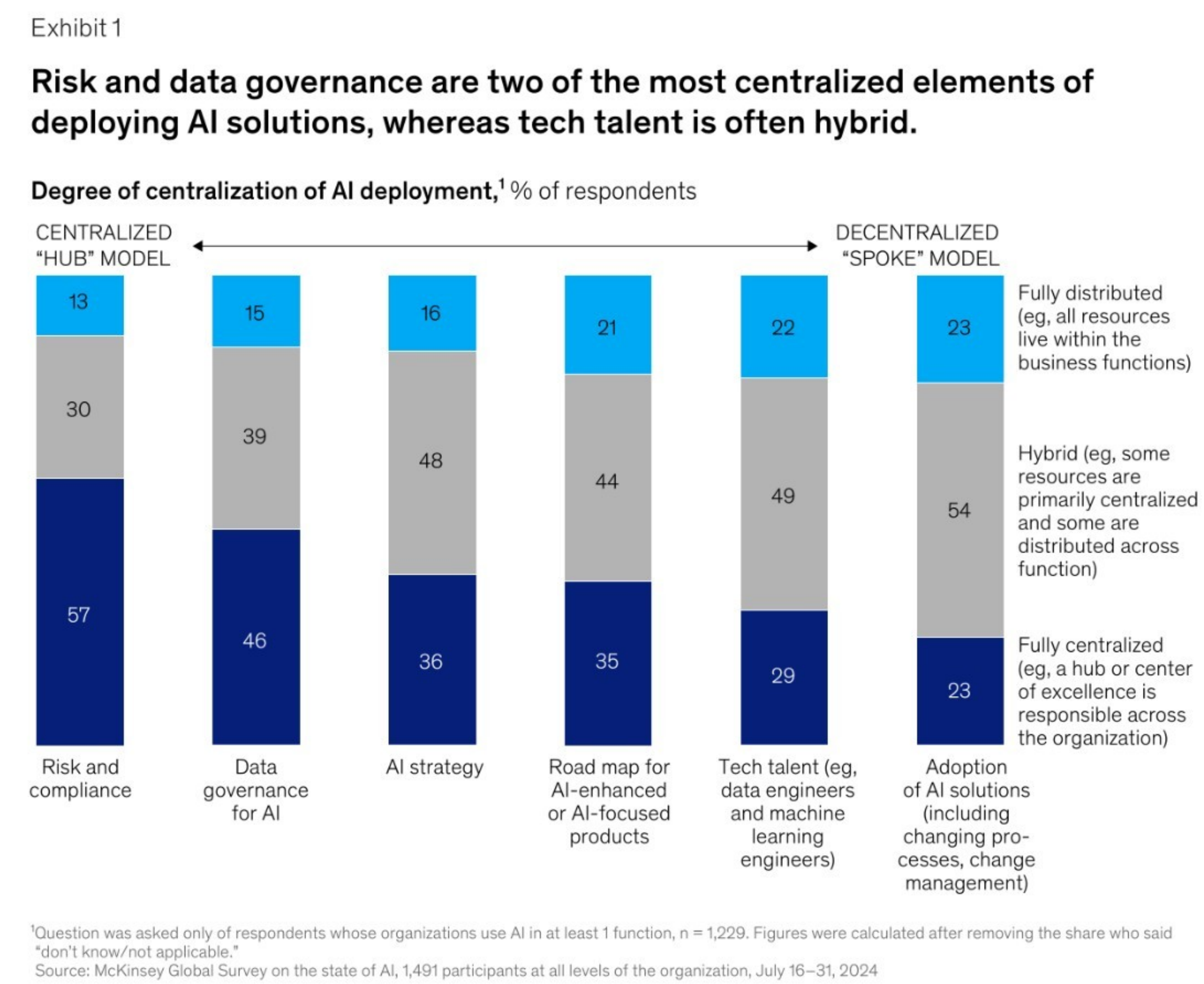
35

29

23

30

39



Extract text from PDF files

Side-by-side comparison: Test N°1

PyMuPDF

Exhibit 1

Degree of centralization of AI deployment,[1] % of respondents

McKinsey & Company

[1]Question was asked only of respondents whose organizations use AI in at least 1 function, n = 1,229. Figures were calculated after removing the share who said

“don’t know/not applicable.”

Source: McKinsey Global Survey on the state of AI, 1,491 participants at all levels of the organization, July 16–31, 2024

Risk and data governance are two of the most centralized elements of

deploying AI solutions, whereas tech talent is often hybrid.

57

46

36

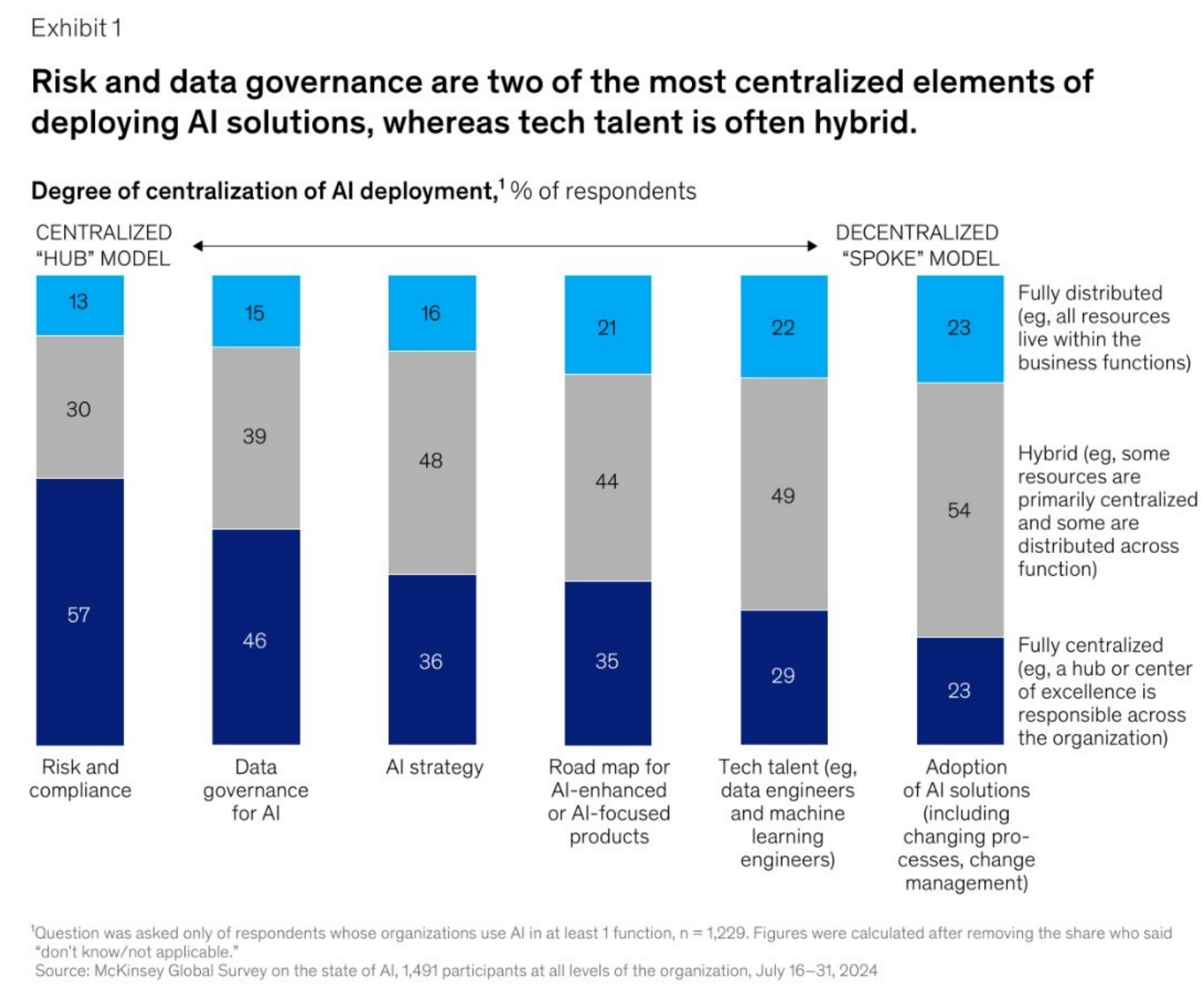
35

29

23

30

39



Extract text from PDF files

Side-by-side comparison: Test N°1

PyMuPDF (Continuation)

49

54

13

15

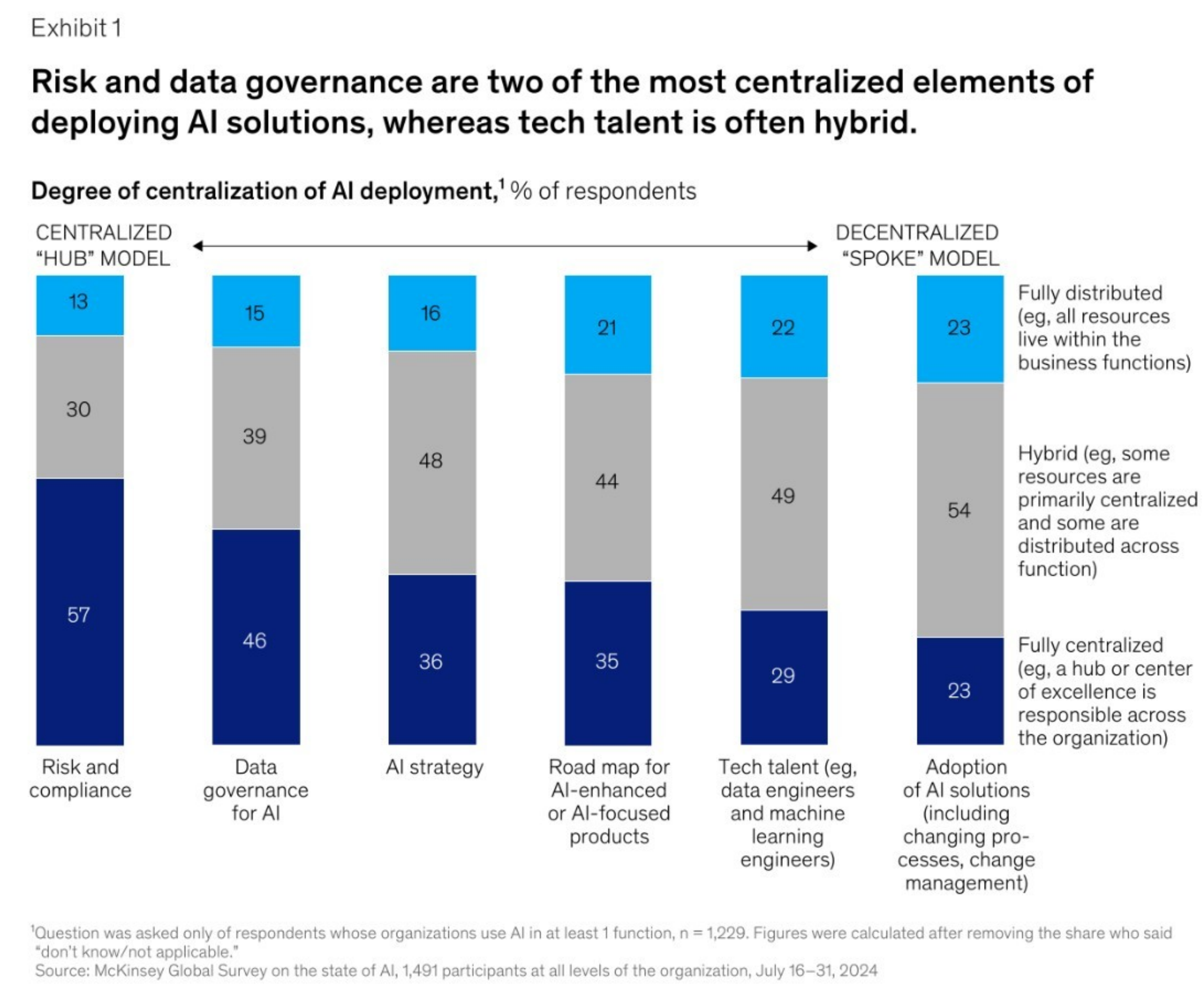
16

21

22

23

Fully centralized
(eg, a hub or center
of excellence is
responsible across
the organization)
Hybrid (eg, some
resources are
primarily centralized
and some are ...



Extract text from PDF files

Side-by-side comparison: Test N°1

VL model (Gemini 2.5 Pro)

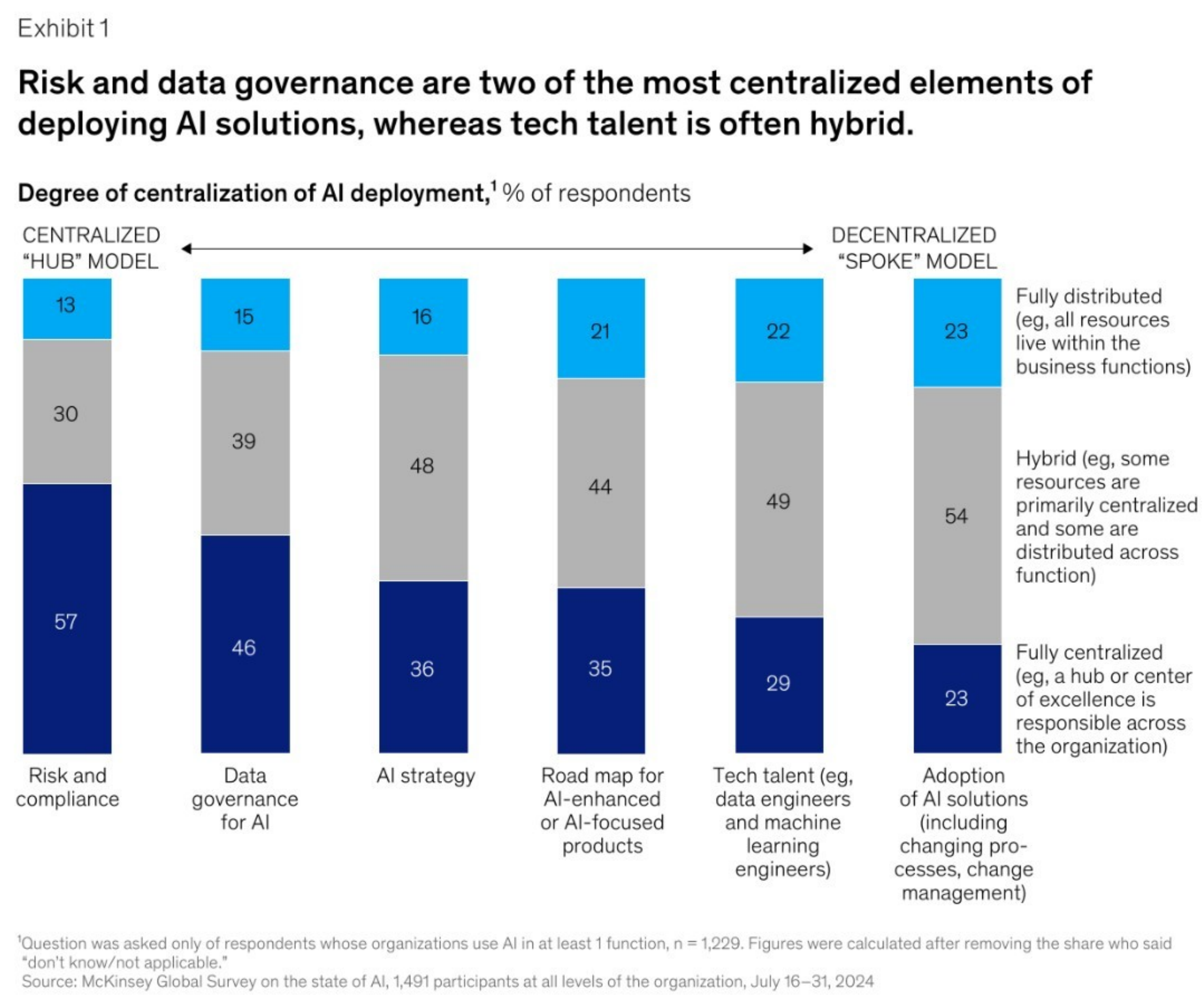
Exhibit 1

Risk and data governance are two of the most centralized elements of deploying AI solutions, whereas tech talent is often hybrid.

The chart shows two models of centralization: a "**CENTRALIZED 'HUB' MODEL**" on the left and a "**DECENTRALIZED 'SPOKE' MODEL**" on the right. The chart has six vertical stacked bars, each representing a different aspect of AI deployment. The legend on the right explains the color coding for the segments of each bar: a dark blue segment represents "**Fully centralized** (eg, a hub or center of excellence is responsible across the organization)", a gray segment represents "**Hybrid** (eg, some resources are primarily centralized and some are distributed across function)", and a light blue segment represents "**Fully distributed** (eg, all resources live within the business functions)".

The six bars are for the following categories, with their respective data breakdowns:

1. **Risk and compliance**: **57%** Fully centralized, **30%** Hybrid, **13%** Fully distributed.
2. **Data governance for AI**: **46%** Fully centralized, **39%** Hybrid, **15%** Fully distributed.
3. **AI strategy**: **36%** Fully centralized, **48%** Hybrid, **16%** Fully distributed.
4. **Road map for AI-enhanced or AI-focused products**: **35%** Fully centralized, **44%** Hybrid, **21%** Fully distributed.
5. **Tech talent** (eg, data engineers and machine learning engineers): **29%** Fully centralized, **49%** Hybrid, **22%** Fully distributed.
6. **Adoption of AI solutions** (including changing processes, change management): **23%** Fully centralized, **54%** Hybrid, **23%** Fully distributed.] ...



Extract text from PDF files

Side-by-side comparison: Test N°1

VL model (Gemini 2.5 Pro)

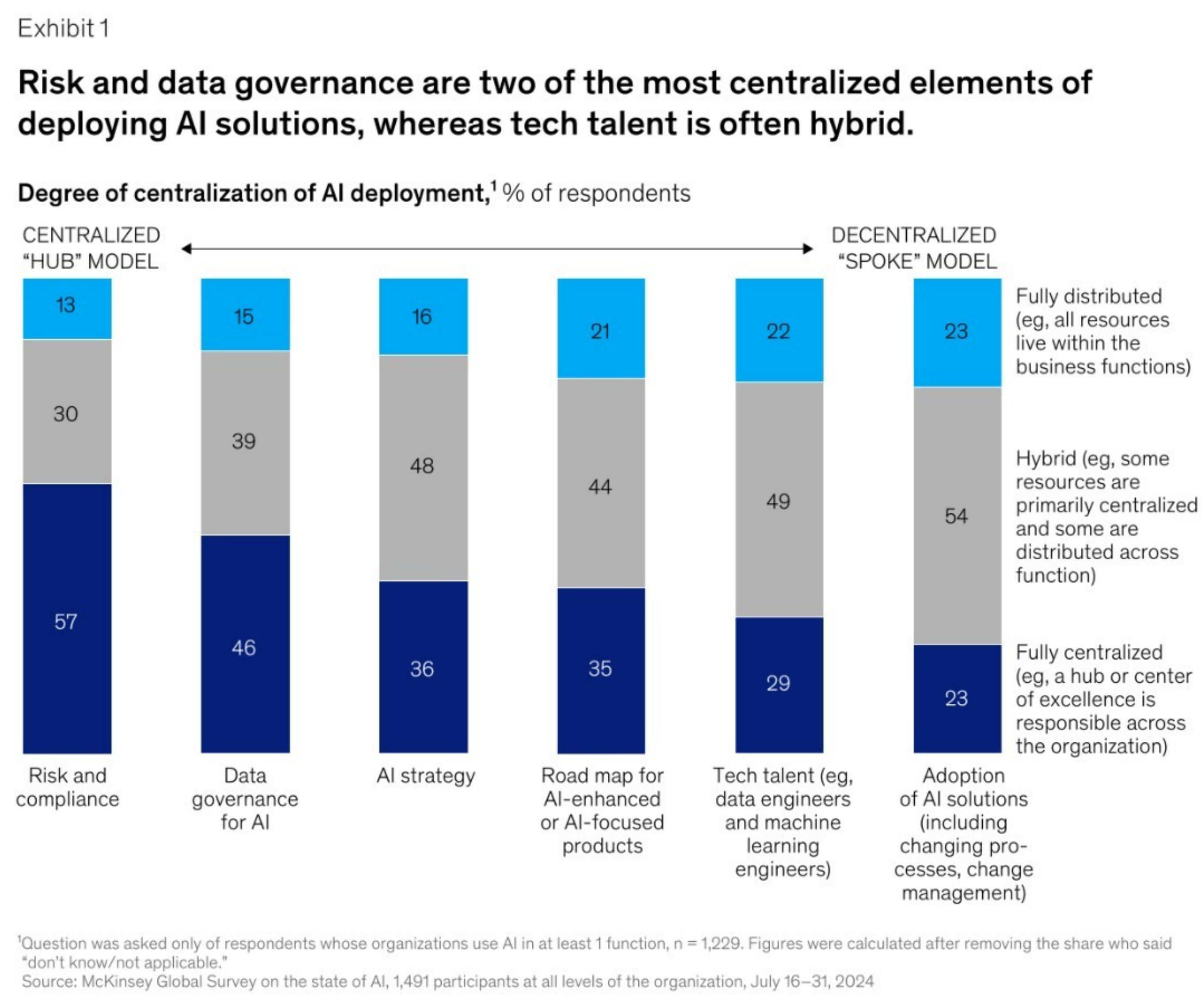
Exhibit 1

Risk and data governance are two of the most centralized elements of deploying AI solutions, whereas tech talent is often hybrid.

The chart shows two models of centralization: a "**CENTRALIZED 'HUB' MODEL**" on the left and a "**DECENTRALIZED 'SPOKE' MODEL**" on the right. The chart has six vertical stacked bars, each representing a different aspect of AI deployment. The legend on the right explains the color coding for the segments of each bar: a dark blue segment represents "**Fully centralized** (eg, a hub or center of excellence is responsible across the organization)", a gray segment represents "**Hybrid** (eg, some resources are primarily centralized and some are distributed across function)", and a light blue segment represents "**Fully distributed** (eg, all resources live within the business functions)".

The six bars are for the following categories, with their respective data breakdowns:

1. **Risk and compliance**: **57%** Fully centralized, **30%** Hybrid, **13%** Fully distributed.
2. **Data governance for AI**: **46%** Fully centralized, **39%** Hybrid, **15%** Fully distributed.
3. **AI strategy**: **36%** Fully centralized, **48%** Hybrid, **16%** Fully distributed.
4. **Road map for AI-enhanced or AI-focused products**: **35%** Fully centralized, **44%** Hybrid, **21%** Fully distributed.
5. **Tech talent** (eg, data engineers and machine learning engineers): **29%** Fully centralized, **49%** Hybrid, **22%** Fully distributed.
6. **Adoption of AI solutions** (including changing processes, change management): **23%** Fully centralized, **54%** Hybrid, **23%** Fully distributed.] ...



Extract text from PDF files

Side-by-side comparison: Test N°1

VL model (Gemini 2.5 Pro)

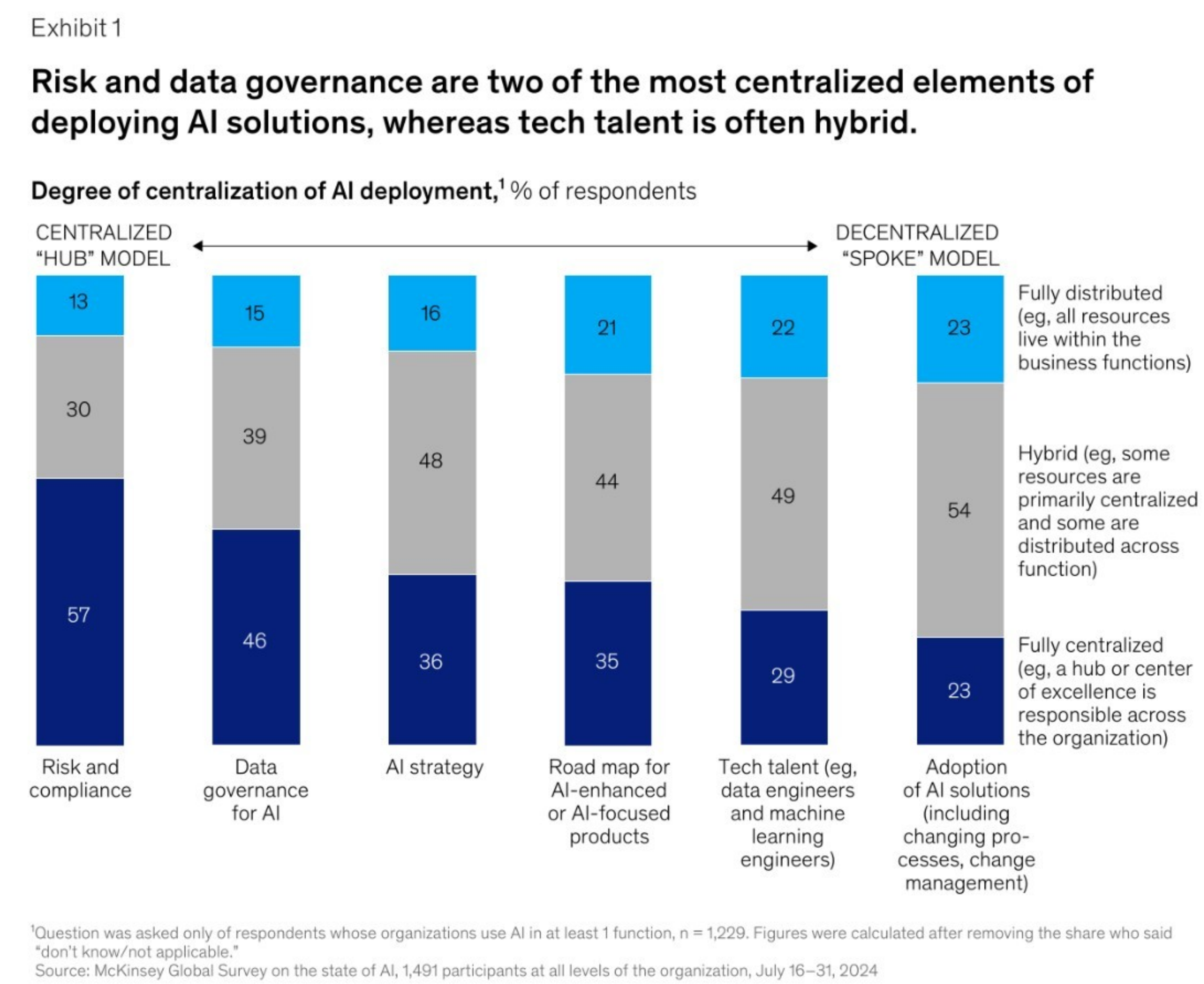
Exhibit 1

Risk and data governance are two of the most centralized elements of deploying AI solutions, whereas tech talent is often hybrid.

The chart shows two models of centralization: a "**CENTRALIZED 'HUB' MODEL**" on the left and a "**DECENTRALIZED 'SPOKE' MODEL**" on the right. The chart has six vertical stacked bars, each representing a different aspect of AI deployment. The legend on the right explains the color coding for the segments of each bar: a dark blue segment represents "**Fully centralized** (eg, a hub or center of excellence is responsible across the organization)", a gray segment represents "**Hybrid** (eg, some resources are primarily centralized and some are distributed across function)", and a light blue segment represents "**Fully distributed** (eg, all resources live within the business functions)".

The six bars are for the following categories, with their respective data breakdowns:

1. **Risk and compliance**: **57%** Fully centralized, **30%** Hybrid, **13%** Fully distributed.
2. **Data governance for AI**: **46%** Fully centralized, **39%** Hybrid, **15%** Fully distributed.
3. **AI strategy**: **36%** Fully centralized, **48%** Hybrid, **16%** Fully distributed.
4. **Road map for AI-enhanced or AI-focused products**: **35%** Fully centralized, **44%** Hybrid, **21%** Fully distributed.
5. **Tech talent** (eg, data engineers and machine learning engineers): **29%** Fully centralized, **49%** Hybrid, **22%** Fully distributed.
6. **Adoption of AI solutions** (including changing processes, change management): **23%** Fully centralized, **54%** Hybrid, **23%** Fully distributed.] ...



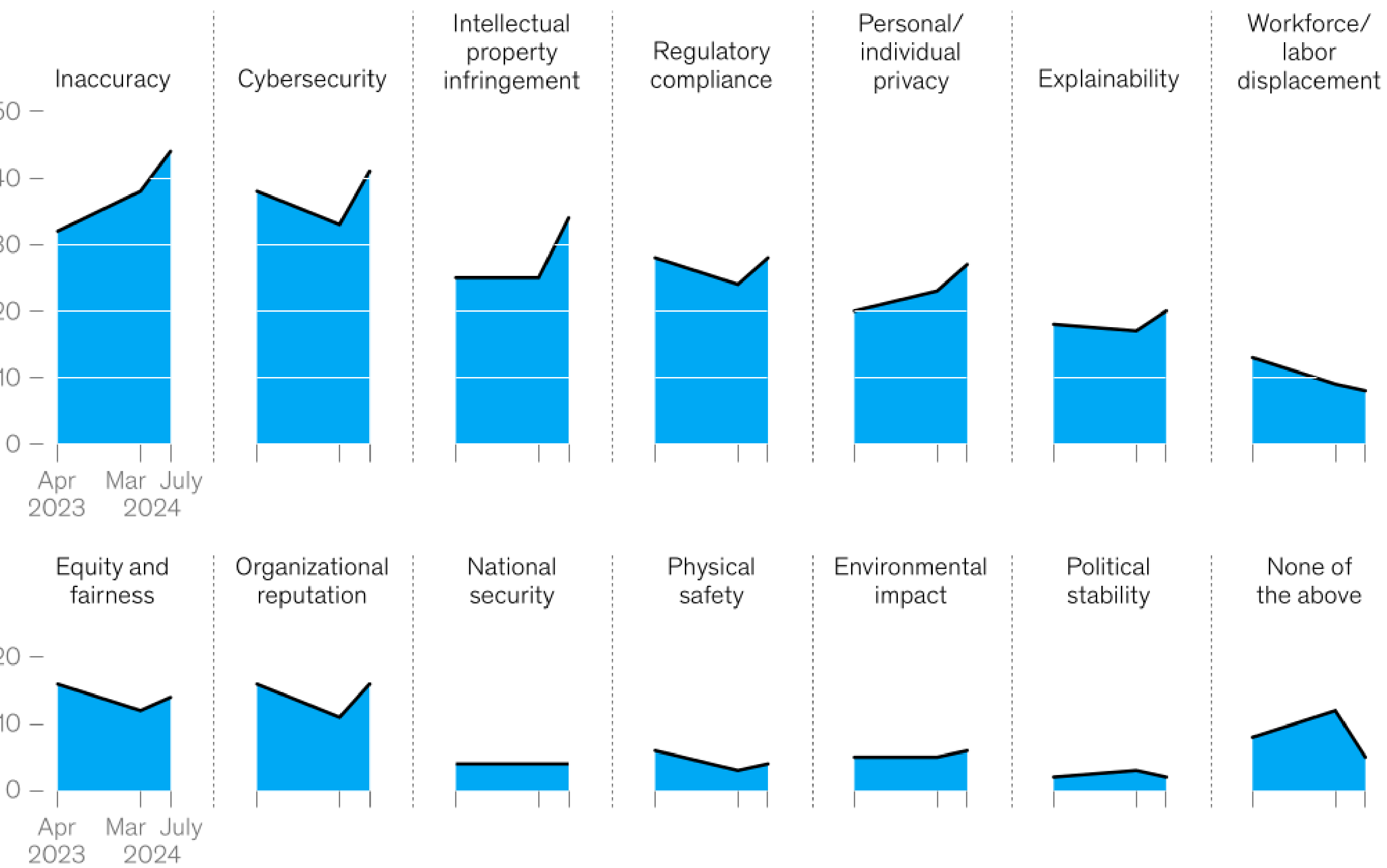
Extract text from PDF files

Side-by-side comparison: Test N°2

Exhibit 3

Respondents report increasing mitigation of inaccuracy, intellectual property infringement, and privacy risks related to use of gen AI.

Gen-AI-related risks that organizations are working to mitigate,[1] % of respondents



[1]Only asked of respondents whose organizations use AI in at least 1 business function. Respondents who said "don't know/not applicable" are not shown.
Source: McKinsey Global Surveys on the state of AI, 2023–24

Extract text from PDF files

Side-by-side comparison: Test N°2

PyMuPDF

Exhibit 3

Gen-AI-related risks that organizations are working to mitigate,1 % of respondents

1Only asked of respondents whose organizations use AI in at least 1 business function. Respondents who said “don’t know/not applicable” are not shown.

Source: McKinsey Global Surveys on the state of AI, 2023–24

Respondents report increasing mitigation of inaccuracy, intellectual

property infringement, and privacy risks related to use of gen AI.

McKinsey & Company

0

10

20

30

40

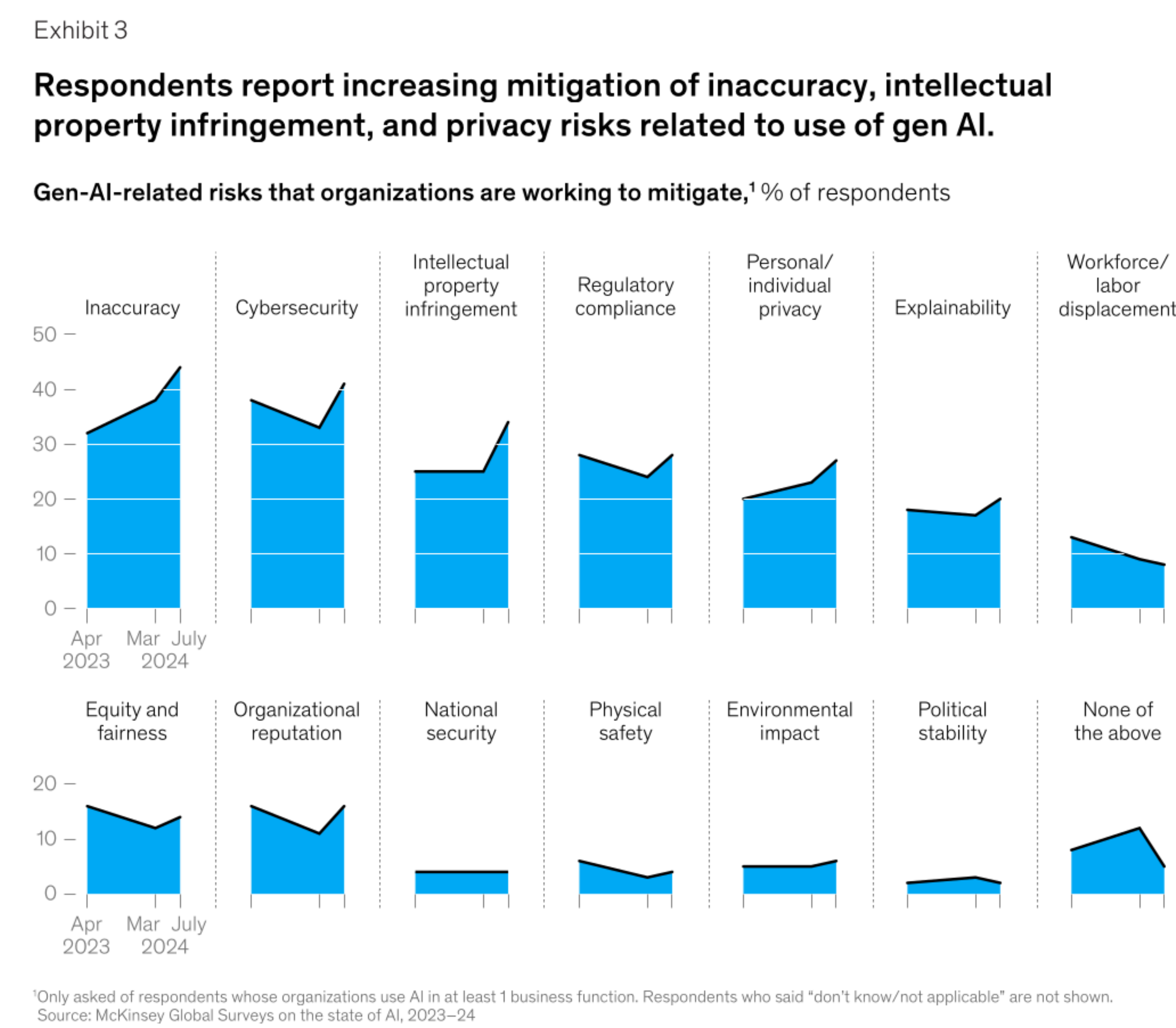
50

0

10

20

...



Extract text from PDF files

Side-by-side comparison: Test N°2

VL model (Gemini 2.5 Pro)

Exhibit 3

Respondents report increasing mitigation of inaccuracy, intellectual property infringement, and privacy risks related to use of gen AI.

Gen-AI-related risks that organizations are working to mitigate,¹ % of respondents

Image of a series of **14 area charts** arranged in **two rows of seven**. The charts show the percentage of respondents working to mitigate various generative AI-related risks over time, from **April 2023 to July 2024**.

The charts in the **top row** have a **y-axis ranging from 0 to 50**. The charts are:

* Inaccuracy: Shows a **significant increasing trend**, starting around **32%** in April 2023 and rising to approximately **45%** by July 2024.

* Cybersecurity: Shows a **slight overall decrease**, starting near **38%** in April 2023 and ending around **35%** in July 2024, with a dip in between.

* Intellectual property infringement: Shows a **clear increasing trend**, starting at about **25%** in April 2023 and rising to roughly **35%** in July 2024.

* Regulatory compliance: Shows a **slight downward trend**, starting at about **28%** in April 2023 and ending around **25%** in July 2024.

* Personal/individual privacy: Shows an **increasing trend**, starting from approximately **22%** in April 2023 and rising to about **28%** by July 2024.

* Explainability: Shows a **slight increase**, starting from about **18%** in April 2023 and ending around **20%** in July 2024 ...

Exhibit 3

Respondents report increasing mitigation of inaccuracy, intellectual property infringement, and privacy risks related to use of gen AI.

Gen-AI-related risks that organizations are working to mitigate,[1] % of respondents

[1]Only asked of respondents whose organizations use AI in at least 1 business function. Respondents who said "don't know/not applicable" are not shown.
Source: McKinsey Global Surveys on the state of AI, 2023–24

Extract text from PDF files

Side-by-side comparison: Test N°2

VL model (Gemini 2.5 Pro)

Exhibit 3

Respondents report increasing mitigation of inaccuracy, intellectual property infringement, and privacy risks related to use of gen AI.

Gen-AI-related risks that organizations are working to mitigate,¹ % of respondents

Image of a series of **14 area charts** arranged in **two rows of seven**. The charts show the percentage of respondents working to mitigate various generative AI-related risks over time, from **April 2023 to July 2024**.

The charts in the **top row** have a **y-axis ranging from 0 to 50**. The charts are:

* Inaccuracy: Shows a **significant increasing trend**, starting around **32%** in April 2023 and rising to approximately **45%** by July 2024.

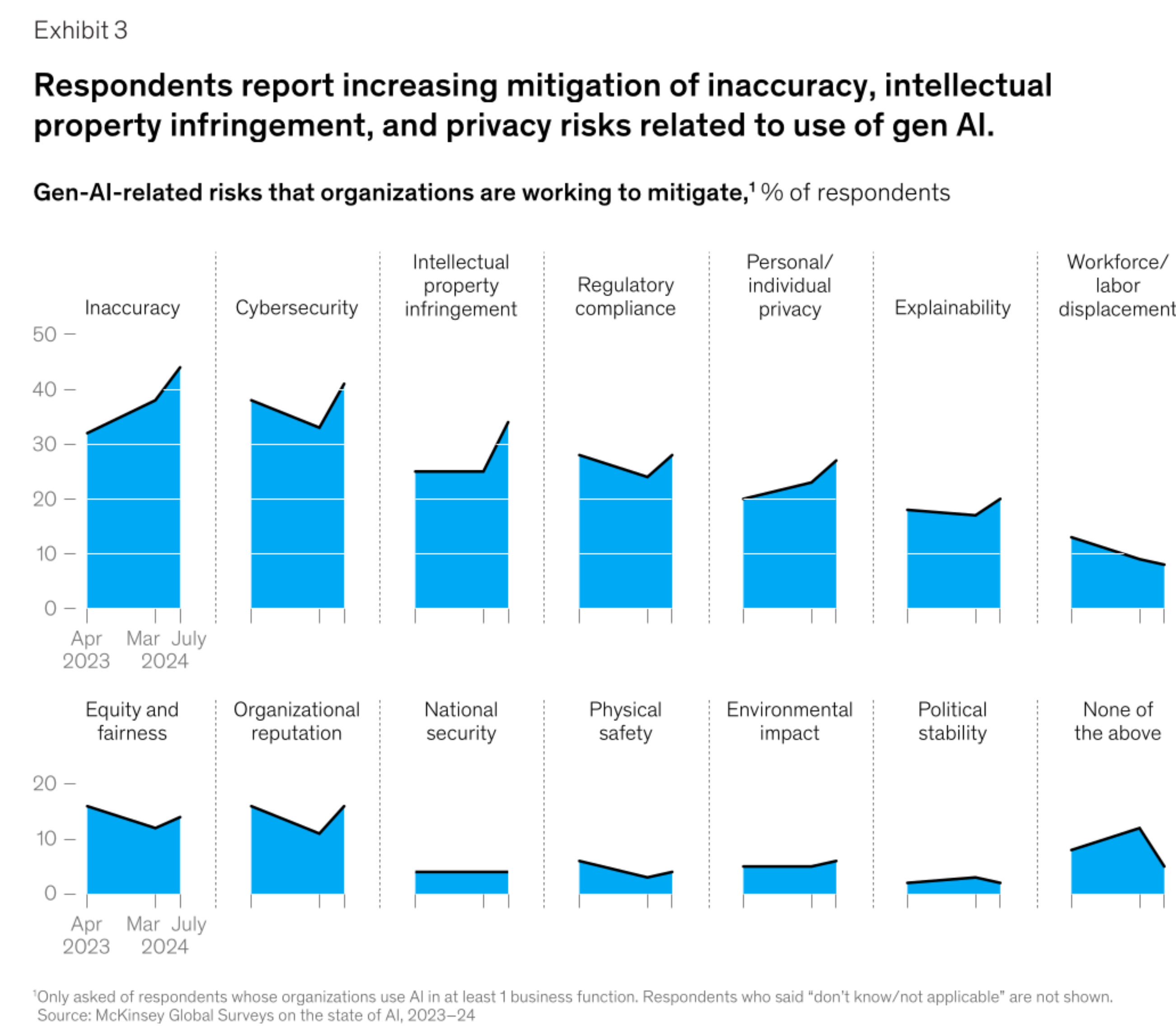
* Cybersecurity: Shows a **slight overall decrease**, starting near **38%** in April 2023 and ending around **35%** in July 2024, with a dip in between.

* Intellectual property infringement: Shows a **clear increasing trend**, starting at about **25%** in April 2023 and rising to roughly **35%** in July 2024.

* Regulatory compliance: Shows a **slight downward trend**, starting at about **28%** in April 2023 and ending around **25%** in July 2024.

* Personal/individual privacy: Shows an **increasing trend**, starting from approximately **22%** in April 2023 and rising to about **28%** by July 2024.

* Explainability: Shows a **slight increase**, starting from about **18%** in April 2023 and ending around **20%** in July 2024 ...



Extract text from PDF files

- VL models will help you create high quality datasets by:
  - **Annotating** images.
  - Describing tables or outputting them in **Markdown**.
  - **Preserve the layout**.
  - **Extract only what you need** from the documents.
  - Working with **scanned** documents.

Extract text from PDF files

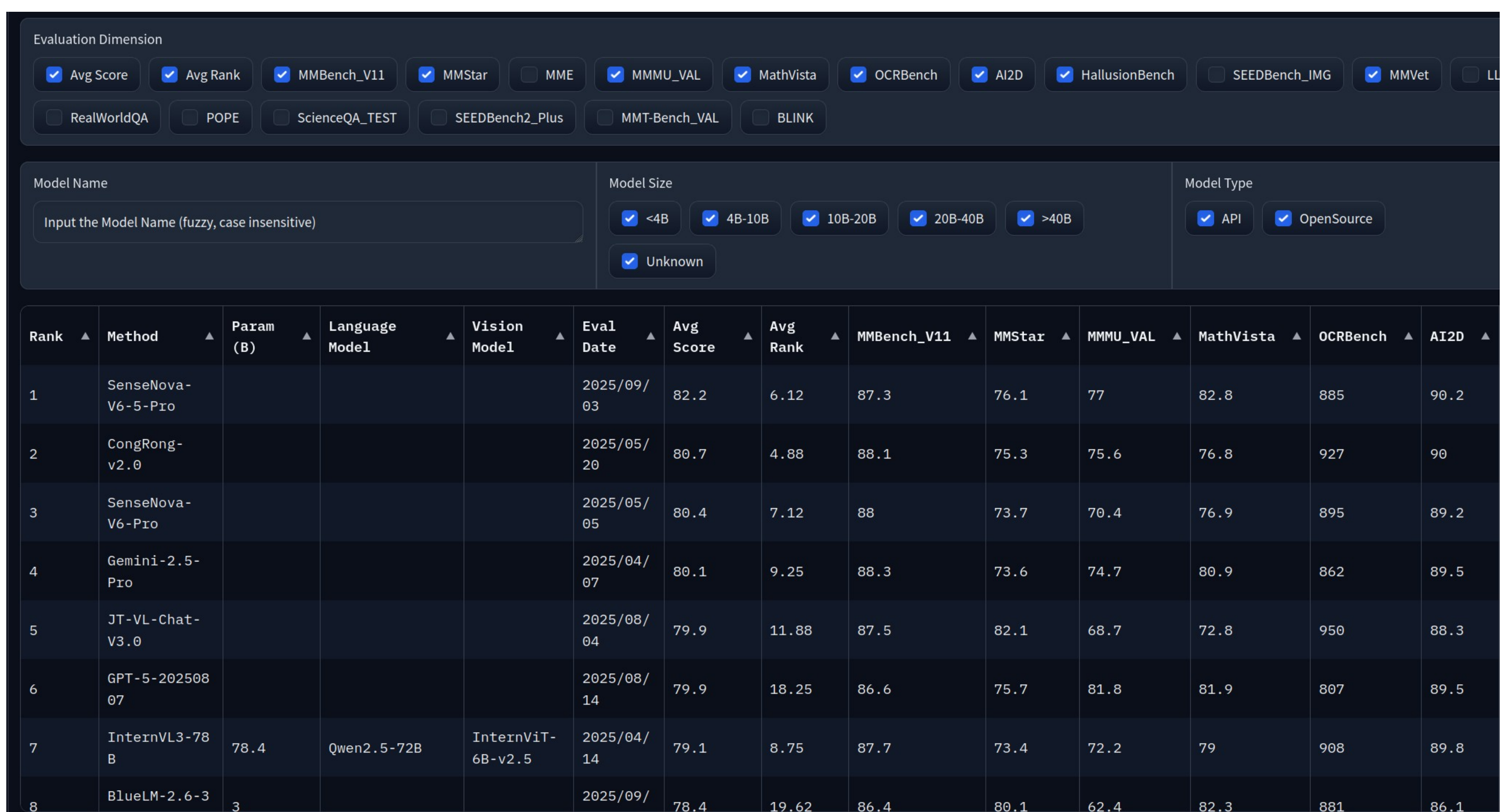
- What model should you use?
- Use proprietary and open models through the API. **Cheaper!**
- Host the models in your own servers. **Expensive!!!**

Proprietary models

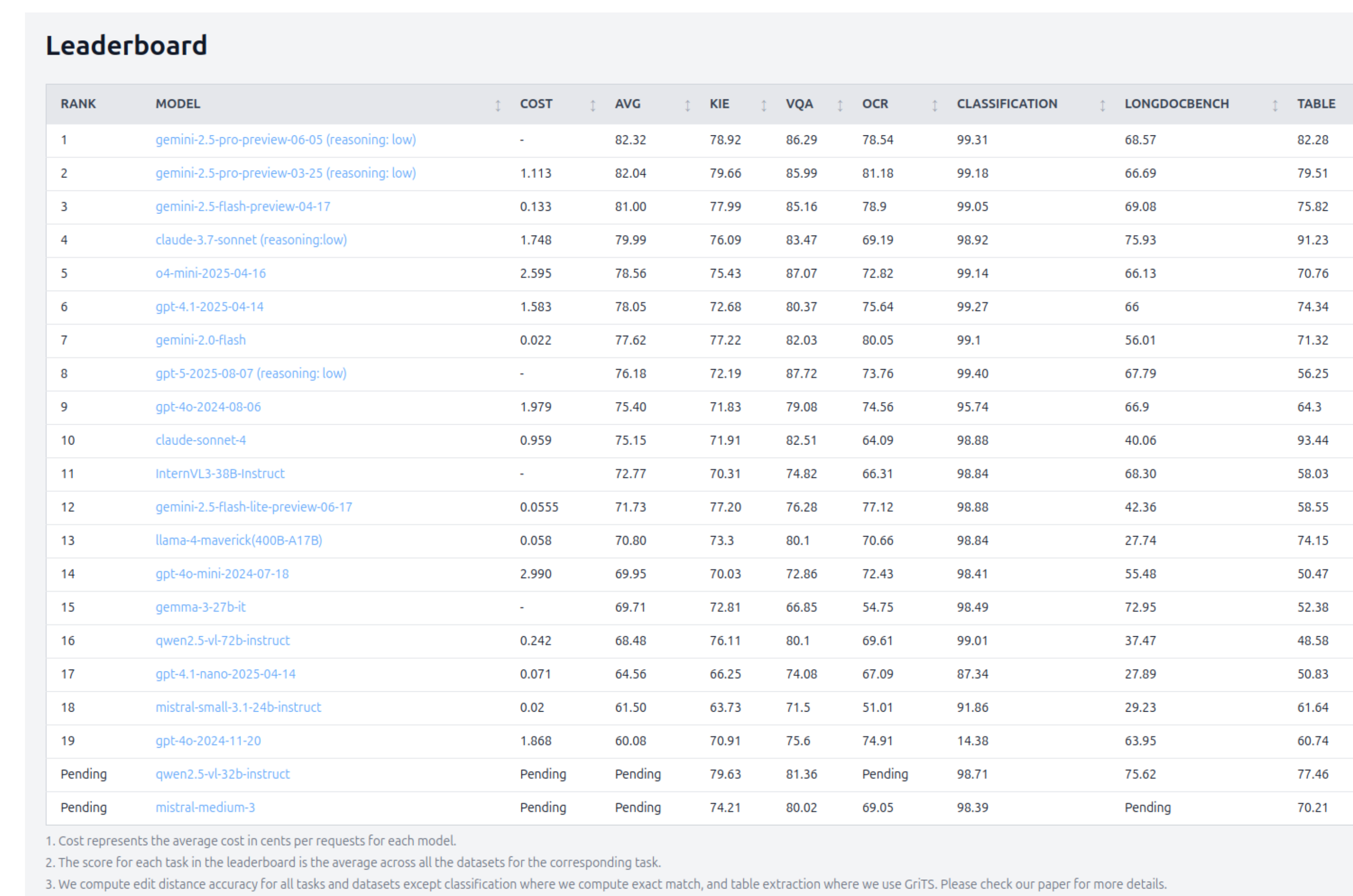
Open models

Extract text from PDF files

- Keep track of the best models with these leaderboards:
  - OpenVLM leaderboard.
  - Intelligent document processing leaderboard.



Rank	Method	Param (B)	Language Model	Vision Model	Eval Date	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OCRBench	AI2D
1	SenseNova-V6-5-Pro				2025/09/03	82.2	6.12	87.3	76.1	77	82.8	885	90.2
2	CongRong-v2.0				2025/05/20	80.7	4.88	88.1	75.3	75.6	76.8	927	90
3	SenseNova-V6-Pro				2025/05/05	80.4	7.12	88	73.7	70.4	76.9	895	89.2
4	Gemini-2.5-Pro				2025/04/07	80.1	9.25	88.3	73.6	74.7	80.9	862	89.5
5	JT-VL-Chat-V3.0				2025/08/04	79.9	11.88	87.5	82.1	68.7	72.8	950	88.3
6	GPT-5-20250807				2025/08/14	79.9	18.25	86.6	75.7	81.8	81.9	807	89.5
7	InternVL3-78B	78.4	Qwen2.5-72B	InternViT-6B-v2.5	2025/04/14	79.1	8.75	87.7	73.4	72.2	79	908	89.8
8	BlueLM-2.6-3	3			2025/09/	78.4	19.62	86.4	80.1	62.4	82.3	881	86.1


Leaderboard

RANK	MODEL	COST	AVG	KIE	VQA	OCR	CLASSIFICATION	LONGDOCBENCH	TABLE
1	gemini-2.5-pro-preview-06-05 (reasoning: low)	-	82.32	78.92	86.29	78.54	99.31	68.57	82.28
2	gemini-2.5-pro-preview-03-25 (reasoning: low)	1.113	82.04	79.66	85.99	81.18	99.18	66.69	79.51
3	gemini-2.5-flash-preview-04-17	0.133	81.00	77.99	85.16	78.9	99.05	69.08	75.82
4	claude-3.7-sonnet (reasoning:low)	1.748	79.99	76.09	83.47	69.19	98.92	75.93	91.23
5	o4-mini-2025-04-16	2.595	78.56	75.43	87.07	72.82	99.14	66.13	70.76
6	gpt-4.1-2025-04-14	1.583	78.05	72.68	80.37	75.64	99.27	66	74.34
7	gemini-2.0-flash	0.022	77.62	77.22	82.03	80.05	99.1	56.01	71.32
8	gpt-5-2025-08-07 (reasoning: low)	-	76.18	72.19	87.72	73.76	99.40	67.79	56.25
9	gpt-4o-2024-08-06	1.979	75.40	71.83	79.08	74.56	95.74	66.9	64.3
10	claude-sonnet-4	0.959	75.15	71.91	82.51	64.09	98.88	40.06	93.44
11	InternVL3-38B-Instruct	-	72.77	70.31	74.82	66.31	98.84	68.30	58.03
12	gemini-2.5-flash-lite-preview-06-17	0.0555	71.73	77.20	76.28	77.12	98.88	42.36	58.55
13	llama-4-maverick(400B-A17B)	0.058	70.80	73.3	80.1	70.66	98.84	27.74	74.15
14	gpt-4o-mini-2024-07-18	2.990	69.95	70.03	72.86	72.43	98.41	55.48	50.47
15	gemma-3-27b-it	-	69.71	72.81	66.85	54.75	98.49	72.95	52.38
16	qwen2.5-vl-72b-instruct	0.242	68.48	76.11	80.1	69.61	99.01	37.47	48.58
17	gpt-4.1-nano-2025-04-14	0.071	64.56	66.25	74.08	67.09	87.34	27.89	50.83
18	mistral-small-3.1-24b-instruct	0.02	61.50	63.73	71.5	51.01	91.86	29.23	61.64
19	gpt-4o-2024-11-20	1.868	60.08	70.91	75.6	74.91	14.38	63.95	60.74
Pending	qwen2.5-vl-32b-instruct	Pending	Pending	79.63	81.36	Pending	98.71	75.62	77.46
Pending	mistral-medium-3	Pending	Pending	74.21	80.02	69.05	98.39	Pending	70.21

1. Cost represents the average cost in cents per requests for each model.
2. The score for each task in the leaderboard is the average across all the datasets for the corresponding task.
3. We compute edit distance accuracy for all tasks and datasets except classification where we compute exact match, and table extraction where we use GriTS. Please check our paper for more details.