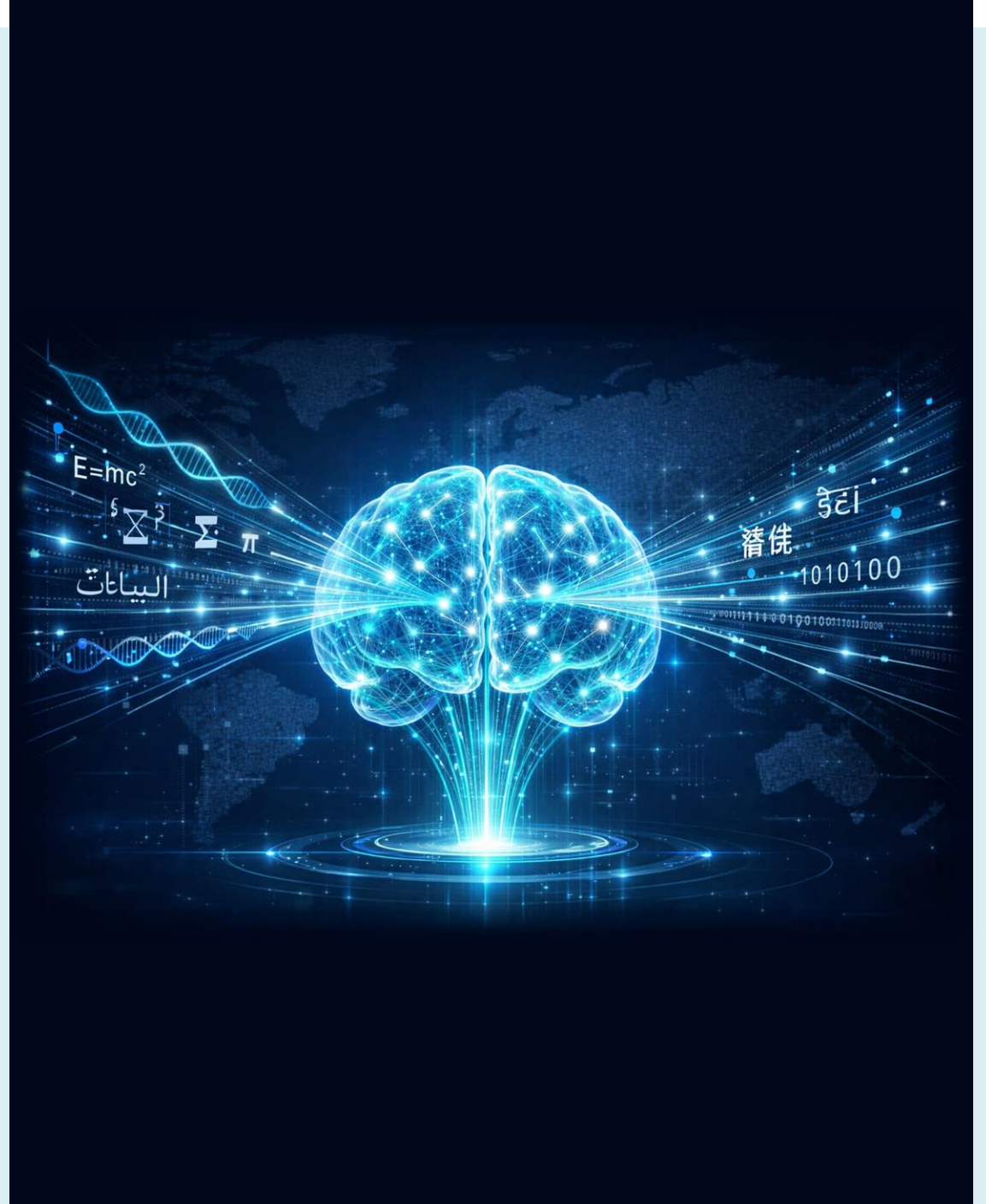


InfoBay.AI

The Synthetic Data Partner Behind Reliable, High-Performance LLMs

*Reducing hallucinations and boosting accuracy
for frontier AI models*



Top LLM builders already spend ~15% of total AI budgets on data; and this share is rising faster than compute spend



95% of executives rank data issues as top-3 AI risks



38% say domain-specific data is their #1 LLM bottleneck



2 in 3 AI builders don't trust their training data

TAM

\$17Bn spent on data & data services²

- Will increase to \$46Bn by 2030



SAM

\$7.6Bn spent on Data preparation market

- Will increase to \$16.2Bn by 2030
- Remaining on Data Annotation and Labelling; and validation



SOM

\$1.62Bn Looking to capture by 2030

- 10% of the current market



1. [Source](#): Data and AI development costs are substantial and growing. High-quality data preparation alone accounts for ~15–25% of total project spend, and most companies start without sufficient training data, increasing time and expense. Custom AI solutions range from ~\$20K to \$500K+ depending on complexity, with infrastructure and tooling adding another material cost layer. This cost intensity is a key driver for demand in synthetic and principled data solutions
2. [\(Mordor Intelligence\), \(Global Growth Insights\), \(IMARC Group\): The global data preparation market — a prerequisite for viable AI and analytics — was valued at ~USD 6.5 B by 2024 and is forecast to grow to ~USD 27.3 B by 2033 at a ~16.4% CAGR. Increasing data volume, rising demand for automated analytics, and the need for high-quality, governance-ready data are driving this expansion.](#)

Why Smarter Data Beats More Data in Advanced LLMs

InfoBay.AI

As LLMs advance, teams are prioritizing specialized and high-quality datasets that are curated, deduplicated, and domain-tailored because that **smarter data yields far better ROI than blind scale**.



Curation

(Signal > Noise)

Carefully curated data outperforms brute-force token scale by 40–50% on downstream tasks



Deduplication

(Eliminate False Confidence)

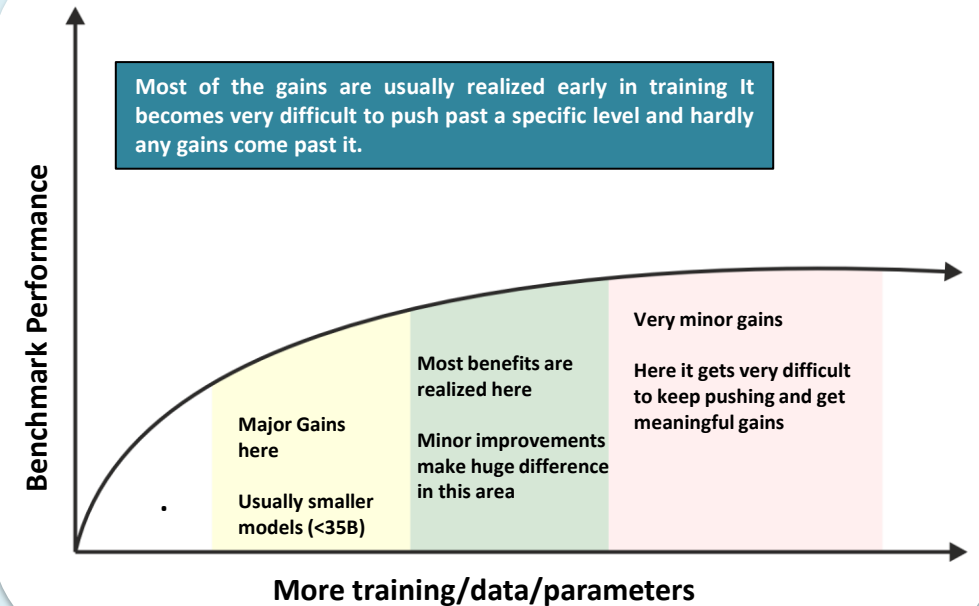
High duplication inflates training loss but reduces real-world accuracy by up to 40%



Domain Tailoring

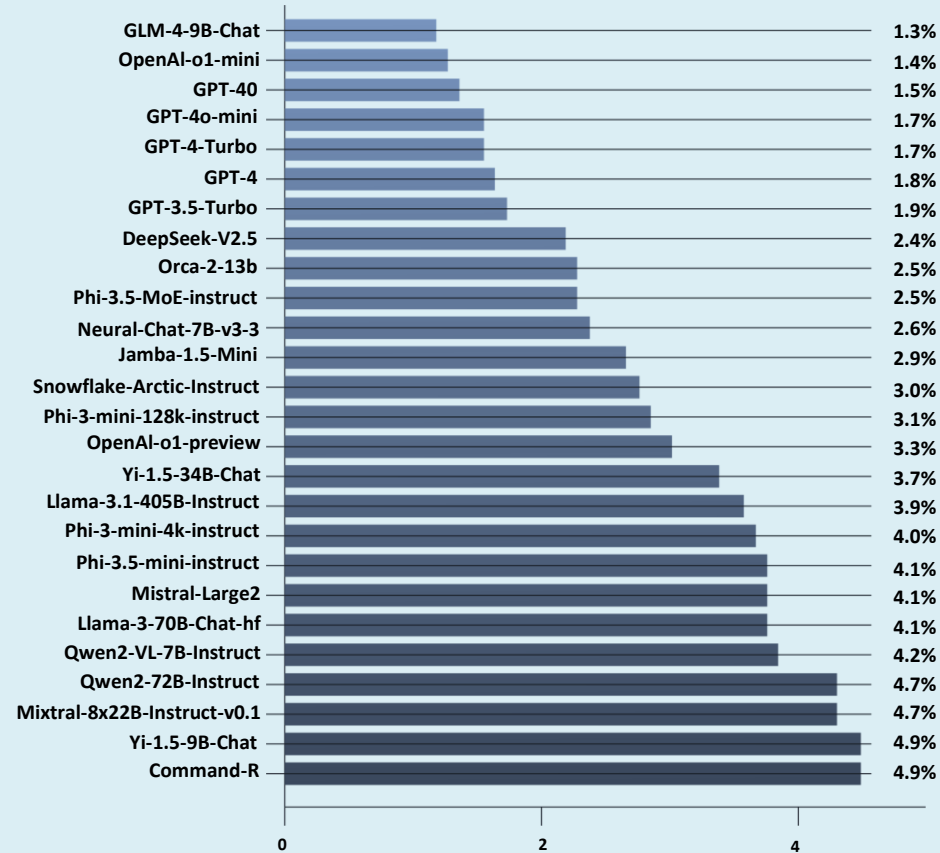
(Where ROI Lives)

70% of LLM practitioners now focus on domain-specific models, not general-purpose ones



Even the best-performing LLMs hallucinate 1–5% on general topics ; far higher in STEM, medical, and reasoning tasks

Rates increase materially as task complexity and reasoning depth increase



Hallucination rates in the Top 25 LLMs for general topics.

For Advanced topics such as STEM, Medical, Reasoning etc. these rates are much higher.

3 methods of solving LLM Hallucinations:

1



Scaling Laws

More data + more compute

Outcome: Marginal gains

Problem: Exponentially rising cost

2



Architectural Changes

Model redesigns

Outcome: Step-function improvements

Problem: Complex, slow, extremely expensive and no guaranteed output

3



Synthetic Data

(InfoBay's approach)

Targeted reasoning & failure-mode correction

Outcome: High-impact, surgical gains

Advantage: Fast, scalable, cost-efficient

Synthetic Data solves Hallucinations Efficiently



Parameter Drift (Data Variation)

Minor input variations silently alter internal reasoning paths



Over-Compression (Data Reduction)

Removing “unnecessary” context breaks latent reasoning chains



Pattern Overfitting (Template Modification)

Models anchor to familiar patterns and miss subtle changes



Multi-Step Collapse (Incremental Complexity)

LLMs fail when multiple simple operations must be composed



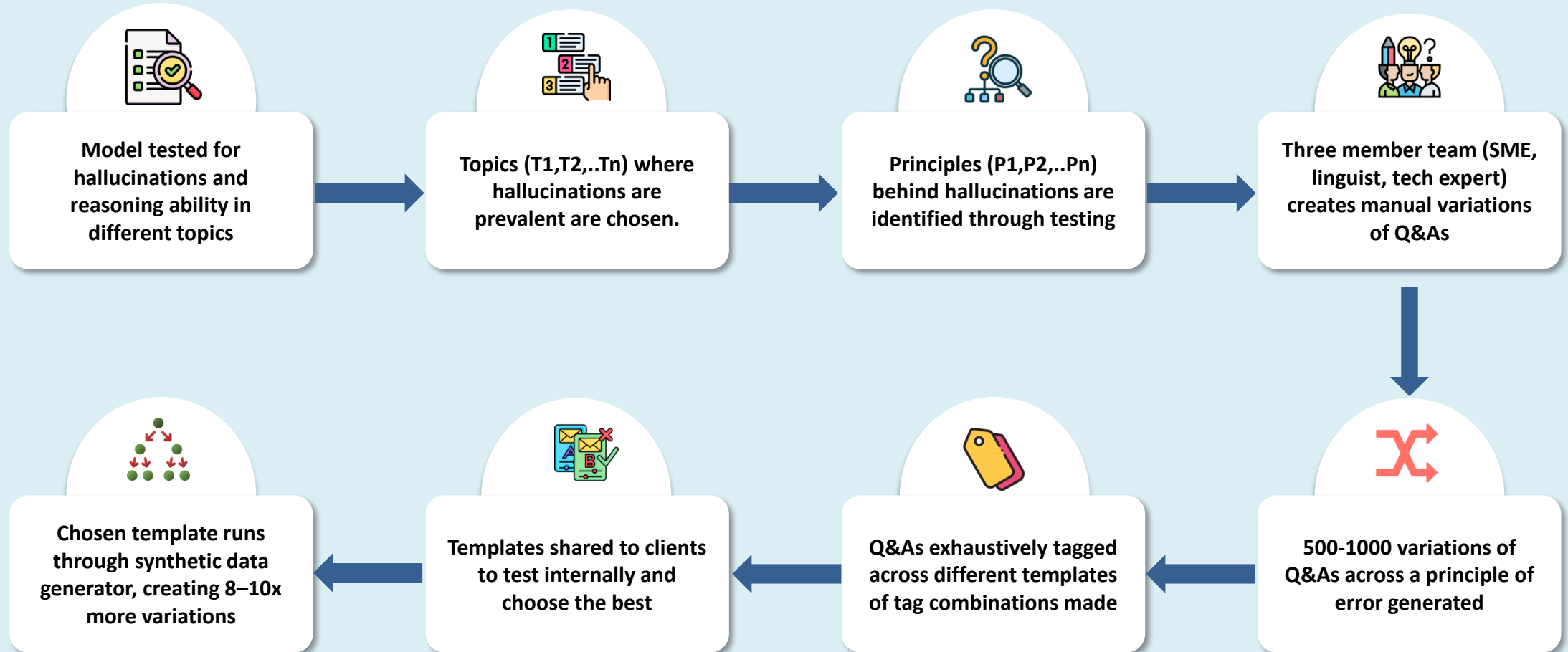
Implicit Constraint Blindness (Implicit Context Training)

Models optimize for explicit instructions but ignore unstated commonsense rules

and much more...

Inside InfoBay's Synthetic Data Generation Engine

InfoBay.AI



Exclusive Tier-1 Only. HQ-to-HQ. No Subsidiaries. No Affiliates.

PILLAR 1

Hyperscaler Only Access

We partner exclusively with companies valued at **\$100B+** market cap. This ensures data is built for frontier-grade models with the highest quality bar.

PILLAR 2

HQ-to-HQ Contracts Only

Microsoft: CEO Mustafa Suleyman | **Amazon:** CTO approved | **Google:** Alphabet-level signing

PILLAR 3 — THE DEAL PROCESS

- 1** **Discovery Call**
Sales teams align on scope & requirements
- 2** **Sample Delivery**
≤2% sample; AI scientists calibrate together
- 3** **Handshake Sign-off**
Scientists approve samples; R&D scopes full dataset
- 4** **Contract Signed**
Legal & commercial teams negotiate; deal finalized
- 5** **7–10 Day Delivery**
Full dataset delivered with 60-day review window
- 6** **Invoice on Day 61**
Clean invoice raised; 30–60 day payment terms

Global Tech Leaders

\$100B+
Client Market Cap (min)

All contracts executed directly with global headquarters

7 global technology leaders

HQ-level contracts (no subsidiaries or affiliates)

Multiple repeat orders



\$17 Mn+
Orders executed

\$37 Mn
In sales
pipeline

80%+
Gross Margin

65%
PBT

Near Term Pipeline



ANTHROPIC



HONDA

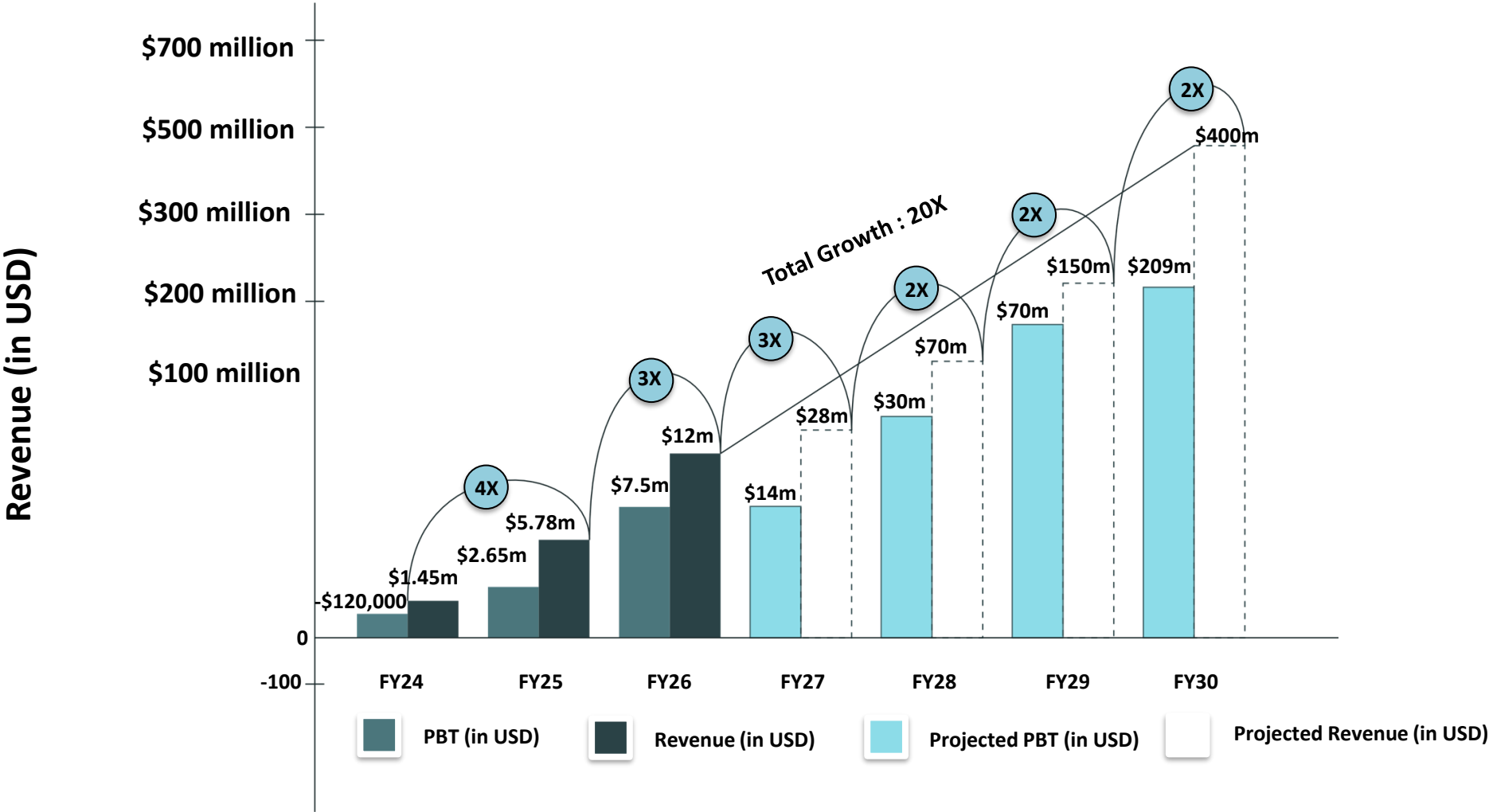


and much more

1 contract per month | Pipeline of \$37M | 15 orders in pipeline

65% PBT Business, 3x annual growth

InfoBay.AI





WHAT WE SERVE TODAY

Education

7.2M Q&A pairs
2.5B training tokens
15 Languages (English, Hindi, Arabic, Bahasa, Javanese, Nepali, Urdu, 8 Indics)

Conversational Audio

2.1M hours audio
57K podcast hours
36 Languages
13 countries

Healthcare

2M Patient data
Oncology,
Tuberculosis,
Infertility,
Dermatology, HIV

Video

4.5M vertical videos



Healthcare Data — Our depth advantage

47

Hospital Networks
India, Indonesia
& Philippines

15

Super specialist Doctors



COMING NEXT



PHYSICAL AI

Robots need data about the physical world.
Sensor fusion · Spatial reasoning
Environment perception
The \$38B physical AI market is exploding; and it needs the same precision training data that has made InfoBay a trusted partner.



DEFENCE AI

The Russia Ukraine conflict is the first war where AI makes split-second decisions: **drone targeting, threat identification, ISR.**
Every military AI model is only as good as its training data.
InfoBay's domain-tailored synthetic datasets are purpose-built to fill that gap: fast, accurate, and at scale.

Our 5 Moats: Why InfoBay Is Structurally Hard to Replicate

InfoBay.AI



1. Patent

Built **Tibet AI engine** in-house for scalable synthetic data generation using proprietary reasoning-failure identification and correction workflows



2. Future Proof

Proactively building deep datasets across **medical, voice, multilingual, and advanced reasoning** domains ahead of market demand



3. Proven scale

Delivered large-scale, high-quality datasets across **multiple modalities and languages** without linear increase in human effort



4. Low cost

Synthetic-first delivery model enables **structurally lower marginal costs** and **80%+ gross margins at scale**

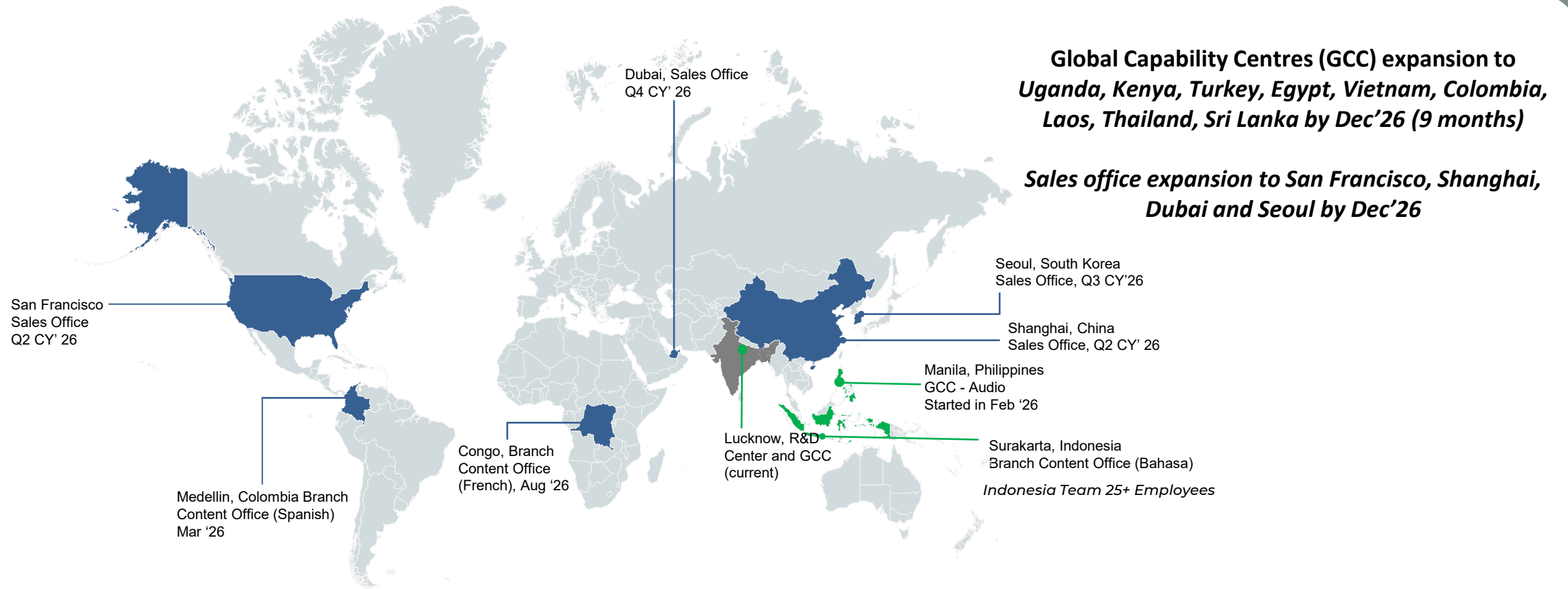


5. Trust & credibility

Trusted by **7 global technology leaders**, with repeat orders and contracts executed directly at **HQ level**

Operations in 5 countries, Expanding 1 country per month

InfoBay.AI



Will scale to 15 countries by Mar'27

Different Approaches to LLM Data; And Where InfoBay Wins

InfoBay.AI

Company	Business Model	Product Roadmap Shift	Funding raised	Type of Data	STEM / Reasoning Focused Smart Data	Hybrid Model	MMLU/Reasoning Boost	Headcount	Data use phase
InfoBay.AI	Depth and diversity of raw data with deep R&D capabilities	Focused on expanding its synthetic data engine 'Tibet' to cover Physical AI	\$3 Mn (In legacy biz.)	Specialized (STEM, Videos, Audio, Medical)	Yes	Yes (SME + synthetic)	Yes – explicitly boosts MMLU Scores	400	Across Pre-training, SFT & RLHF phase
	Premium AI training data platform for advanced LLM alignment, delivering expert-led RLHF, SFT, evaluations, red-teaming, and benchmarking.	Evolved from basic data annotation into a full-stack LLM alignment and evaluation platform, spanning RLHF, multimodal tasks, adversarial testing, and 70+ languages.	Bootstrapped; ~\$1–1.2 Bn revenue (2024); raising ~\$1 Bn (2025)	Specialized, expert-generated reasoning & alignment data	Yes	Yes (expert humans + AI tooling)	Yes (reasoning & alignment benchmarks)	~110 employees; ~1M expert contractors	SFT, RLHF, alignment & evaluation
	Transitioned from labelling to full-stack AI infrastructure, including evaluation and tooling.	Expanding into defense and government sectors with projects like Thunderforge; launched 'Scale Evaluation' platform for LLM benchmarking	\$2.6 Bn (Last round: Series F, \$1 Bn, May 2024; Meta stake \$14.3 Bn, Jun 2025)	General & specialized	No	Yes (human + ML)	NA	1,500 employees 100k+ contractors	Pre-training, Fine-tuning
	Provides AI training data and tech services, leveraging a vast network of human experts for model fine-tuning and development.	Launched LLM Lab and ALAN tools to accelerate fine-tuning and evaluation; expanded into enterprise AI consulting with a 30-day AI transformation accelerator.	\$247 Mn (Last Round: Series E, \$111 M, Mar 2025)	General & specialized	Yes	Yes(Human + synthetic)	Claims unspecified boost in reasoning ability	667 employees 4M+ contractors	Pre-training, Fine-tuning
	Provides ethical labelling services with focus on GenAI and vision tasks	Repositioning services to cater to GenAI applications, emphasizing ethical data sourcing and annotation	\$485.6 Mn (Last round: Series C, \$350 M, Oct 2025)	General & specialized	Yes	Yes (human + synthetic)	NA	30 employees 30k+ experts	Pre-training, Fine-tuning

Our Team

InfoBay.AI



Rohit Manglik
Founder & CEO

- B.Tech, NIT Surathkal | Stanford Seed Program
- 16+ years across AI, data systems, and scalable platforms
- Ex-Oracle, Morgan Stanley, D. E. Shaw | 6 patents



Shashwat Vikram
Co-founder & CTO

- B.Tech, IEM Kolkata | 6 patents
- 10+ years in full-stack systems and AI infrastructure patents
- Leading product and synthetic data engine development



Sanjeev Pandey
Head Of Content

- BCA, SRMU Lucknow
- 9+ years in large-scale, multilingual datasets
- Leads content operations and quality assurance



Ankit Singh
Head of Global Data Procurement

- BBA (Professional), BBDU Lucknow
- 7+ years in scalable data supply networks
- Leads data sourcing across regions and languages



Yashveer
Head of Delivery

- Ex Scale and Ex Turing
- 25+ years in Project Management and Data
- Leads data sourcing across all clients



Ashish Goel
Chief Financial Officer

- CA, CS, CMA -
- Ex Indux Intex, Sabyasachi, Datahash
- Raised upwards of 500cr from PEs
- 16+ years of Experience



India



Indonesia



Kenya



Philippines



Uganda

THANK YOU

