Yujia Bao Ph.D.

bao@yujia.io | (617)-386-9729 | LinkedIn | Google Scholar | GitHub | US Permanent Resident

Summary

I manage 80 research scientists and engineers at Accenture's Center for Advanced AI in Mountain View. I initiated and lead the engineering and research development of AI Refinery, Accenture's first AI Agent Platform for enterprise. Over the past year, I grew the team from 3 to 80, coached 9 tech leads, created hiring pipelines, formalized engineering standards, and promoted research innovations with 7 patents pending. Since its Sep 2024 internal launch, AI Refinery has driven hundreds of millions in sales over 30 Fortune 500 clients, supporting 3,000+ projects and adoption by 10,000+ developers worldwide (details available upon request).

In addition to technical leadership, I defined the product roadmap and shaped the go-to-market strategy for AI Refinery. I engaged directly with client C-suite and VP stakeholders to remove blockers, and our collaboration was featured by Jensen Huang at NVIDIA GTC 2025.

Education

Massachusetts Institute of Technology, Ph.D. in Computer Science 2022

Massachusetts Institute of Technology, M.S. in Computer Science 2019

University of Wisconsin-Madison, M.A. in Mathematics 2017

Shanghai Jiao Tong University, B.S. in Mathematics 2016

Experience

Accenture - Center for Advanced AI Mountain View, CA

AI Refinery Platform Lead Sep 2024 – present

Pioneered Accenture's AI Refinery platform, lead day-to-day engineering and research execution, managed a team of 80 engineers and researchers, and drove enterprise sales and adoption. Over past 12 months, personally contributed over 150k lines of production code and merged over 900 pull requests.

- **From zero to one:** In Aug 2024, I recognized a market opportunity for an agent-as-a service product. At the time, Accenture's global teams built agents in silos with various frameworks, making re-use and central governance impossible. I prototyped the first AI Refinery platform SDK, which was used to deliver a PoC for a Fortune 500 client within two weeks. After securing the deal, Accenture executives elevated AI Refinery platform to a company-wide priority [Accenture Newsroom].
- **Team building:** Built four functional teams—agent, training, infra, testing-at the beginning, and later on expanded to nine by adding security, data, UI, experience, and support. Established engineering standards across the platform, including code maturity levels; design and PR review policies; and a release cycle that spans development, staging, pre-prod, and prod.
- **Hiring pipeline:** Designed interview processes and talent pipelines for research scientists and engineers (directly hired ∼40 external candidates). Ran recruiting events at NeurIPS 24, ICLR 25 and ICML 25 (80-120 candidates attended each event), and organized campus-recruitment drives at leading engineering schools (Columbia, Stanford).
- **Developer community:** Organized 100+ workshops for internal developers to engage with AI Refinery SDK since Sep 2024, reaching over ten thousand active users to date. Hosted developer competition (Sep 2025), with participation from 76 teams worldwide.
- **Design principles:** Established core design philosophy for AI Refinery platform.
  - **Ease of use.** Heavy-lifting infrastructure resides on service side, enabling an intuitive client-side API, with agent configuration through YAML schema.
  - **Extensibility.** Developers have access to pre-built agents (deep research, data analytics), *and* the ability to fully customize and define new agents/tools for different industry solutions [Accenture Newsroom].

- **Enterprise-ready features.** Customers have access to PII masking, global/local RAI governance, and built-in integration with third-party agents from providers such as Amazon Bedrock, Azure AI Foundry, and Databricks [Accenture Newsroom].
- **Stability.** Achieved 99.99% availability for the platform over the past year.
- **Multi-cloud agnosticism.** Low-level infrastructure (databases, storage, etc.) is abstracted from agent design, making AI Refinery applicable to customers from different cloud providers or on-premises environments. [Accenture Newsroom, Azure Marketplace, AWS Marketplace]

Uber Tech Lead, Multi-Agent Application Jun 2024 – Aug 2024

Developed multi-agent AI system to support Accenture's internal marketing and campaign processes, which later became the foundation of AI Refinery. Product was delivered under a three-month deadline, with a PoC demonstration to Accenture board members and Jensen Huang at 1.5 months. Since launch (Aug 2024), the solution has served the team daily, reducing manual steps for managing a marketing campaign by 55% [Accenture Case Study]. Managed a team of 28 engineers and researchers, and led four technical leads.

- **Cross-functional collaboration:** Identified design objectives through week-long workshop with marketing experts and defined execution roadmap. Mapped common tasks to AI agents, including information collection and data analysis.
- **Data foundation:** Leveraged over 50 data sources, including: digital assets (PDFs, PPTs, videos), SQL databases, and static storage (S3, Blob). Built fully-automated data cleaning pipelines, including de-duplication, quality filtering, and conflict resolution.
- **Agent framework – Distiller:** Developed multi-agent framework composed of three agent types:
  - **Orchestrator** – routing the user query to the underlying agents; providing safety guardrail;
  - **Utility agent** – completes a single standalone task (e.g., RAG, image generation, report authoring, etc.).
  - **Super agent** – agent managers that decomposes complex tasks into standalone tasks that can be handled by the utility agents. Support both dynamic planning and pre-defined workflows.

  Introduced a shared memory layer for standard chat history, agent-specific storage, cross-agent memory, and structured memories (ICLR 2025). Distiller became the foundation of the AI Refinery platform [Accenture Newsroom].
- **Agent development:** Supervised the development of 14 marketing-focused agents [AWS Marketplace], including:
  - **Research agent (utility agent):** Productionizes agentic RAG with multiple data sources. Innovations: inference-time reasoning for complex logical queries (ACL 2025) and source-specific query transformation for multiple vector databases.
  - **Analytics agent (utility agent):** Selects relevant databases; generates and executes SQL; aggregates results; and produces final answers or plots.
  - **Strategy advisor (super agent):** Generates a marketing planning brief that spans research, analytics, calendar, writing, and reasoning agents. Ensures no timeline conflicts and information coverage.
- **Iterative refinement:** Integrated UI elements for user suggestions, and organized biweekly workshops to discuss performance and limitations before product launch. Improved agent prompt coverage, security and responsible AI layers over time after the release.

Tech Lead, LLM Customization Jun 2023 – May 2024

Led Accenture's first client project on LLM customization—Fortune Analytics [Accenture Case Study]. Product was delivered in May 2024 and won the iF Design Award. Managed a team of 14 researchers and engineers. Pretrained 7B model on 3T tokens; mid-training on in-domain data, context length extension 4k to 32k on 200k samples; SFT on 100k samples; DPO on 2000 samples over 3 rounds.

- **Infrastructure:** Developed in-house distributed training system on hundreds of nodes and inference pipeline. Both pipelines have been actively used since Jan 2024.
- **Pre-training:** Built automated pipelines to download and prepare data from open sources such as Wikipedia, arXiv, and CommonCrawl. Implemented multiple data-taggers to filter low-quality content for pre-training.
- **Mid-training:** Processed Fortune's proprietary data across modalities (online articles, magazine text, videos, tables, and ranking lists). Performed bias-detection to eliminate political biases.

- **Supervised fine-tuning:** Conducted 19 iterations of SFT on public and client data.
  - **General SFT data mix:** Curated open-source data sets (FLAN, Tulu, ShareGPT, etc.), identified biases such as time and identity bias, and iteratively refined data mixture.
  - **Fortune-specific data mix:** Generated Fortune-specific instruction pairs. Examples include fact-checking, trends analysis over time, and code generation for data visualization [Fortune Interview].
- **RLHF:** Conducted 3 cycles of training and evaluation, with human preference alignment through DPO.
- **Evaluation:** Designed evaluation benchmarks for both verifiable and open-ended questions. Performed LLM judge and human evaluations, with comparisons to external state-of-the-art models.

Insitro - Advanced Machine Learning South San Francisco, CA
Lead Machine Learning Scientist Nov 2022 – Nov 2023

- **Contextual Vision Transformer**: Addressed covariate shift in cell imaging by dynamically inferring and adjusting for batch-level covariates, mitigating unwanted confounding effects. Presented at SBI2 2023.
- **Channel Vision Transformer**: Enabled flexible adaptation to varying input fluorescence channels by flattening channel dimensions for cross-channel interaction, enhancing model accuracy and versatility. Presented at ICLR 2024.
- **Vision Foundation Model Pipeline**: Developed a self-supervised training framework for large-scale vision models, leveraging unlabeled data to learn robust feature representations for cell imaging at Insitro, which contributed to Insitro's ALS phenotype analysis report.

Emerald Innovations Cambridge, MA
Senior Machine Learning Scientist Jun 2022 – Oct 2022

- **Wireless sensing**: Built a dataset of human actions with paired video and RF signal data. Developed a deep learning algorithm mapping RF signal sequences to 3D human skeletons using 4D CNNs and Transformers.
- **Cloud infrastructure**: Built internal infrastructure to manage AWS training job submission and monitoring.

Selected Publications

(Please refer to my Google Scholar page for the full list.)

PromptBridge: Cross-Model Prompt Transfer for Large Language Models
arXiv:2512.01420
Yaxuan Wang, Quan Liu, Zhenting Wang, Zichao Li, Wei Wei, Yang Liu, ***Yujia Bao***

WebDART: Dynamic Decomposition and Re-planning for Complex Web Tasks
arXiv:2510.06587
Jingbo Yang, Bairu Hou, Wei Wei, Shiyu Chang, **Yujia Bao**

KVLink: Accelerating Large Language Models via Efficient KV Cache Reuse
NeurIPS 2025
Jingbo Yang, Bairu Hou, Wei Wei, ***Yujia Bao****, Shiyu Chang*

MCP-Bench: Benchmarking Tool-Using LLM Agents with Complex Real-World Tasks via MCP Servers
arXiv:2508.20453
Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, ***Yujia Bao****, Eugene Siow*

Collaborative Memory: Multi-User Memory Sharing in LLM Agents with Dynamic Access Control
arXiv:2505.18279
Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, ***Yujia Bao***

SFT-GO: Supervised Fine-Tuning with Group Optimization for Large Language Models
arXiv:2506.15021
Gyuhak Kim, Sumiran Singh Thakur, Su Min Park, Wei Wei, ***Yujia Bao***

Sample, estimate, aggregate: A recipe for causal discovery foundation models
TMLR 2025
*Menghua Wu, **Yujia Bao**, Regina Barzilay, Tommi Jaakkola*

Enhancing Retrieval Systems with Inference-Time Logical Reasoning
ACL 2025
*Felix Faltings, Wei Wei, **Yujia Bao***

Improving Data Efficiency via Curating LLM-Driven Rating Systems
ICLR 2025
*Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, **Yujia Bao**, Wei Wei*

LLM Unlearning via Loss Adjustment with Only Forget Data
ICLR 2025
*Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, **Yujia Bao**, Yang Liu, Wei Wei*

From Isolated Conversations to Hierarchical Schemas: Dynamic Tree Memory Representation for LLMs
ICLR 2025
*Alireza Rezazadeh, Zichao Li, Wei Wei, **Yujia Bao***

Harnessing Business and Media Insights with Large Language Models
arXiv:2406.06559
***Yujia Bao**, Ankit Parag Shah, Neeru Narang, Jonathan Rivers, et al.*

Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words
ICLR 2024
***Yujia Bao**, Srinivasan Sivanandan, Theofanis Karaletsos*

Contextual Vision Transformers for Robust Representation Learning
SCIS @ ICML 2023
***Yujia Bao**, Theofanis Karaletsos*

Learning to Split for Automatic Bias Detection
arXiv:2204.13749
***Yujia Bao**, Regina Barzilay*

Learning Stable Classifiers by Transferring Unstable Features
ICML 2022
***Yujia Bao**, Shiyu Chang, Regina Barzilay*

Predict then Interpolate: A Simple Algorithm to Learn Stable Classifiers
ICML 2021
***Yujia Bao**, Shiyu Chang, Regina Barzilay*

Few-shot Text Classification with Distributional Signatures
ICLR 2020
IBM Outstanding Poster Award
***Yujia Bao**, Menghua Wu, Shiyu Chang, Regina Barzilay*

Deriving Machine Attention from Human Rationales
EMNLP 2018
***Yujia Bao**, Shiyu Chang, Mo Yu, Regina Barzilay*

Awards

Exchange & Visiting International Student Academic Excellence Award
University of Wisconsin-Madison, 2016

Excellent Thesis Award
Shanghai Jiao Tong University, 2016

Scholarship for Academic Excellence
Shanghai Jiao Tong University, 2015

Broader Impact

- Accenture Pioneers Custom Llama LLM Models with NVIDIA AI Foundry
- Accenture and NVIDIA Lead Enterprises into Era of AI
- Accenture Launches AI Refinery for Industry to Reinvent Processes and Accelerate Agentic AI Journeys
- Accenture Expands AI Refinery and Launches New Industry Agent Solutions to Accelerate Agentic AI Adoption
- Accenture Introduces Trusted Agent Huddle to Allow Seamless, First-of-its-Kind Multi-System AI Agent Collaboration Across the Enterprise
- UOB and Accenture Collaborate to Transform Customer Experience Using Advanced Technologies Including GenAI
- Accenture Collaborates with Dell Technologies and NVIDIA to Accelerate Enterprise AI Transformation with AI Refinery
- Accenture Teams with NVIDIA to Advance AI Agenda for Europe with AI Refinery for Sovereign and Agentic AI
- Accenture Launches Distiller Agentic AI Framework to Accelerate Scalable Industry AI Solutions