

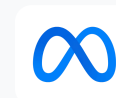
Simulation-driven AI Agent Testing and Improvement



Dr. Zhou (Jo) Yu

Co-Founder, CEO

- Columbia University CS Professor (CMU PhD)
- Open-source GenAI models (1M+ downloads)
- Forbes 30 Under 30, 17+ AI experience.
- Outstanding papers in ACL & NAACL 2024



amazon prize

Current Agent Ecosystem

AI Agent Verticals Applications

(Leverage the entire agentic infra to further enhance agents' usability through UI/UX, verifiers, human intervention, human teaching, etc to provide real ROIs to enterprise and individuals)

Agent Developer Tool

(Dev tools handel: memory, tool use (such as mcp), information extractor, fine-tuning, testing, evaluation, observation, security guardrails, etc)

Agent Foundation Models

(Continue training, reinforcement learning to adapt Foundation Models for agent tasks)

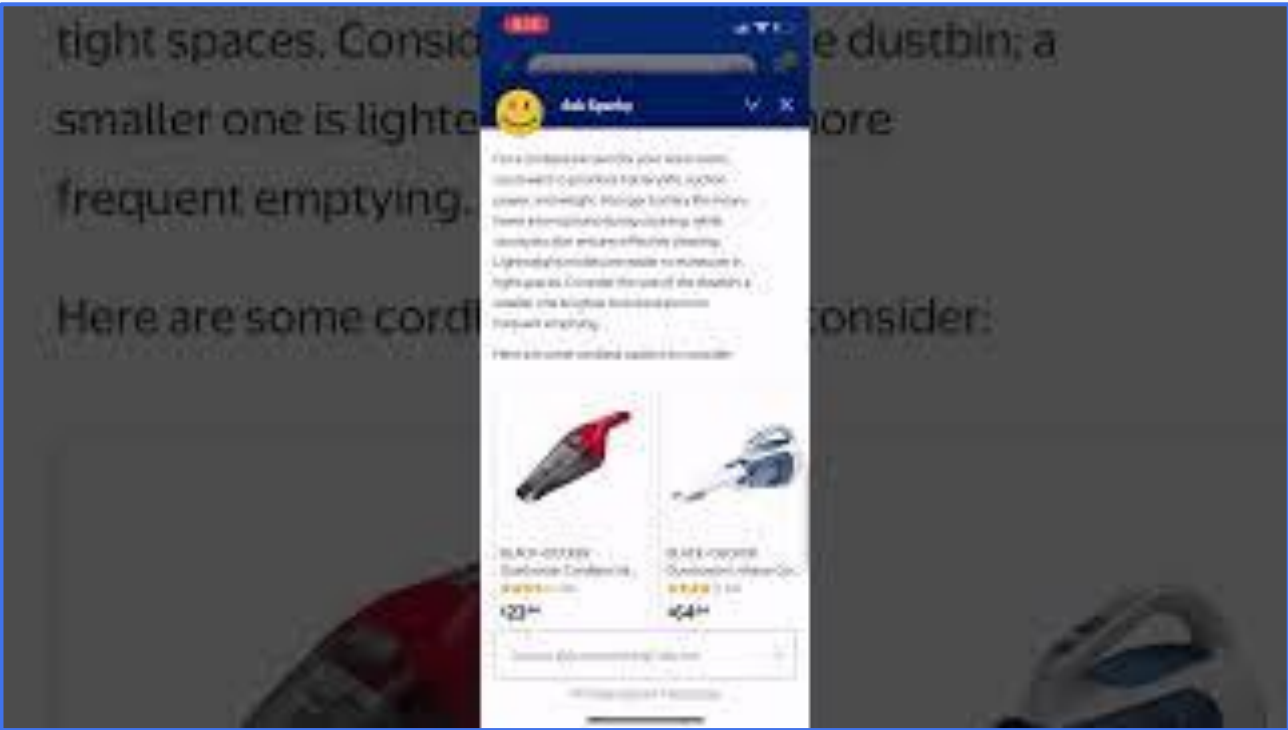
Foundation Models (openAI, Anthropic, etc)



95% Agents Stay as Demos

-MIT review

Walmart Sparky Agent x Arklex AI



Agent Recommendation by User Preferences

The screenshot shows a web browser window with the URL `arklex.ai/hub/bot/W8DXP1Zv9wLk239Mjqh65M/v1alpha1/voice-call/manual-calls`. The interface includes a sidebar with navigation options like 'Flow', 'Knowledge', 'Integration', 'Website Agent', 'Voice Agent', 'Settings', 'Outbound Calls', 'Scheduled Calls', 'Inbound Calls', 'Testing', and 'SMS'. A central 'Voice Call Details' modal is open, displaying the following information:

- Call Sid:** CA6cc5b5d8af0fa29fe5f86ca6b5d61bf
- Time:** Tue Jul 08 2025 15:44:30 GMT-0400 (Eastern Daylight Time)
- Recording:** 0:00 / 1:46
- Transcript:**
 - bot: Hi, this is Jordan from iChase. How can I help you today?
 - user: Hi, I'm interested in getting a credit card and one that fits my lifestyle.
 - bot: Absolutely, I'd be happy to help.
 - bot: Could you share what types of things you usually spend on or any features you're looking for?
 - user: So I order a lot of food delivery especially through Uber Eats, I shop online often, usually Amazon, and I occasionally book travel. I want something flexible when it comes to redeeming rewards.

Arklex AI Voice Agent



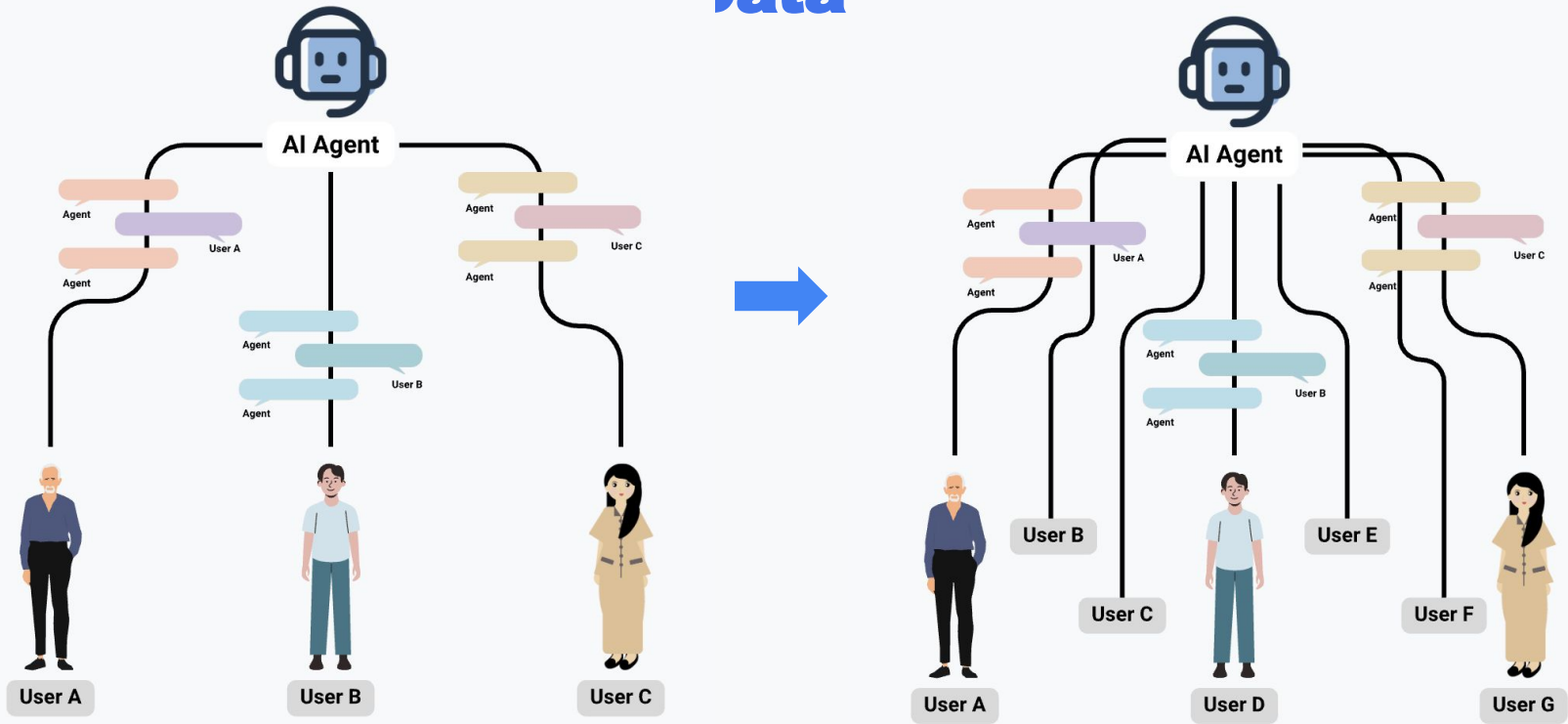
ARKLEX AI
AGENT: RECOMMENDATION BASED ON USER PREFERENCES

Gaps in Current AI Agent Evaluation

- **Single-turn evaluation**, where agents receive a single query and response it directly without multi-turn context consideration.
- **Static benchmark** that lacks mechanisms to accommodate iterative agent evaluation.
- **Lack runtime simulation environments**, no tool use, state tracking, and iterative task execution.
- **Human testing**: low coverage, slow turnaround, high cost, varied quality

SAGE: A Top-Down Bottom-Up Knowledge-Grounded User Simulator for Multi-turn Agent Evaluation, EACL 2026

Require Diverse and Realistic User Behavior Data



Solution: Synthetic Users

Synthetic user → *data for analysis and model improvement*



Scalable

cover diverse scenarios



Fast & On-demand

no human annotators



Easy to Adopt

validation, testing, analytics



80% faster development cycle



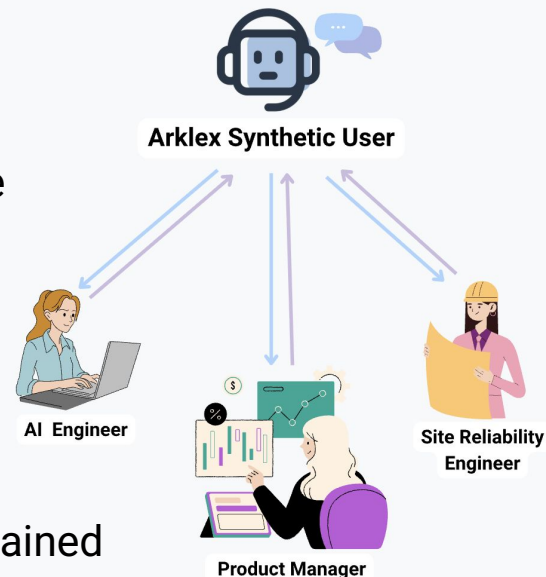
90% of testing automated



95% risk coverage



100% easier maintained



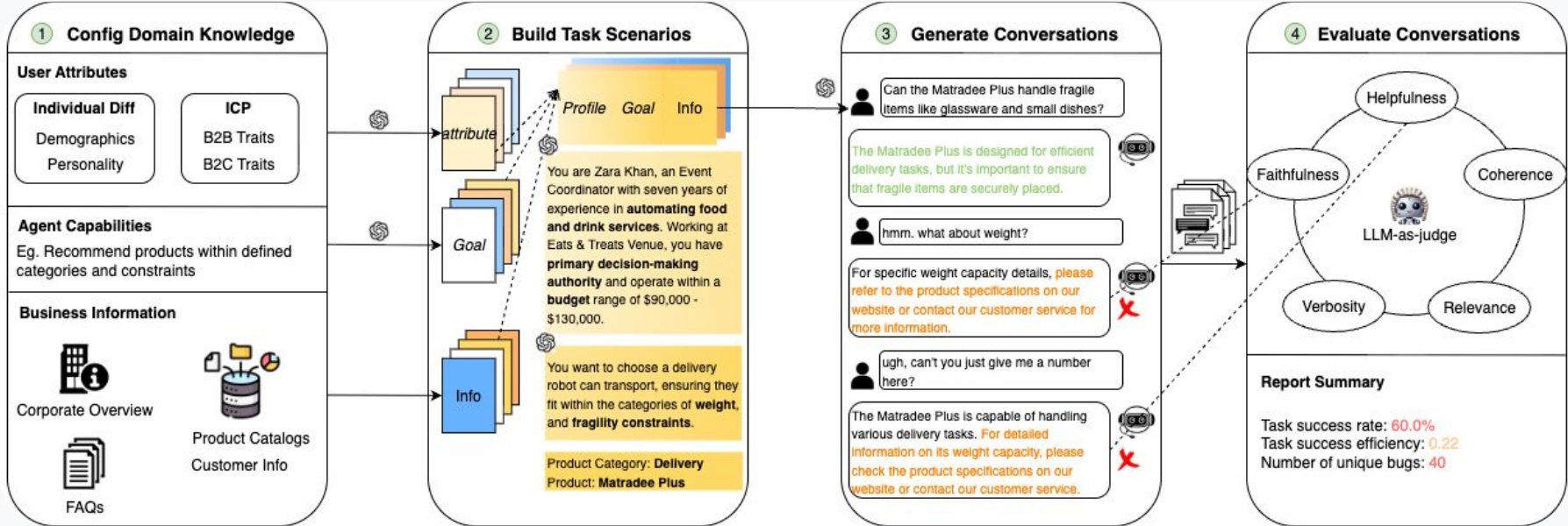
Solution: Arklex AI Simulator Demo



Open Source
Available

High-Level Architecture

Simulate real-world user interactions to evaluate the agent's performance.



User Simulation Config

1. Company Information

Describe your company

ABC Electronics is an online retailer, specializing in consumer electronics such as smartphones, tablets, and smart home devices.

2. User Needs (Agent capability)

What should the agent help your customers with?

- *General product information (e.g. Do you sell tablets?)*
- *Specific product details (e.g. Does your tablets use windows system?)*
- *Product recommendations (e.g. Which phone would you recommend for my dad who is 70 and is not that tech savvy?)*
- *Order tracking (e.g, Can you check when my tablet will arrive?)*

3. Agent call to actions

What specific actions do you want the agent to encourage your customers to take during the conversation?

- Deliver promotion code
- Encourage mailing list subscription
- Upsell and Cross-sell

4. Agent knowledge

Structured database schema

Unstructured data

5. Additional tools

Identify Agent Issues

Synthetic User: Hi, want to check my order status

Agent: Hi! I see you have two recent orders. Order #1299 has been fulfilled, and it includes a Textile City Faux Cashmere Sofa Blanket. Order #1300 is still pending fulfillment and includes a Nordic Knitted Blanket. Which order would you like to check?

Synthetic User: When will order 1300 ship?

Agent: Order #1300 is still pending fulfillment, which means it hasn't been shipped yet. Unfortunately, I don't have an exact shipping date at the moment. You can keep an eye on the status using the [Order #1300 Status] link. If you have any further questions or need more assistance, feel free to ask

Synthetic User: Can you tell me when it might arrive in Seattle?

Agent: I'm sorry, but I don't have the exact shipping date for Order #1300 at the moment, so I can't provide an estimated arrival time in Seattle. The order is still pending fulfillment. You can keep checking the status using the [Order #1300 Status] link.

Issue: The agent cannot access to the shipping information

Synthetic User: Can you give me the customer service email?

Agent: I'm sorry, but I can't provide the customer service email directly. However, you can usually find contact information on the \"Contact Us\" page of the Sweet for Everyone website.

Issue: The agent doesn't have the customer service email information

Evaluate Simulated Interaction Quality

- **Coverage**
 - Tool-calling Transition Entropy
 - Tool-calling Distribution Entropy
 - Trajectory Distance
- **Realism**
- **Cost**
- **Agent Failure Identification**

Evaluate Agent Performance (LLM-as-judges conditioned on scenarios)

Example result:

- **"customer_needs_completion_rate": 0.6**

Measures how often the agent completes the assigned goals

- **"customer_needs_completion_efficiency": 4.5**

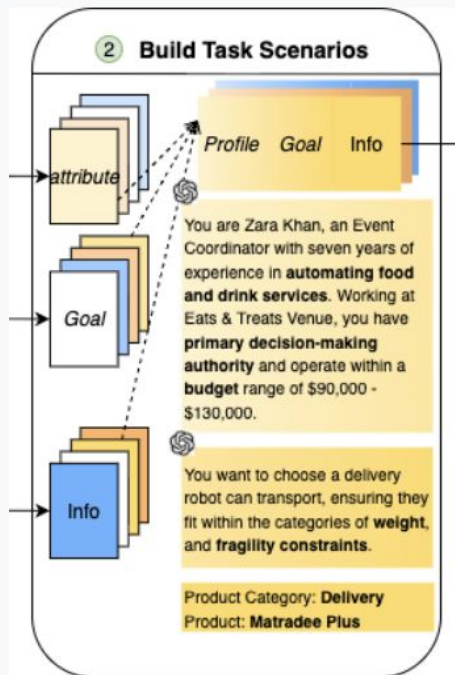
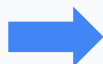
Quantifies how efficient is the agent to complete user simulator's goal

- **"call_to_action_completion_rate": 0.8**

Measures how often the agent achieves its intended call to action.

An Ever Evolving Golden Scenario Set

Production Logs



Arklex Business Model



Open-Source

- Free -> adoption funnel
- For developers & early adopters
- Fast prototyping




API

- Usage-based (\$1/interaction)
- Tight security constraints



Container

- Annual tiered license fee (\$100K+/year)
- Security-sensitive enterprises 
- Compliance & scale

Enterprise-Focused with Proven ROI



Lindsay Calder

Senior Program Manager



Pearson GED

Largest education company

“With Arklex, development are much faster and maintaining post-deployment system quality is easier”

How We Achieve This?



Deep Research: 10+ top tier publication with Amazon, Google, Microsoft (ACL & NAACL Best papers)



Proprietary Data: Rich personas, behavior data, and conversations



Domain Knowledge: Trusted by enterprise leaders (Walmart, TD Bank, etc)



Continuous improvement: Post-release real user data enhances simulation fidelity

Next Step: Continual Learning

1. Automatic Improvement

- Optimize the sales agent based on the simulated interactions on
 - Agent Flow Change
 - Prompts Optimization

Agent Flow Update Example:

- For the user query - I want to placing an order
- The original complete process is: 1. Get the user's account information. 2. Ask user which product do they want. 3. Add the product to cart. 4. Order placement
- However, by using the user simulator, the first two steps are always useless since the agent has user's login information and know which page user are looking at, so the agent don't need these two steps, so the final optimal process will skip 1&2 step but directly jump to step 3.

Next Step: Continual Learning

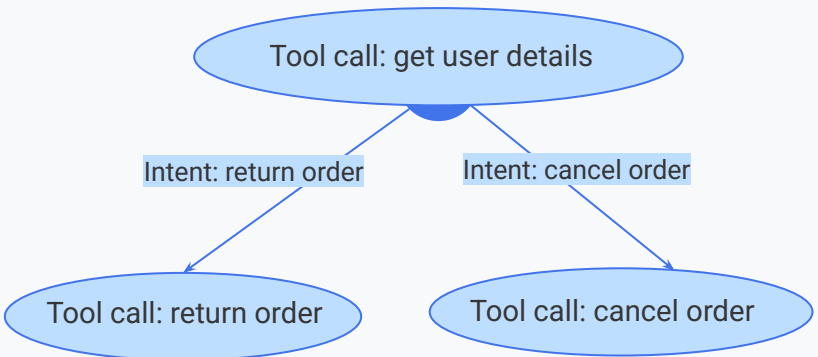
2. Manual Improvement

- Generate report to provide agent improvement suggestions for the developer
 - Missing RAG knowledge
 - Missing tools
 - Bad tools

Example:

```
“RAG”: [  
  {“issue”: “No estimated shipping info”, “suggestion”: “Add default fulfillment + shipping  
  estimates to RAG documents or assistant fallback responses”},  
  {“issue”: “Customer service email missing”, “suggestion”: “Include support email in RAG docs or  
  a static fallback response pattern”}  
],  
“<Tools>”: <XXX>
```

Before



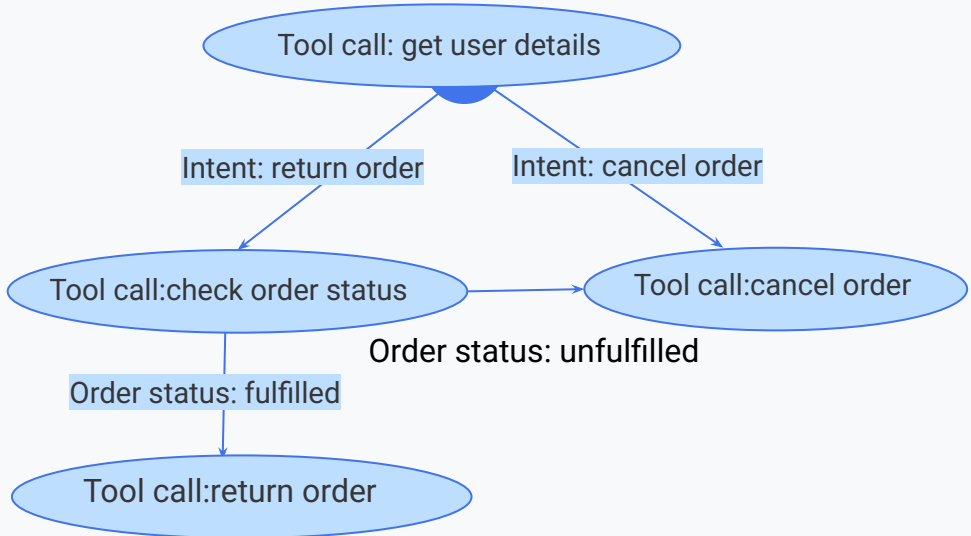
User: I want to return my order

Agent: Sorry there is nothing to return

Error: no valid order that is within return period.

Suggestion: This user might mean cancel order instead of return order, given the context that he has an active order but not delivered yet.

After



User: I want to return my order

Agent: I saw that your order #1001 kid's cap hasn't shipped it, would you like me to cancel it?

User: Oh, yeah, sure.

**Do we also need
simulation for workflow
agents?**

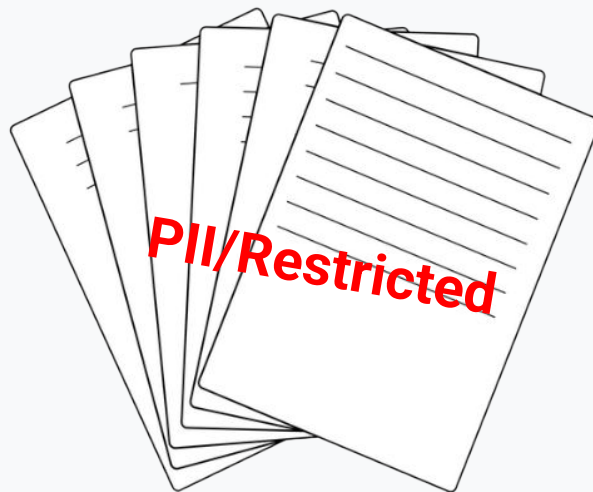
Where Underwriting Breaks Down

- Mortgage decisions depend on multiple documents
- Most errors come from cross-document inconsistencies
- Today's models are rarely trained on complete, realistic cases



Why Current Data Fails

- Real data is restricted and hard to scale
- Borrower stories are fragmented across documents
- Edge cases and real-world complexity are not controllable



How to safely train and evaluate underwriting systems without using real customer data, while still reflecting real-world complexity?

Persona-Based Document Synthesis

We generate **privacy-safe, end-to-end** synthetic borrower profiles that produce **complete, internally consistent** mortgage application documents for training and testing cross-document underwriting systems.

One Borrower

Persona

Eden Baker

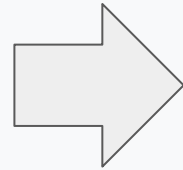
SIN Number: 570 809 590

Employer: Yukon Territory Systems

Employer Address: 165 Elm Street, Brampton, ON L6V 1A1

Employee Address: 2285 Maple Avenue, Saguenay, QC G7H 3A1

Employment Income: \$68,626.41 Income Tax Deducted: \$19,260.16



A Full Application

A grid of 10 document icons representing a full mortgage application. The icons are arranged in two rows of five. Each icon has a status indicator in the bottom right corner: a green checkmark for a complete document and a red 'X' for a missing document.

- Row 1: T4 (Tax) [checkmark], T5 (Tax) [checkmark], BANK STATEMENT [checkmark], ID DOCUMENT [X], DOWN PAYMENT PROOF [checkmark]
- Row 2: EMPLOYMENT LETTER [checkmark], CREDIT REPORT [X], PAYSTUB [X], HOME INSURANCE [X]

Synthesized Tax Documents

Persona

Eden Baker

SIN Number: 570 809 590

Employer: Yukon Territory Systems

Employer Address: 165 Elm Street, Brampton, ON L6V 1A1

Employee Address: 2285 Maple Avenue, Saguenay, QC G7H 3A1

Employment Income: \$68,626.41
Income Tax Deducted: \$19,260.16

Employer's name - Nom de l'employeur
Yukon Territory Systems

Canada Revenue Agency / Agence du revenu du Canada

Year / Année: **2025**

T4

Statement of Remuneration Paid / État de la rémunération payée

	Employment income / Revenu d'emploi: 14 68626.41	Income tax deducted / Impôt sur le revenu retenu: 22 19260.16
54	Employer's account number / Numéro de compte de l'employeur: 803547629RP0001	Employee's CPP contributions - see over / Cotisations de l'employé au RRC - voir au verso: 17 4160.00
12	Social insurance number / Numéro d'assurance sociale: 570809590	Employee's second CPP contributions - see over / Deuxièmes cotisations de l'employé au RRC - voir au verso: 17A 5.06
28	Exempt - Exemption: QC	El insurable earnings / Gains assurables d'AE: 24 63600.00
29	Employment code / Code d'emploi: 10	Employee's EI premiums / Cotisations de l'employé à l'AE: 18 839.52
Employee's name and address - Nom et adresse de l'employé Last name (in capital letters) / Nom de famille (en lettres majuscules): Baker First name / Prénom: Eden Address: 2285 Maple Avenue, Saguenay, QC G7H 3A1		Union dues / Cotisations syndicales: 44
		RPP contributions / Cotisations à un RPA: 20
		Charitable donations / Dons de bienfaisance: 46
		Pension adjustment / Facteur d'équivalence: 52
		RPP or DSPSP registration number / N° d'agrement d'un RPA ou d'un DSPSP: 50
		Employee's RPP premiums - see over / Cotisations de l'employé au RPAP - voir au verso: 55
		RPP insurable earnings / Gains assurables du RPAP: 56

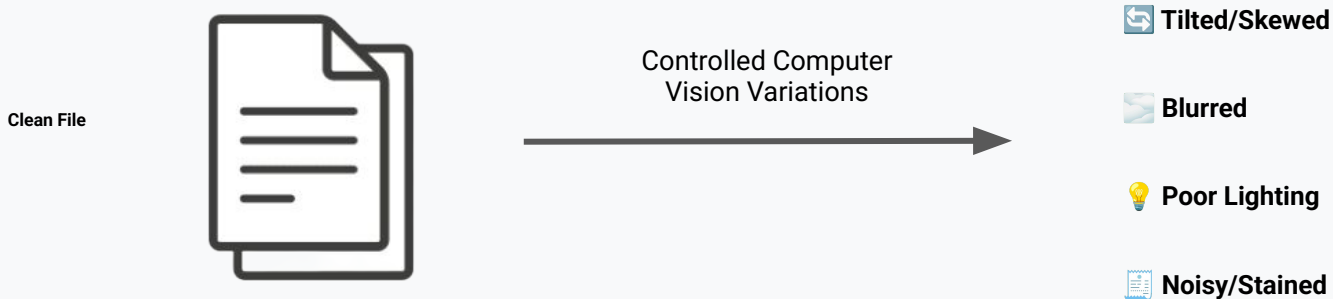
Other information (see over)	Box - Case: 38	Amount - Montant: ---	Box - Case: 34	Amount - Montant: ---	Box - Case: 40	Amount - Montant: ---
Autres renseignements (voir au verso)	Box - Case:	Amount - Montant:	Box - Case:	Amount - Montant:	Box - Case:	Amount - Montant:

T4 (24) Protected B when completed / Protégé B une fois rempli

Making It Real: Imperfect Documents

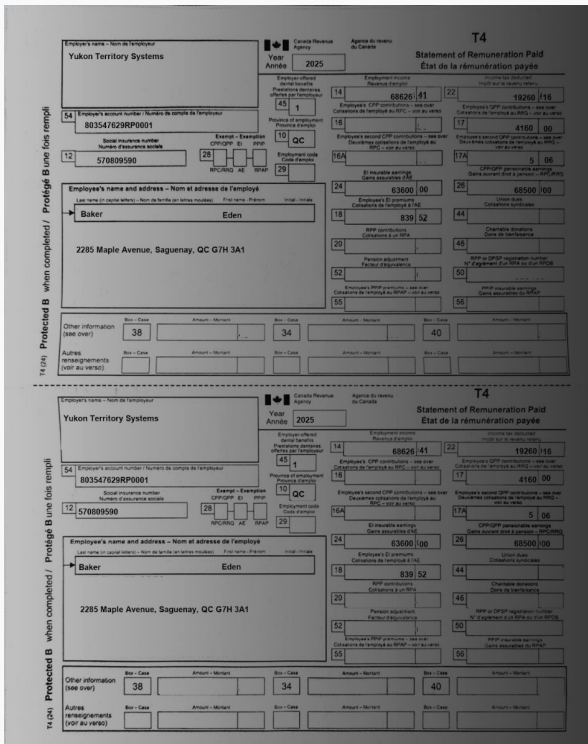
- **Problem:** Mortgage documents are often **scanned, photographed, and imperfect**, which significantly reduces extraction accuracy compared to clean PDFs.

- **How Arklex Simulates**

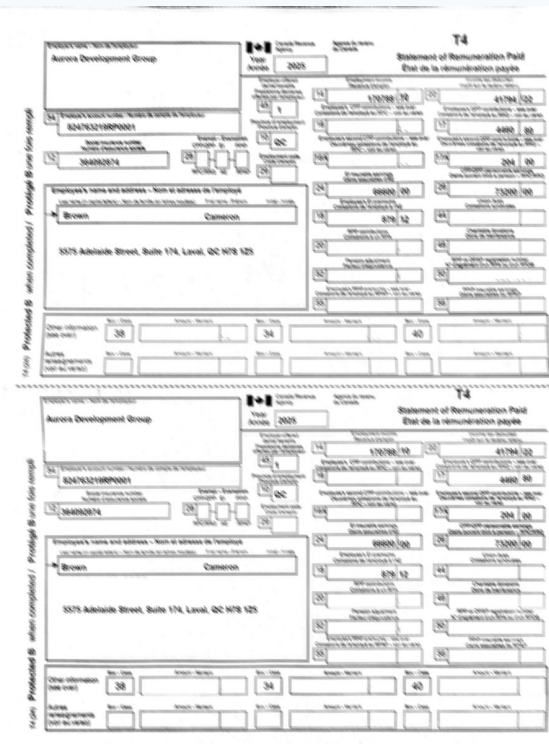


- **Why It Matters:** Models are trained on documents that match real customer submissions, not idealized inputs, leading to **more reliable underwriting decisions**.

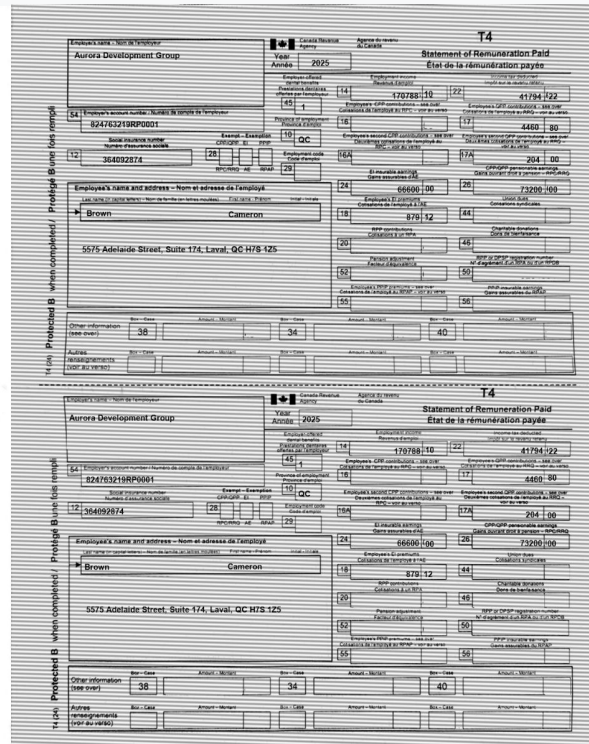
Synthesized Documents with Variations



Lightning



Blur (motion blur)



Noise (moiré effect)

Other Documents

Paystubs

PAY STUB / ADVICE OF DEPOSIT

Prairie Power Electrical Ltd.
11210 178 Street NW
Edmonton, AB T5G 298

Pay Frequency: Bi-weekly
Pay date: Jan 24, 2025
Pay period end date: Jan 17, 2025

SN 642-059-174

EARNINGS			Gross Pay	
Description	Rate	Hours		\$ 3,311.00
Regular Pay	38.50	80.00	Total Deductions	935.00
Overtime (1.5x)	57.75	43.00	Net Pay	\$ 2,376.00
Bonus			Year-to-date Gross	3,311.00
Commission	—		Year-to-date Net	2,376.00
RSU/Stock	—			
Gross Pay Totals	3,311.0	3,311.00		
DEDUCTIONS AND TAXES			Direct Deposit	Direct-Amount
Description	Current			
Federal Income Tax	425.52		Bank Transit/ Institution	10984 XXXXX4182
Provincial Income Tax	232.63		Account	2,376.00
CPP (Canada Pension Plan)	188.99			
EI	54.96		ACCRUALS/LF/AE	
Union Dues	—		Vacation accrued this period	3.09 hrs
Health Benefits	35.00		Vacation used	0.00 hrs
Life Insurance	8.00		Vacation balance	56.00 hrs
Other Deductions	—		Sick leave balance	40.00 hrs
Total Deductions	935.00			

Prairie Power Electrical Ltd.	Jan. 24, 2025	Pay date	Jan 24, 2025
Ethan McLeod		Net Pay	\$ 2,376.00
Employee ID PFE: 0374		Account	XXXXXX4182
Pay date	2,3761.000		

Prairie Power Electrical Ltd. Pay date Jan. 24, 2025

Bank statement

Notice details

Name: Ethan McLeod Address: 8429 110 Street NW, Unit 205
SIN: 642 059 174
Edmonton: AD T6G 2P8

Date of assessment: May 27, 2026

Tax assessment summary

We calculated your taxes using the amounts below. The summary is based information you provided or we received. Reep all slips and receipts.

Tax assessment summary

Line	Description	Final amount	CR/DR
15000	Total income	4,221.97	—
	Deductions from total income	0.00	—
23000	Net income	4,221.97	—
26000	Taxable income	4,221.97	—
35000	Total federal non-refundtital tax credits	—	—
42000	Net federal tax	0.00	—
42800	Net provincial/territorial tax	0.00	—
43500	Total payable	648.05	—
48500	Payment on filing	—	—
	Total payable minus total credits	-648.05	CR
	Penalties	—	—
	Arrears interest	—	—
	Balance from this assesament	-648.05	CR

Property tax

City of Edmonton

Assessment and Tax Levy
Ethan McLeod
9425 Lip Street NW, Unit 205
Edmonton, AB T6G 295
Legal Description Con 132-4557

2025 Final Property Tax Bill

Billing Date: May 24, 2025
Tax Year: 2025
Roll Number: 00133245705
Account Number: 48231/10205
Access Code: \$35412

ASSESSMENT AND TAX LEVY

Tax Class	Assessed Value	Municipal	Municipal	Region	Education
Residential	\$315,000	0.09695	\$2,815.25	—	\$999.25
Subtotal		\$2,813.25	\$2,815.25	\$0.00	\$905.25

LOCAL IMPROVEMENTS

Name	Amount	Tax Vet Levy Sub-Total	\$ 3,622.50
	\$0.00	Final 2025 Taxes	\$ 9925.50
		Less Interim Taxes Paid	\$0.00
		Arrears	\$0.00
		Penalties & Interest	\$0.00
		Credits/Adjustments	\$0.00
		TOTAL AMOUNT DUE	\$ 3,622.50

INSTALLMENT SCHEDULE

Due	June 30	\$,311.25	INSTALMENT	\$1,811.25
Sept-rit	Sept. 30, 2025	5,311.25	Payment meth das: 6:59	02/h

Payment methods: Online banking (City of Edmonton - Property tax, pre-authorized, monthly payments (PP), mail-in cheque, or in person of financial institutions. Please see reverse for arrears, penalties, and payment method details. Payments must arrive by the due date.

Installment 1 Receipt

Roll Number: 001632457205
Account: 48231/10205
Civic Address: 8423 110 Street NW, Unit 205
Due Date: Jun 20, 2025
Amount: \$1,811.25
Amount Paid:



Installment 2 Receipt

Roll Number: 00162245705
Account: 40231/10205
Civic Address: 8428 110 Street NW, Unit 205
Due Date: September 20, 2025
Amount: \$1,811.25
Amount Paid:



Self-Learning Agents



Reinforcement Learning is Search & Learn

Main Idea:

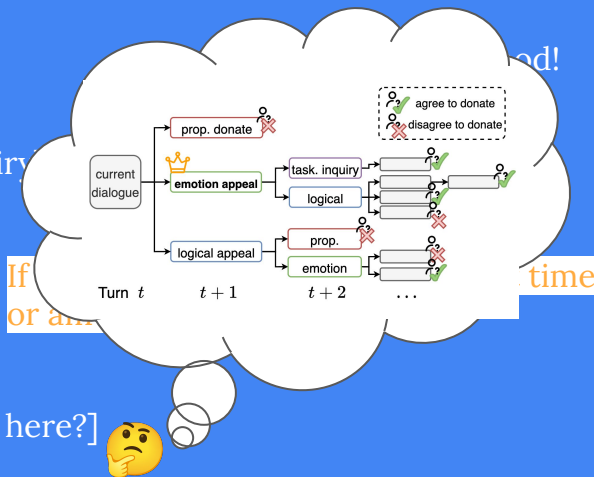
Many agents tasks are essentially about decision making
↳ We can use look-ahead search from games like GO to enhance this



[greet] Hello. How are you doing today?

[task-related inquiry]

[what should I say here?]



Persuadee

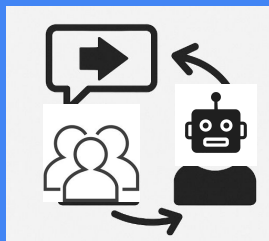
Self-improving AI agents optimized for user experience

User Simulator



- Accelerates iteration cycles
- Improves real-world robustness
- De-risks deployment

RL-based Agent Optimizer



- Enhances experience quality
- Learns from every interaction
- Maximizes downstream business impact

Application-First Auto-Scaling



- Reduces latency and cost
- Ensures performance under variable traffic
- Designed for AI Agents

Columbia DAPlab



Faculty



David Blei

CS ML
Causal inference



Carl Vondrick

CS Vision



Yunzhu Li

CS Robotics



Shipra Agrawal

IEOR
Reinforcement learning



Daniel Hsu

CS
Machine learning theory



Adam Elmachtoub

IEOR
Pricing and market design



Baishakhi Ray

CS Software
Neurosymbolic learning



Junfeng Yang

CS Systems
Security



Lydia Chilton

CS HCI



Zhou Yu

co-director CS
NLP



Kostis Kaffes

CS Systems



Eugene Wu

co-director CS
Data



Tianyi Peng

Columbia Business
Decision making



Xuhai Orson Xu

DBMI HCI
AI for Health



Richard Zemel

Advisory Board CS
ML



Michael Franklin

Advisory Board CS
UChicago

Arklex Synthetic Users Unlock Scalable Enterprise AI Agents Deployment

Dr. Zhou (Jo) Yu - Co-Founder & CEO

 zhouyu@arklex.ai

 <https://arklex.ai>



THANK YOU