



How to ensure your model does not drift

From the human-in-the-loop concept to building fully adaptive ML models using crowdsourcing

Fedor Zhdanov
Head of AI at Toloka




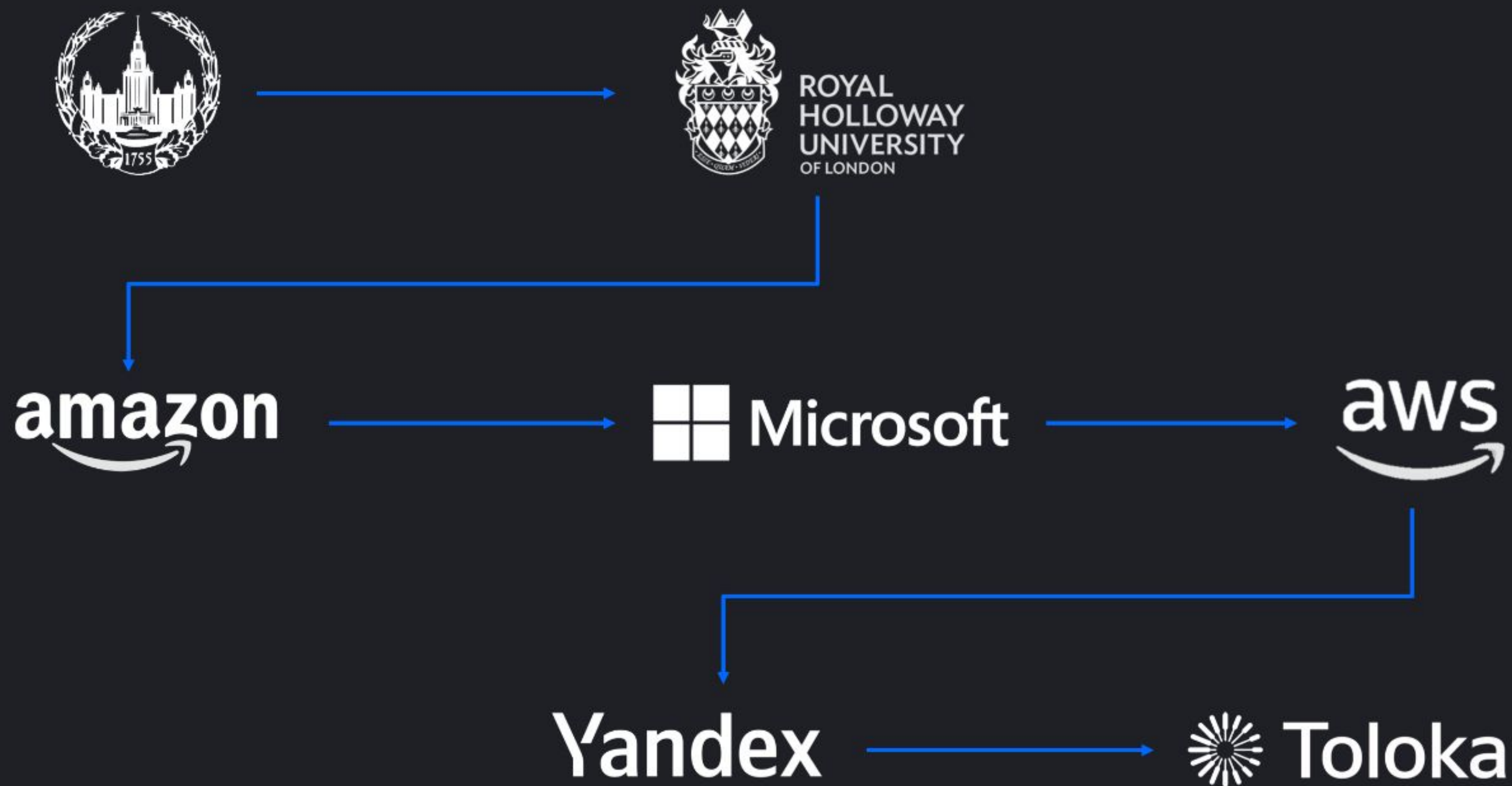
About me



Fedor Zhdanov

Head of AI at Toloka

 fedor@toloka.ai



How to ensure your model does not drift

1

Continually changing environment

2

How does it show itself for deployed ML models?

3

Toloka's approach with Adaptive ML Models

4

Industry & research use cases with adaptive ML and human oversight

The world changes every day

1 Changes in context

Data drift example: Scented candles and COVID-19

Before COVID-19

- “No scent” review and low rating
→ likely the product is bad

After COVID-19

- Spike in “No scent” reviews
- “No scent” review and low rating
→ doesn’t mean the product is bad

Review content didn’t change!
Very hard to infer problem from data alone



The world changes every day

1 Changes in context

2 The emergence of new phenomena

Data drift example: New spam and phishing schemes

Common spam letters years ago

from "Win 800,000:00 EUROS" <[redacted]> ☆
subject **You have Win 800,000:00 EUROS" Reply the departmet urgently"** 12/10/2011 03:05 AM
to [redacted] ☆ other actions ▾

[http://www.l\[redacted\]](http://www.l[redacted])

From the Department of LOTERIA PRIMITIVA in affiliation with Euro Million in Spain Branch has declared your Email/ profile on net WON; amounting the sum of 800,000:00 EUROS.
as our end of year's WINNING PROMOTION with LOTERIA PRIMITIVA " You can check on site for more update"
[http://www.l\[redacted\]](http://www.l[redacted])
You have been advise to send your Email and Tel/Mobil number to this departmental email address for the Claims through the Nominated Bank here in Spain.

Contact Email: [\[redacted\]](mailto:[redacted])
Tel: +34 [redacted]
Fax: +34 [redacted]

Recent spam schemes

Apple

Hi Customer,

Your Apple ID will Be disable Because of Some Violated Policies

The following changes to your Apple ID were made on 6 November, 2018

We have noticed that your account information appears to be invalid and unverified.

We need to verify your account information in order for you to continue using your Apple ID Account

You need to sign and verify it as soon as possible, you should do this soon because disabled accounts are eventually deleted along emails, iCloud, and other data stored with Apple

you should change your password as soon as possible from your Apple ID account page at <https://appleid.apple.com>.

Please verify your identity, we recommend that you go to [Verify Now](#)

[Sign In](#)

Sincerely,
Apple Support

Apple ID | Support | Privacy Policy
Copyright © 2018 One Apple Park Way, Cupertino, CA 95014, United States All Rights Reserved.

The world changes every day

1 Changes in context

2 The emergence of new phenomena

3 New meanings

Data drift example: Semantic changes

Then

Tesla



Now

Also Tesla



Types of data drift → Need to monitor ML

Distribution shift

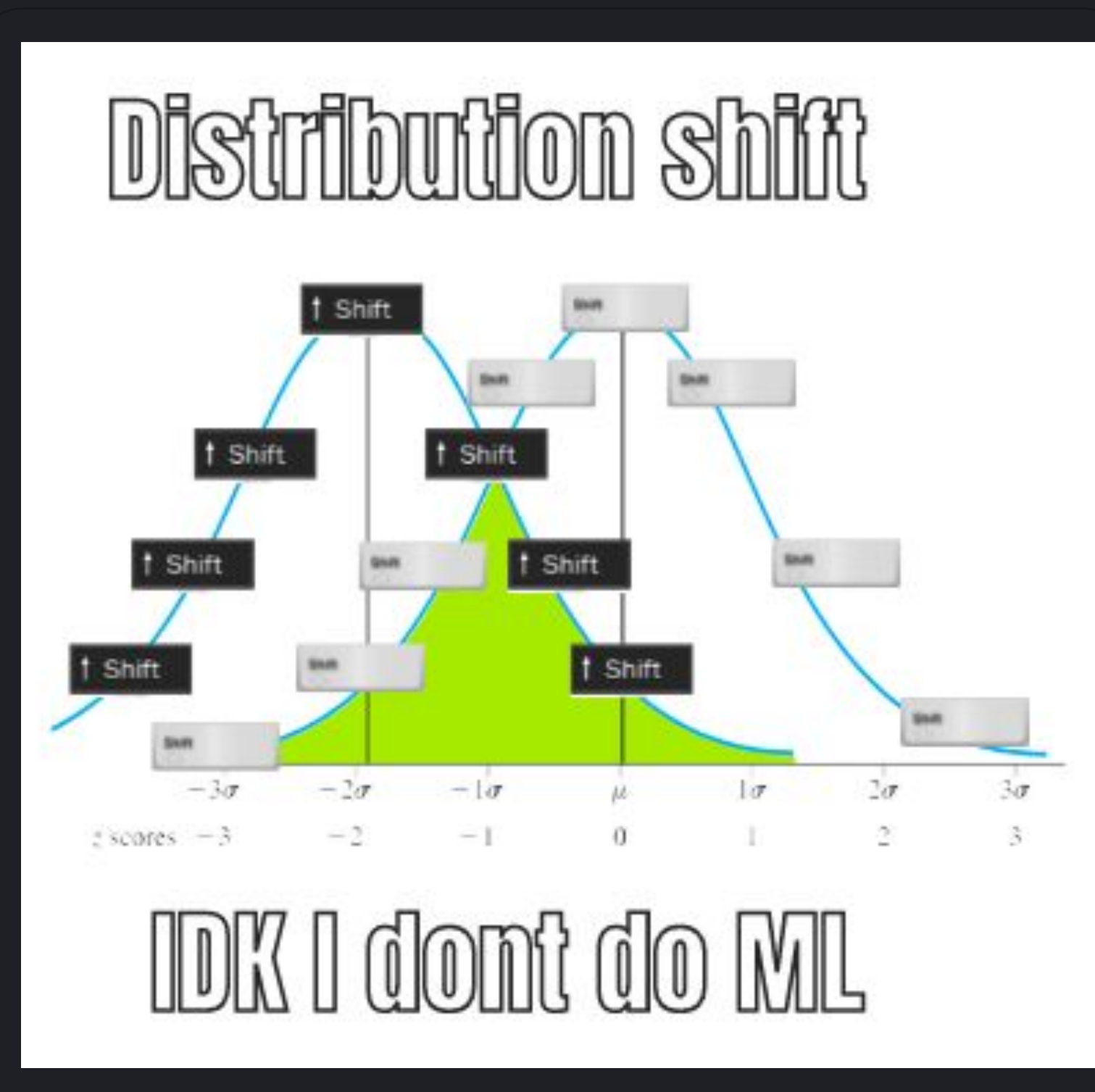
Data changes over time
(e.g. conversation styles change)

Train/test skew

Training data might not be representative

Label (concept) drift and more

Task details change over time



How do data drifts show themselves in deployed ML models?

Real-life example of changes in training data for self-driving cars

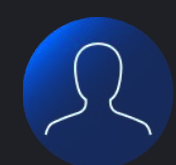


Challenge

Label images of roads with snow within a couple of days after sudden snowfall



Real-life case on outdated model for SMM classification



Challenge

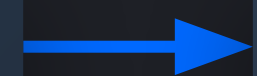
The client was handling 36K items per day with an outdated model trained **on 1 year old data**, and sending an additional 19K items per day for manual labeling.

The final goal was to optimize the process for classifying social media mentions as spam/feedback/neutral/news and retrain model on updated data.



Result

19K
items



9K
items

per day for manual labeling

Even ChatGPT has limitations



How recent is your training data?



My training data was collected and used to create my programming before September 2021.



I do not have access to any new information or updates since that time and my knowledge is based on the data that was available at the time of my training. As a language model, I am not able to browse the internet or access new information, so I am unable to provide information about events or developments that have occurred after my training data was collected. However, I can provide general knowledge and information about a wide range of topics that were current at the time of my training. If you have any specific questions, please don't hesitate to ask and I will do my best to provide helpful and accurate information.

Example with recent earthquakes in Turkey



When was the last earthquake in Turkey?






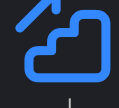


The most recent earthquake in Turkey occurred on February 27, 2021, with a magnitude of 5.6 on the Richter scale. The earthquake struck in the province of Bursa, located in the northwestern part of the country. There were no reports of serious injuries or damage from the earthquake, but it caused some panic and concern among local residents. Turkey is located in a seismically active region, and earthquakes are not uncommon in the country. The government has taken steps to improve earthquake preparedness and response, but earthquakes continue to pose a significant risk to the safety and well-being of people in Turkey.



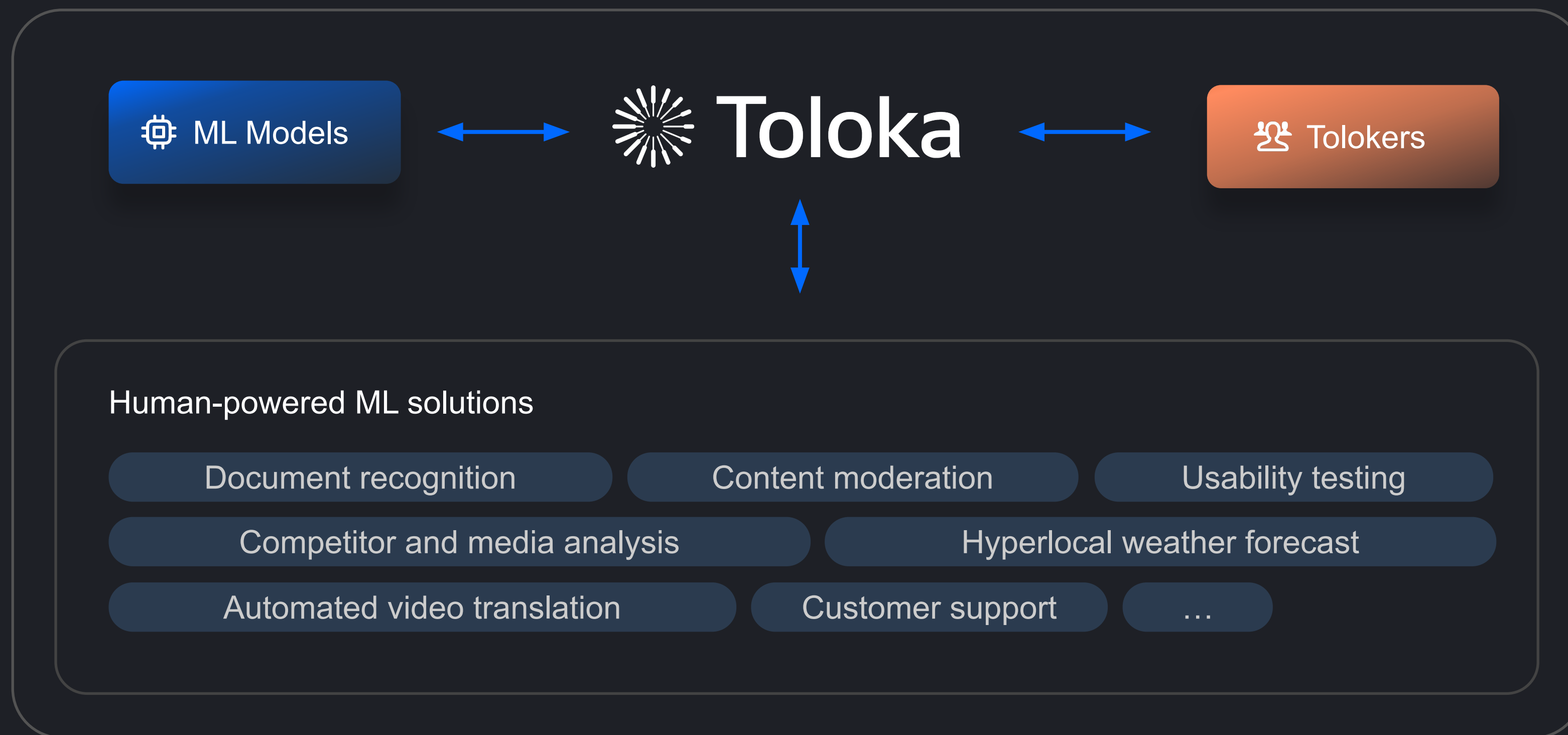
Counting on human oversight: Intro to Toloka and how we use data labeling in ML projects

Toloka supports data-related processes across the entire ML lifecycle

From data collection and annotation to model training, deployment and monitoring

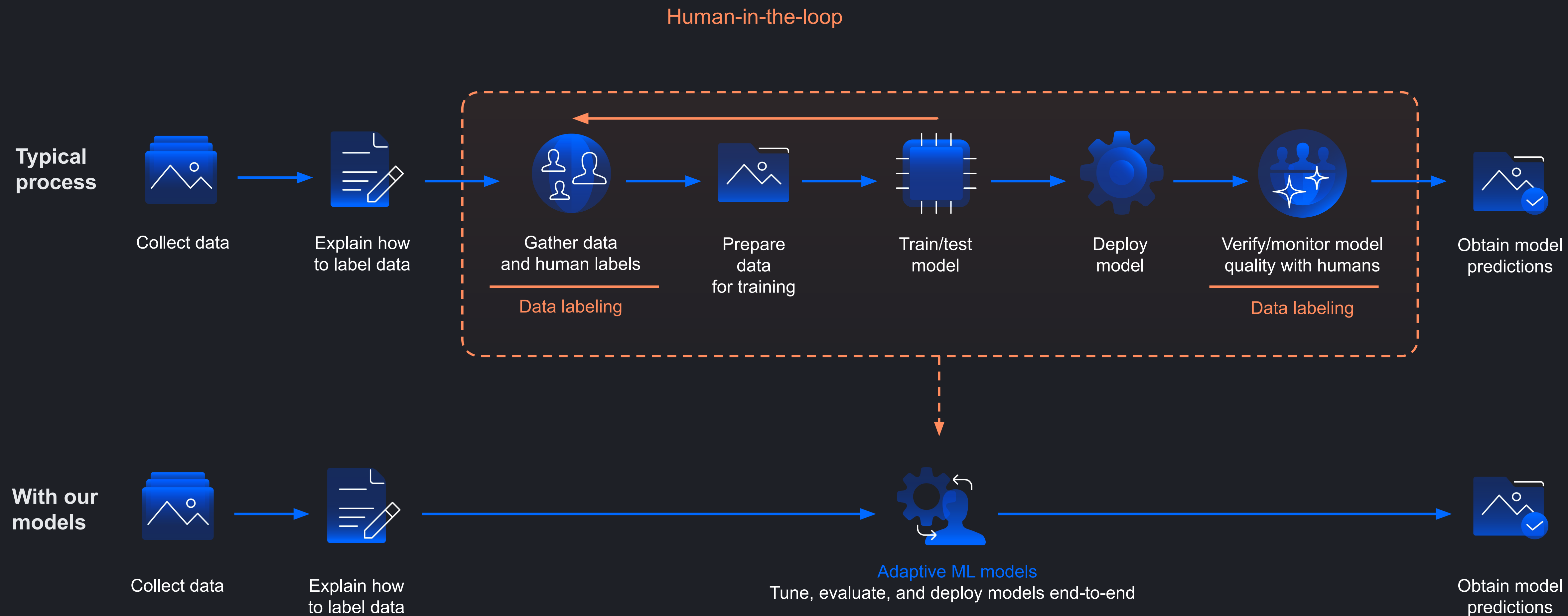
-  Data collection
-  Data processing
Store, process and clean data
-  Data annotation
-  Data analysis
-  Model training, deployment, and evaluation
-  Model monitoring

Intelligent platform and technological infrastructure

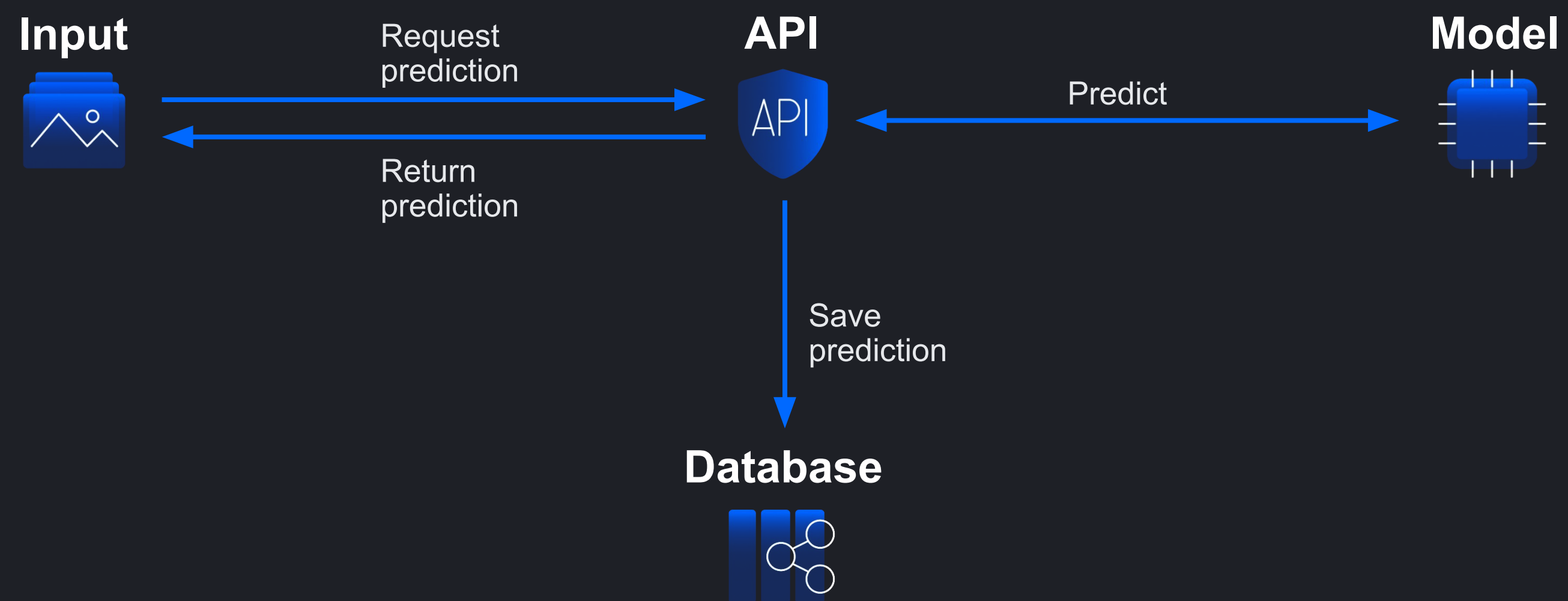


Introducing Adaptive ML models

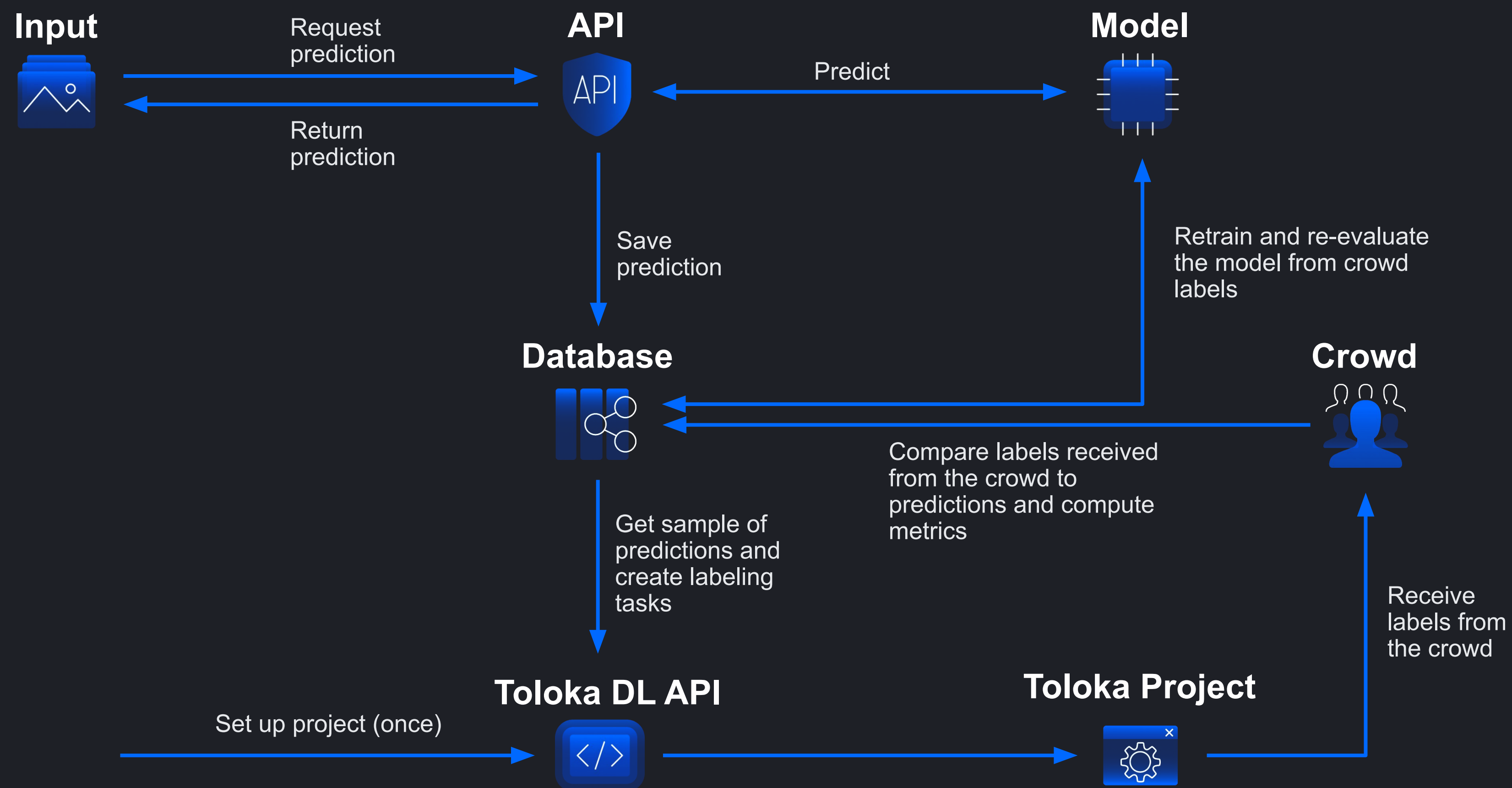
Adaptive ML models with human-in-the-loop flows



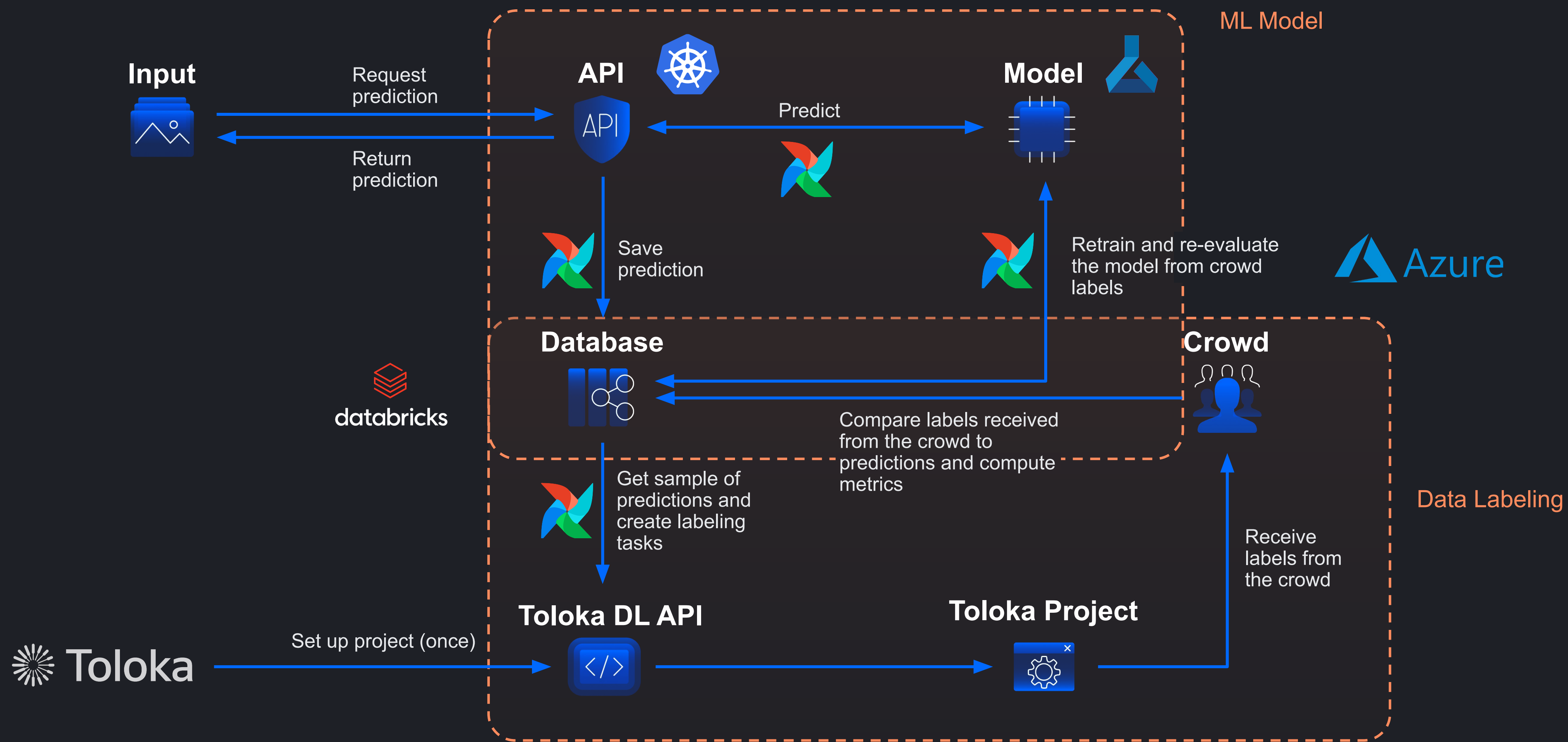
Non-adaptive model



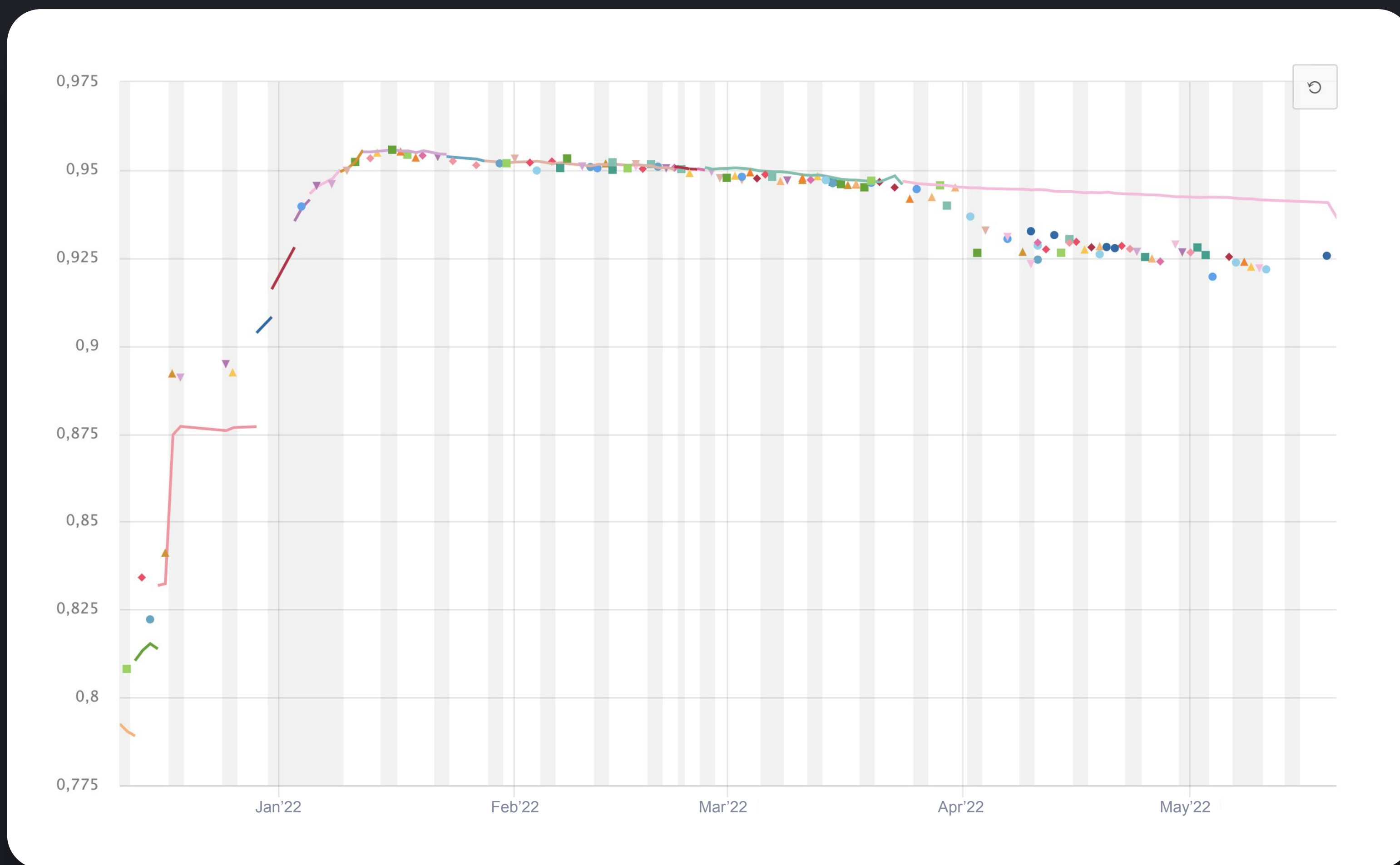
Adaptive model with human-in-the-loop process



Adaptive model with human-in-the-loop process



Report: model retrains

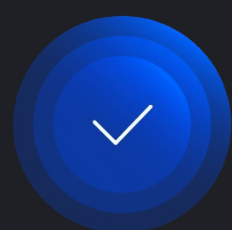


Advantages and disadvantages of customizable vs adaptive

	Ready to use models (AWS Rekognition, Azure Content Moderation)	Toloka Adaptive ML	Custom built models
Cost to launch, time to market	\$	\$\$ Pre-trained on our huge corpora collected over 10+ years of research The infrastructure for training and hosting is built once	\$\$\$\$\$
Accuracy	Low	Medium-High Adapt pre-trained models to your labeled data	High
Human verification	⊗	✔ Monitor models with human verification post deployment, avoid data drifts	⊗

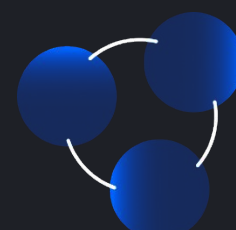
Industry & research use cases with adaptive ML and human oversight

Content moderation for public messages



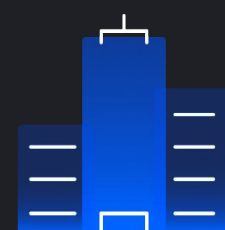
Why it's important

- Protects the messenger from illegal content
- Helps detect and prevent fraud
- Enables good user experience and retention



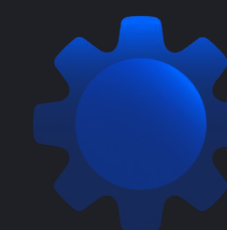
Toloka's solution

Human in the loop pipeline with Adaptive ML Models for moderating messages



Business impact

- Full compliance with company policies
- Enable growth in new regions
- Better customer experience with improved safety



Technical details

- Data: messages and threads
- For initial pipeline setup: 1000 messages and instructions for the classes
- For pilot: 100K messages
- Production volume: 10K-20K messages/day

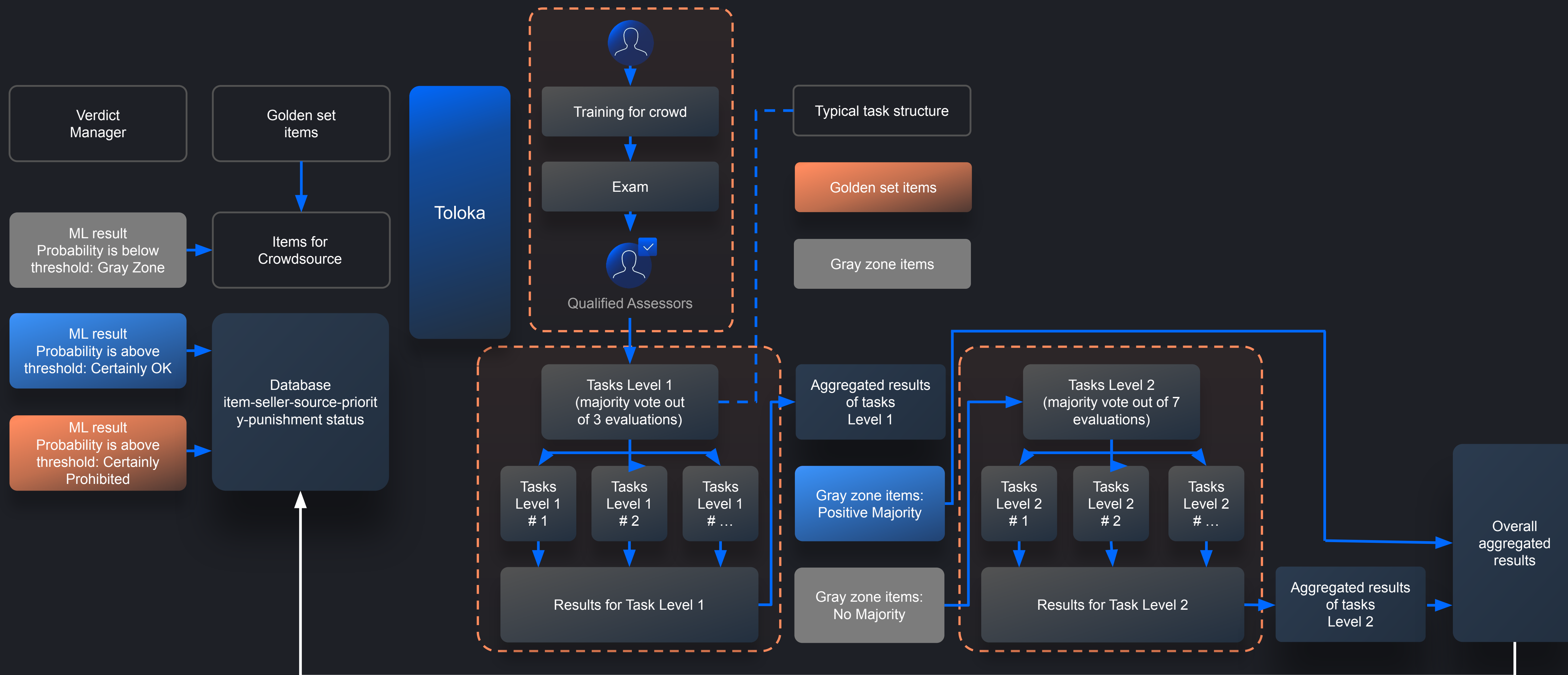
Growth

10K items/day for EN, UK, RU

30-40%

reduction in escalations

Moderation pipeline example



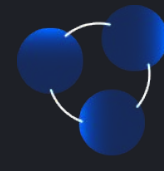
Document recognition AI with human-in-the-loop

Tech startup



Client challenge

AI-powered data extraction tool helps large companies analyze, categorize, and retrieve customer information from scanned documents in seconds. The algorithm needs continual retraining



Solution offered

The company uses Toloka to build human verification into its pipeline with human input 24/7 from around the globe. Results are received in minutes



Business impact

More than 4500 verification tasks are sent to Toloka via API every day. The human-in-the-loop flow ensures accuracy of over 99% for the company's document recognition software

6–8 min
per document

99%
accuracy

Are the string values the same
in each pair below?

Claim number

198 - 24 - 3401

Yes

No

198 — 24 — 3401

Phone number

(507) 638 - 1485

Yes

No

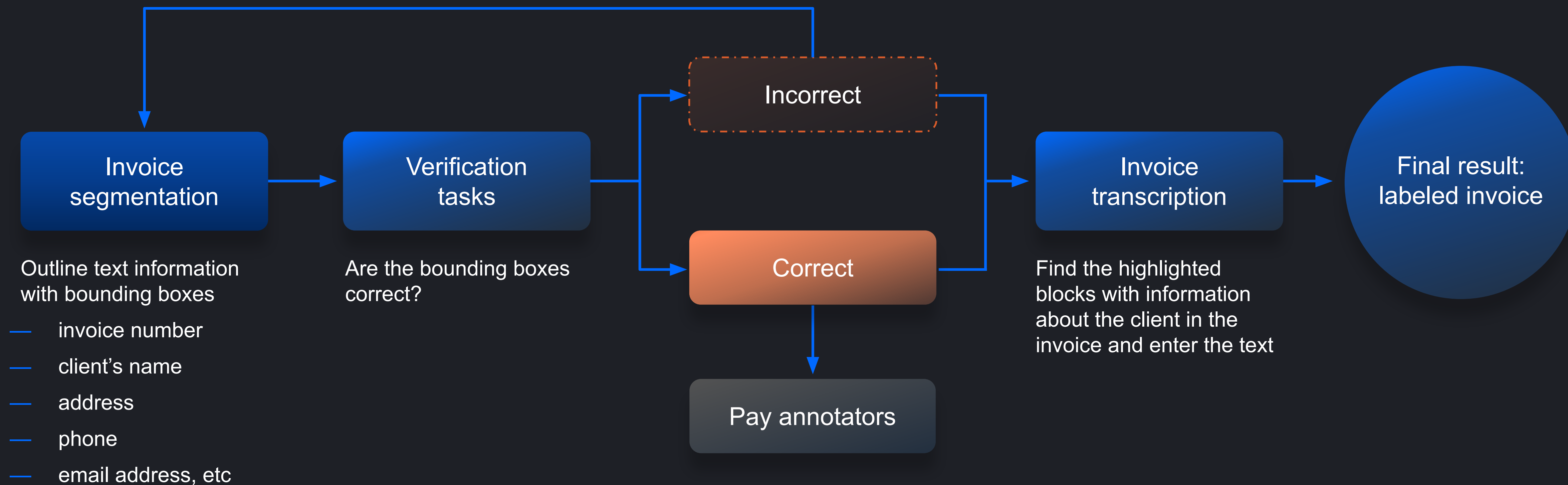
0501) 639 — 1486

(507) 638-1485 |

Done

Text recognition for document processing

Execution pipeline

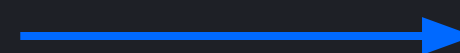


So why do we need human oversight here?

- Having reliable accuracy metrics
- Verifying that the model performs as expected over time

Best prompts for Text-to-Image models

Find best prompts for Text-to-Image models

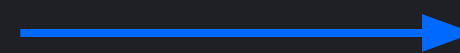


[keyword1, ..., keywordM-1] “A portrait painting of daenerys targaryen queen” [keywordM, ..., keywordN]

Find best prompts for Text-to-Image models



Top 15 keywords



Our approach

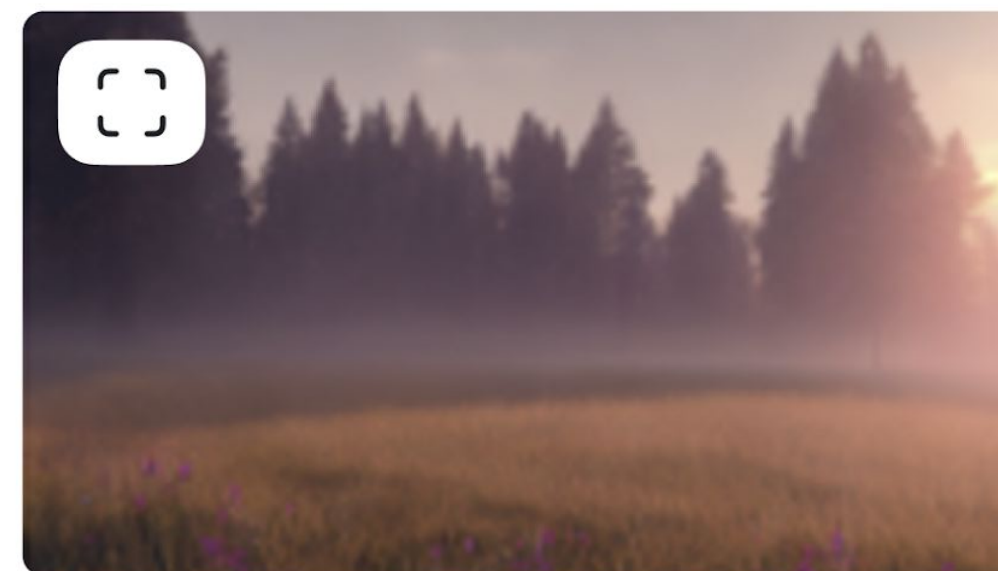
Find best prompts for Text-to-Image models

The process

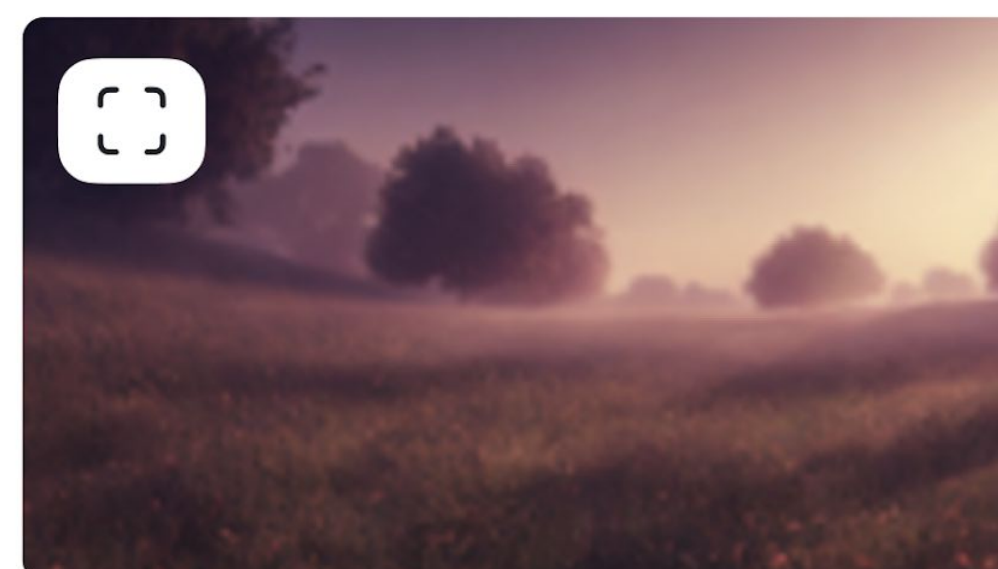
- Tolokers were shown the description and two sets of images, produced with a different set of keywords
- Tolokers chose the most aesthetically-pleasing image for each description
- Images generated by another text-to-image model were used for qualification tests

Beautiful meadow at sunrise, thin morning fog hovering close to the ground

A



B



Which set of images is aesthetically better?

1



A

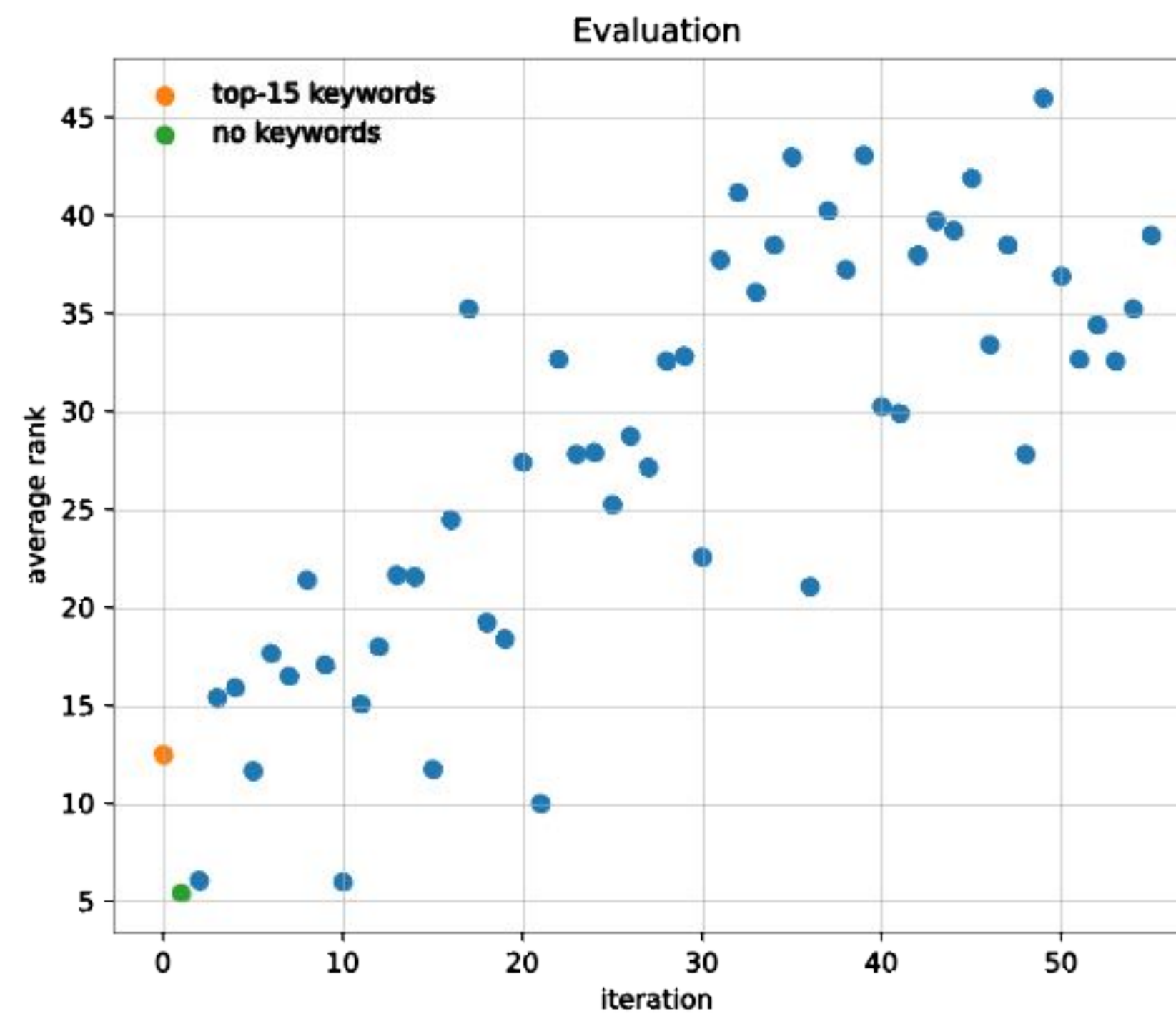
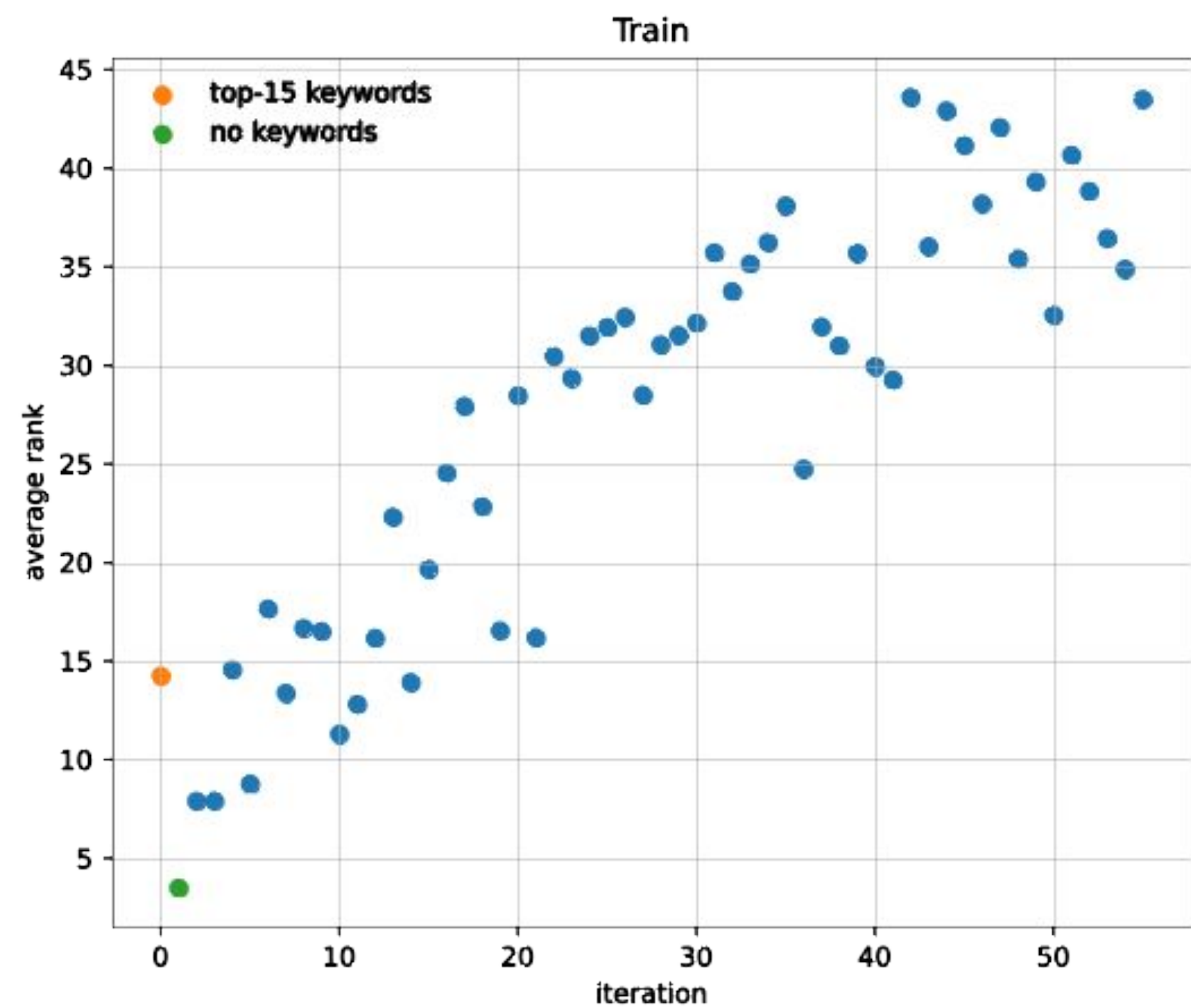
2



B

Genetic algorithm iteratively improves the choice of keywords

It works better than using top 15 keywords or no keywords at all



The best keywords

As selected by our method

- cinematic
- colorful background
- concept art
- dramatic lighting
- high detail
- highly detailed
- hyper realistic
- intricate
- intricate sharp details
- octane render
- smooth
- studio lighting
- trending on artstation

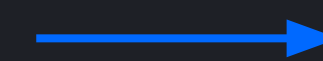


Learning from subjective data: IMDB-WIKI-SbS

Learning from subjective data

Dataset

- Most crowd tasks are classification, which is objective
- Information retrieval and recommender systems need subjective opinions of humans
- Pairwise comparisons work well for gathering subjective opinions, but these methods need evaluation



IMDB-WIKI-SbS

New large-scale dataset for evaluation of pairwise comparisons

Learning from subjective data

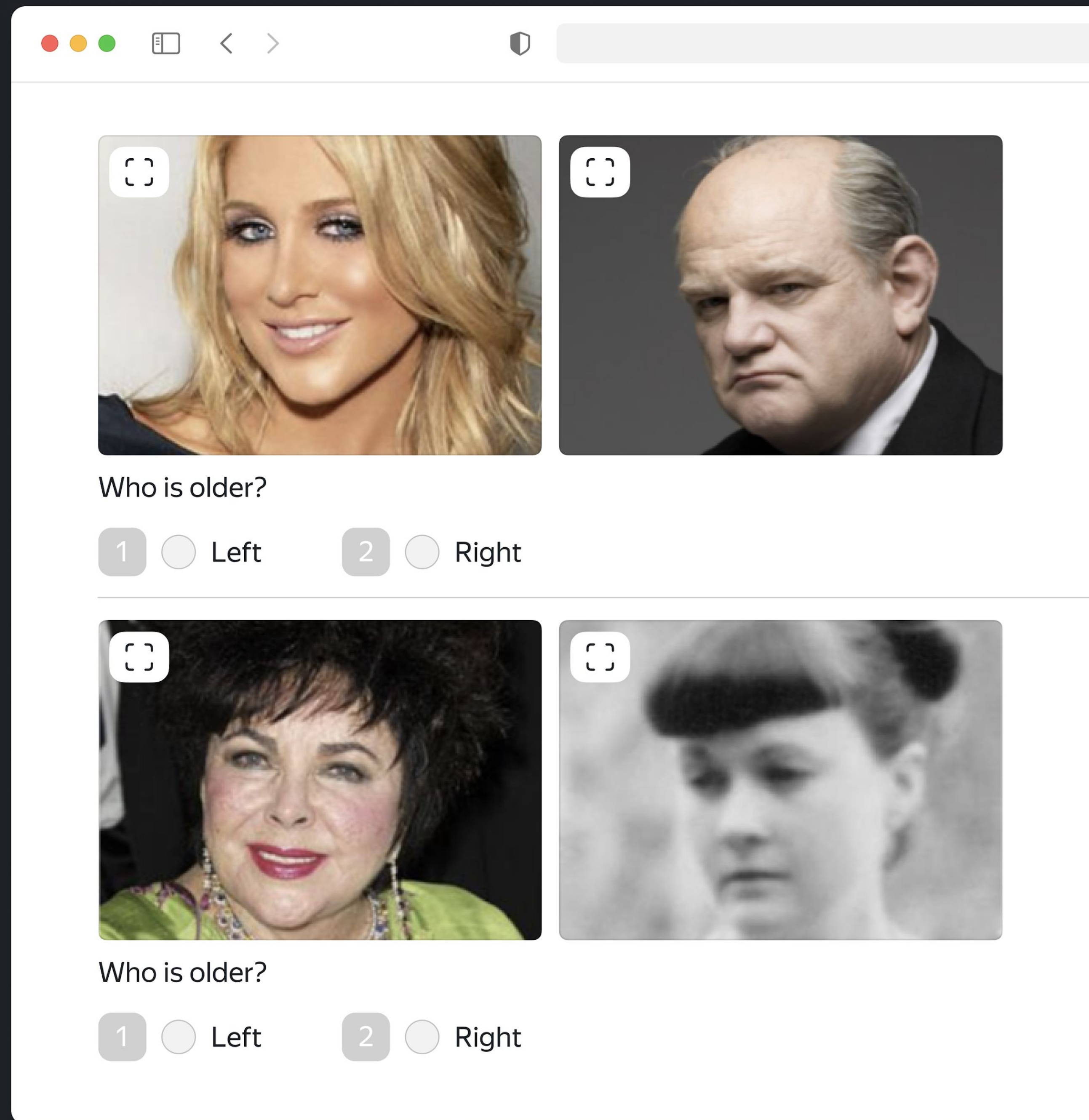
Dataset

- The IMDB-WIKI-SbS dataset uses the age information offered by IMDB-WIKI as ground truth
- It has a balanced distribution of ages and genders of people in photos
- We can use it to evaluate methods for gathering subjective opinions

250K
comparisons

9K
objects

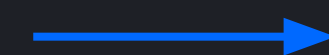
4K
Tolokers



Reinforcement learning with human feedback

Can the crowd teach a "robot" to do a backflip?

- In the Reinforcement Learning problem, an algorithm interacts with an environment and receives feedback — a reward
- The agent takes actions that change the environment to maximize the reward
- Ready reward function can use ready-to-go implementation of RL algorithms
- The reward function needs to be defined by the engineer, but it requires a lot of time and effort



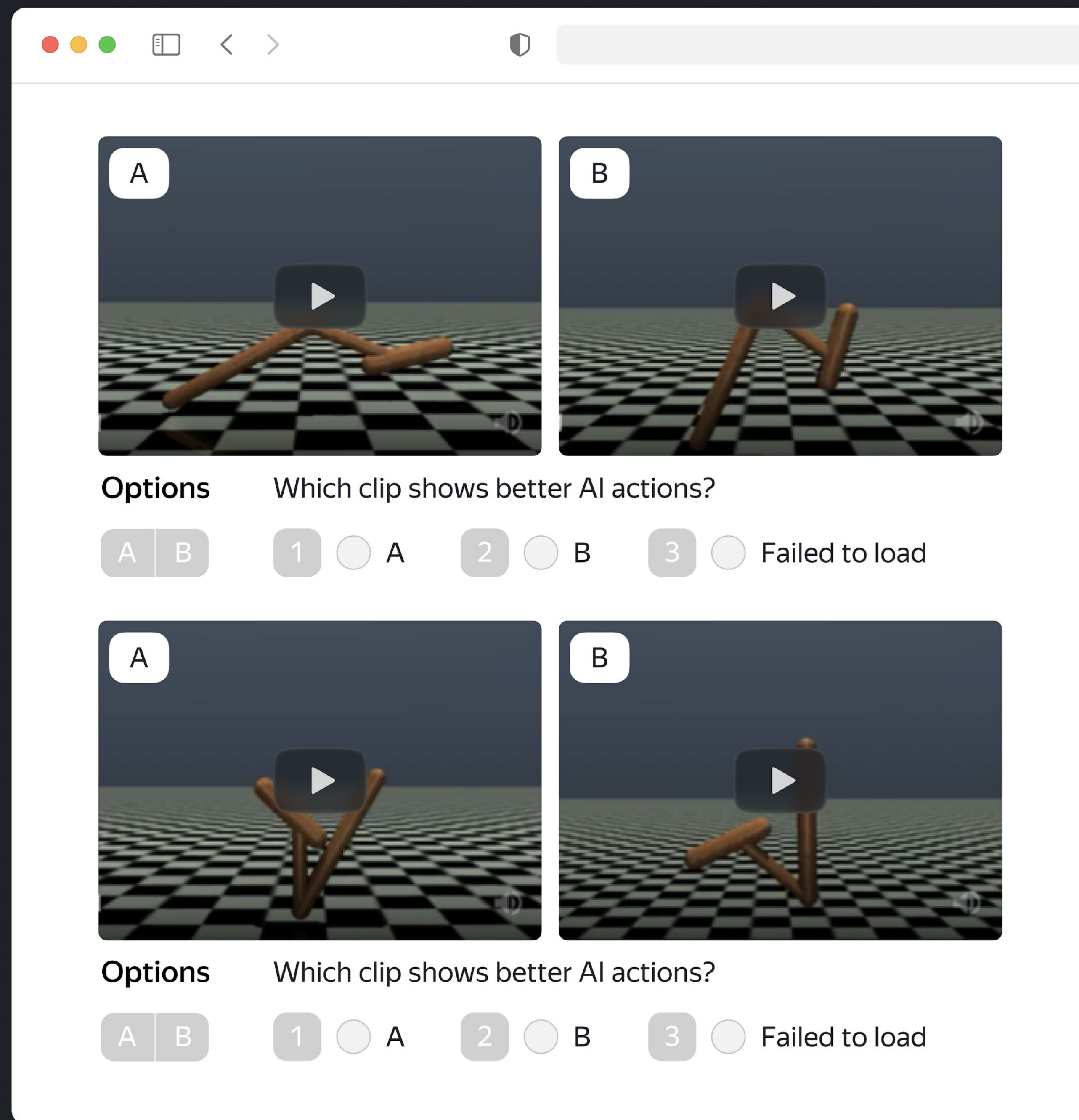
Don't define the reward yourself — ask the crowd!

Learning from subjective data

Replication of the paper “Deep Reinforcement Learning from Human Preferences” by OpenAI and DeepMind using Toloka

The process:

- Let the agent do a backflip
- Sample random trajectory pairs
- Run pairwise comparisons with crowd
- Train a network to predict a reward based on comparison results
- Train agent to maximize predicted reward



So why do we need human oversight here?

Best prompts for Text-to-Image models

- Coming up with good keywords is hard but can scale with large amounts of small tasks for humans

Learning from subjective data: IMDB-WIKI-SbS

- Datasets with subjective comparisons are rarely available
- There are few if any out of the box models available capable of predictive subjective perception

Reinforcement learning with human feedback

- Reward functions can be hard to develop — human opinion can be instrumental in making progress in RL

Human vs Generative Models: place your bets

Our preliminary experiments

Review classification: binary classification with clear patterns (also easy)

GPT-3: Accuracy = 88%

Tolokers: Accuracy = 96%

Our preliminary experiments

Message classification, mostly simple multiclass classification

GPT-3: Accuracy = 69.4%

Tolokers: Accuracy = 100%

What's next?

- Human insight offers [textual feedback for the instructions and the project](#), while an LLM only follows the prompt, which could be wrong
- LLMs can be p-tuned (“adapted”), requiring hundreds of examples (OpenAI offers that), but it is still prone to [data drift](#)
- Difficult tasks require more diligent Tolokers. [LMs and synthetic data can help us select and train Tolokers](#), write task instructions, etc.

Thank you!



Fedor Zhdanov

Head of AI at Toloka

 fedor@toloka.ai



www.toloka.ai

LinkedIn



GitHub

Toloka Adaptive Models: reach out

- You are an engineer who wants to try and build their system with ML services
- Try our solution and influence our roadmap
- Reach out to us about your problem where you want to apply ML



<https://tolokamodels.tech>