Generative Models For Data Augmentation

Brandon Trabucco, 4/9/25

Machine Learning Department
Carnegie Mellon University
School of Computer Science

Towards Generative Data Augmentation

Horizontal Flip

RandAugment
(Cubuk et al., 2019)

CutMix
(Yun et al., 2019)

[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] Yun et al., CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, ICCV 2020.

Towards Generative Data Augmentation

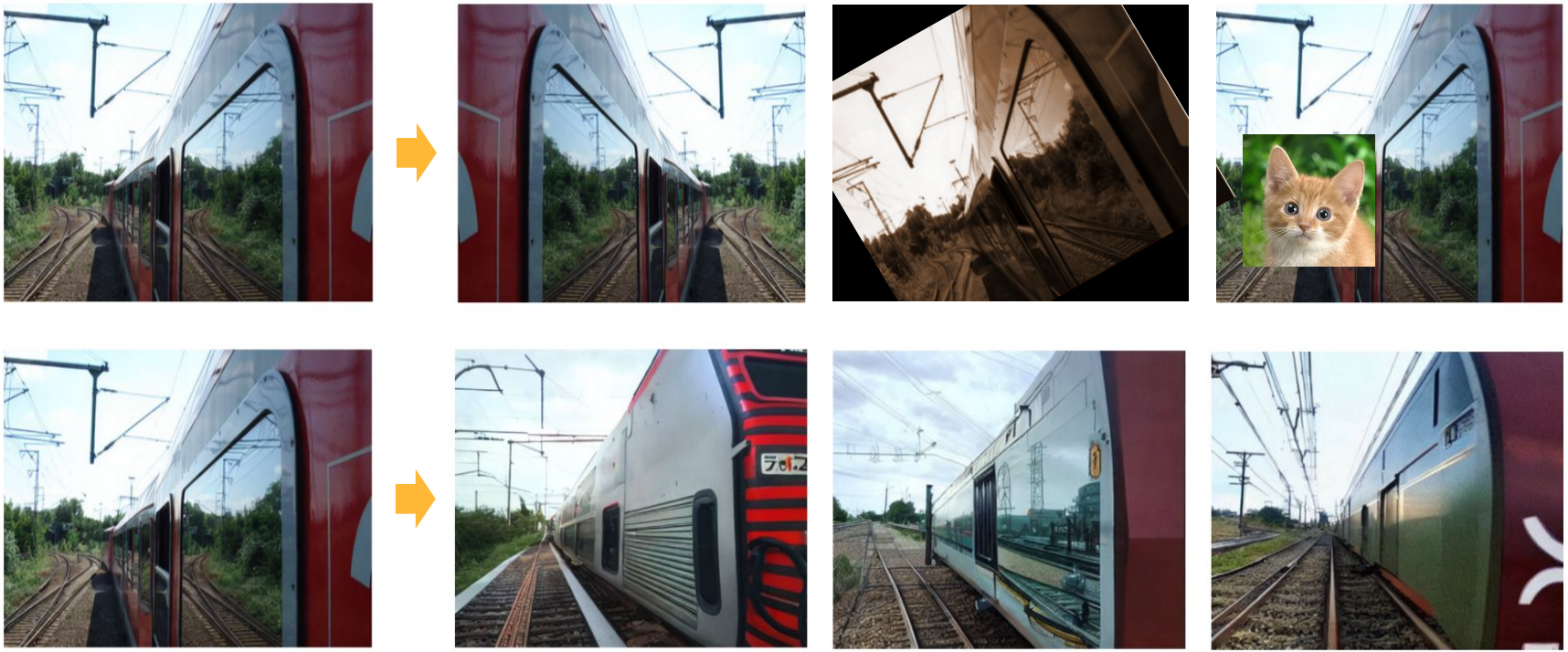
- Data augmentation often **requires a good intuition** about your dataset.

[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] Yun et al., CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, ICCV 2020.

We Need Augmentations That **Adapt To Your Dataset**

[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] Yun et al., CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, ICCV 2020.

We Need Augmentations That **Adapt To Your Dataset**



[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] Yun et al., CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, ICCV 2020.

Are Generative Models Capable Enough?

GAN, 2014 [3]

DCGAN, 2016 [4]

BigGAN, 2019 [5]

StableDiffusion, 2022 [6]

Nightmares before final exams **Works of art**

- Generative models have developed **astounding levels** of **photo-realism**.

[3] Goodfellow, Ian, et al., Generative Adversarial Networks, NeurIPS 2014.
[4] Radford, Alec, et al., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, ICLR 2016.
[5] Brock, Andrew, et al., Large Scale GAN Training for High Fidelity Natural Image Synthesis, ICLR 2019.
[6] Rombach, Robin, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

Can We Harness Image-Editing Techniques?

Why is this a good idea?

- **Data scarcity**: we can sample as many images as we need.
- **Semantics**: we can choose what to edit.

Key Idea: target and modify key semantic properties of images.

- Generations from a recent image-editing technique that respects high-level structure [7].

[7] Hertz, Amir, et al., Prompt-to-Prompt Image Editing with Cross Attention Control, arXiv 2022.

Generative Models Capture New Invariances

Rotate

Flip + Saturate

Standard Augmentation

Lemon → Apple

Lemon → Pistachio

"Data Augmentation" With Generative Models

- Generative models can target **semantic properties** of real images that we expect and desire models to be invariant to.

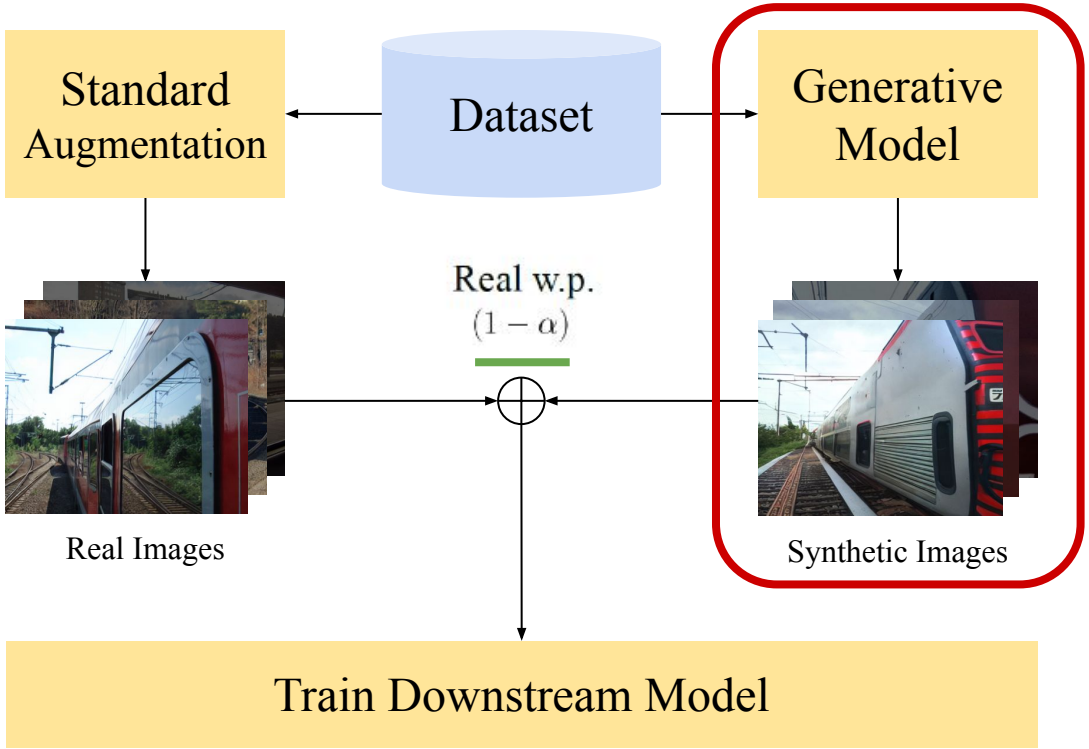
[7] Hertz, Amir, et al., Prompt-to-Prompt Image Editing with Cross Attention Control, arXiv 2022.

The Generative Data Augmentation Pipeline

- Mix real and generated images, weighted by an **augmentation probability**.

New Axes To Explore:

- What image-editing technique shall we use?
- How much structure do we preserve in the data?

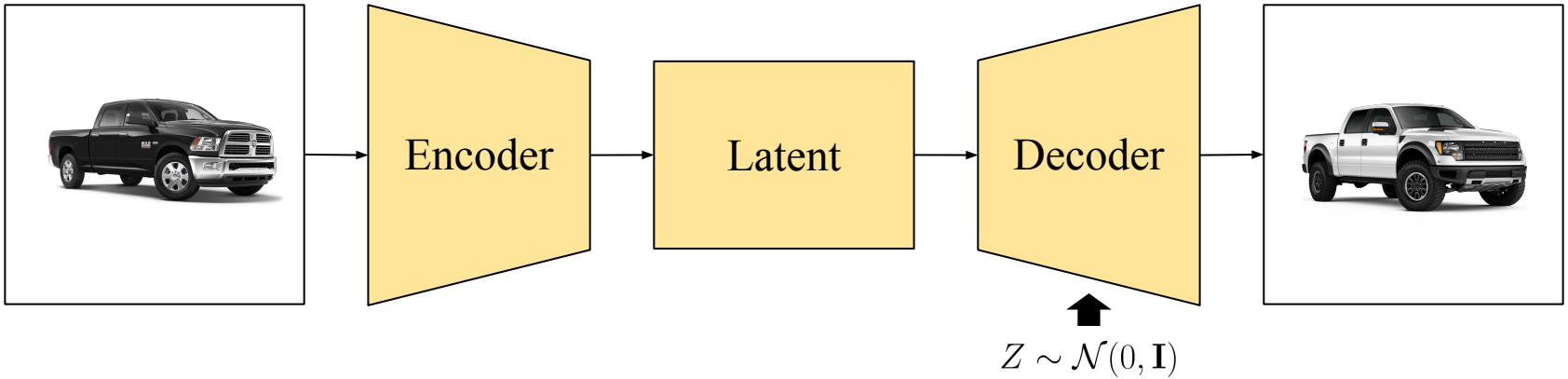


How Do We **Control** The Edited Property?

- Modern generative models have prompts, but this is a **recent development**.
- Before such models, researchers had explored methods to **infer the distribution of edits** purely from images.

How Do We **Control** The Edited Property?

- How can we **infer the distribution of edits** purely from images.
- **DAGAN: model this as an auto-encoding problem.**



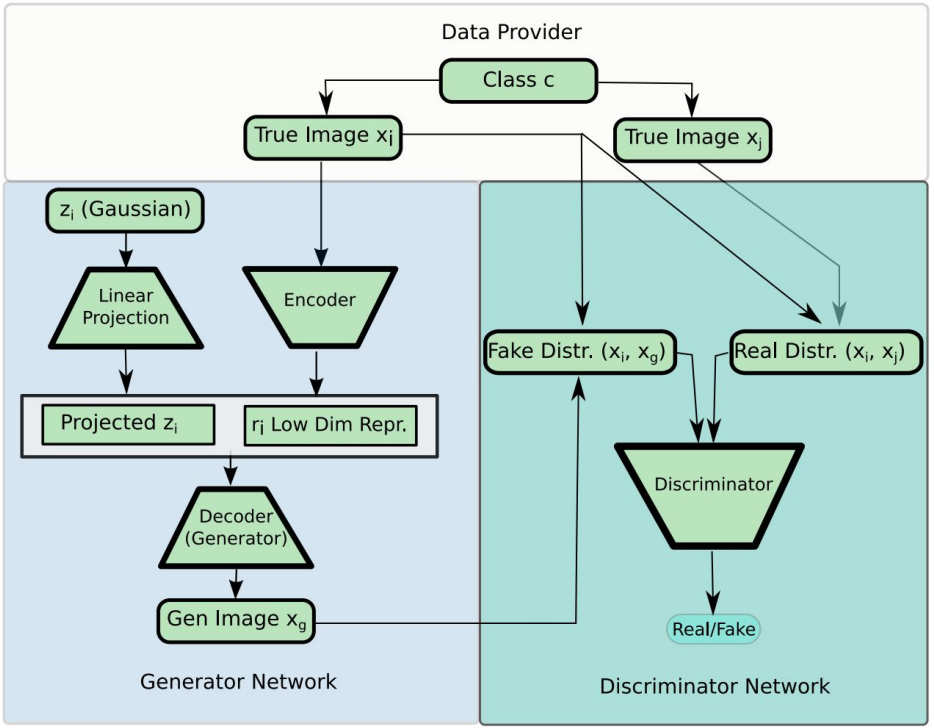
[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

Data Augmentation Generative Adversarial Networks

Modelling Task:

Generate an image from the same class as image Xi, *without observing the class*

- Generator implemented as a **UNet** mapping Xi to a generation Xg.
- Discriminator predicts if Xi, Xg are the same class.



[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

DAGAN: Example Generations & Results

Real

Fake

Face DAGAN Augmented Classification

Experiment ID	Samples Per Class	Test Accuracy
VGG-Face_Standard	5	0.0446948
VGG-Face_DAGAN_Augmented	5	**0.125969**
VGG-Face_Standard	15	0.39329
VGG-Face_DAGAN_Augmented	15	**0.429385**
VGG-Face_Standard	25	0.579942
VGG-Face_DAGAN_Augmented	25	**0.584666**

- Consistent improvement when tested on held-out classes.
- Diminishing gains when many samples per class are available.
- Generations are **inflexible**.

[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

How **Flexible** Is Data Augmentation GAN?

Rotate

Flip + Saturate

- **Controllable** extent and randomness.
- Highly modular.

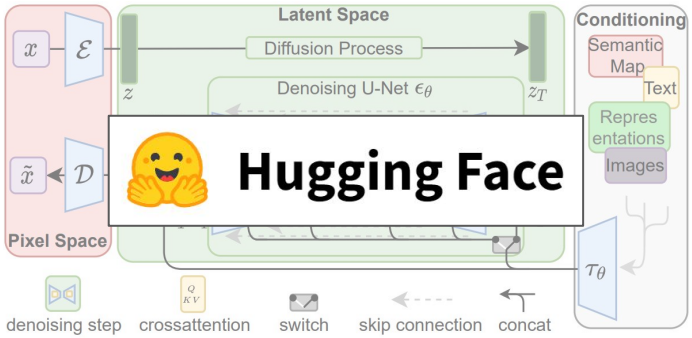
- Generates images **all at once** with little control over the layout and content.
- Requires **training** a GAN.

[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

DA-GAN Questions?

Can We **Avoid Training** A New Generative Model?

Don’t Have



~~Collect A Big Dataset~~

~~Train A Big Model [6]~~

Generate

- Let’s use **pre-trained** image generative models for data augmentation.
- Several powerful models: Imagen, GLIDE, **Stable Diffusion**.

[6] Rombach, Robin, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

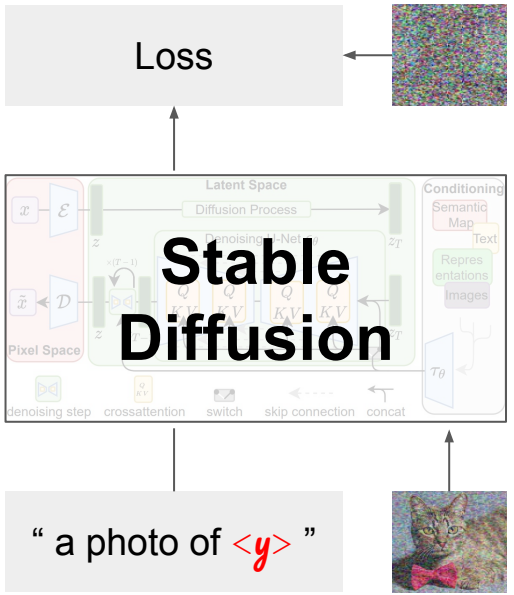
$$\max_{y} \log P(\ \text{  }\ |\ \text{a photo of a } y\ )$$

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

$$\max_{y} \log P(\ \text{  }\ |\ \text{a photo of a } y\ )$$

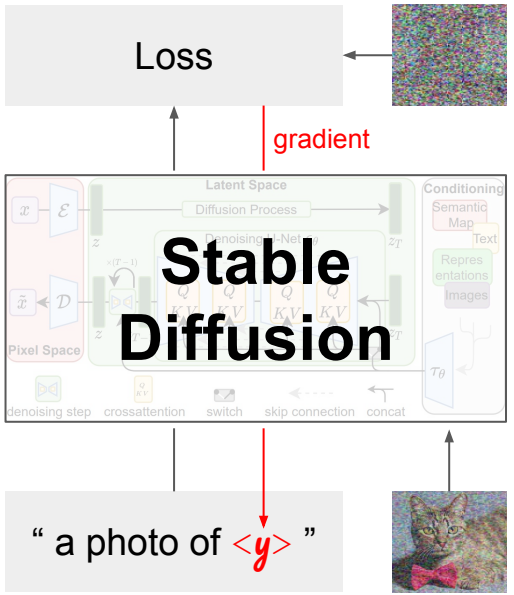


[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

$$\max_{y} \log P(\ \text{  }\ |\ \text{a photo of a } y\ )$$



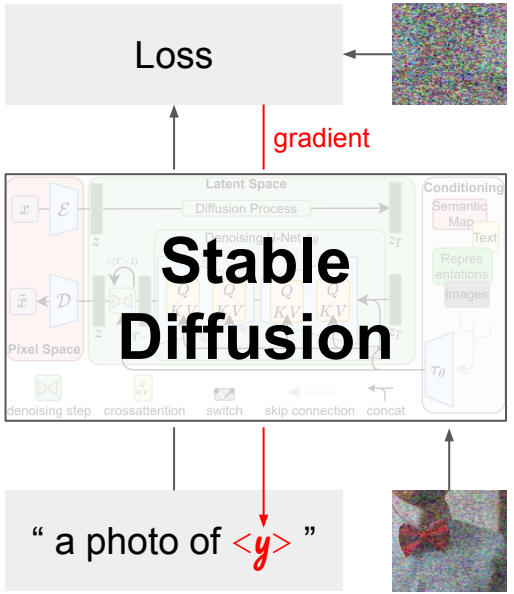
<y> ~ cat wearing a red bow-tie

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

$$\max_{y} \log P(\text{  } \mid \text{a photo of a } y \;)$$



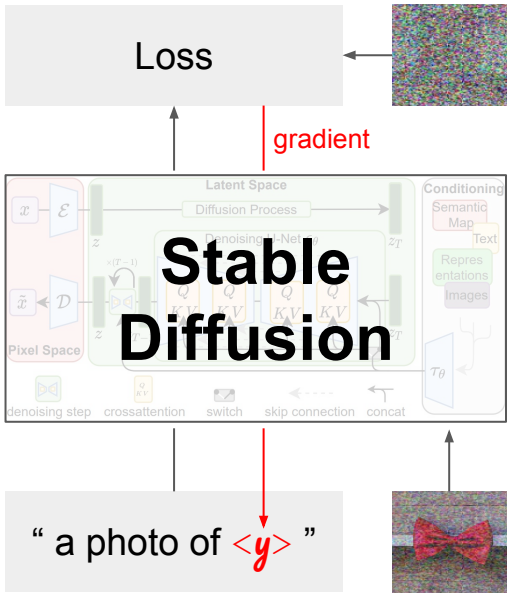
<y> ~ wearing a red bow-tie

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

$$\max_{y} \log P(\ \text{  }\ |\ \text{a photo of a } y\ )$$



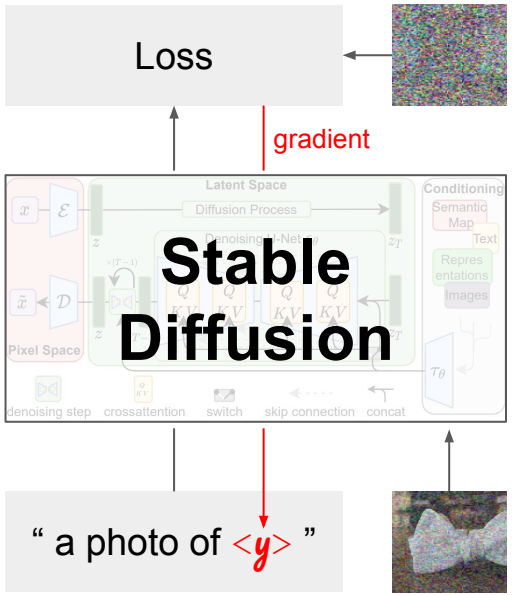
<y> ~ red bow-tie

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion

- **Key idea**: **shared context** in your images controls the augmentation.

$$\max_{y} \log P(\text{  } \mid \text{a photo of a } y \;)$$



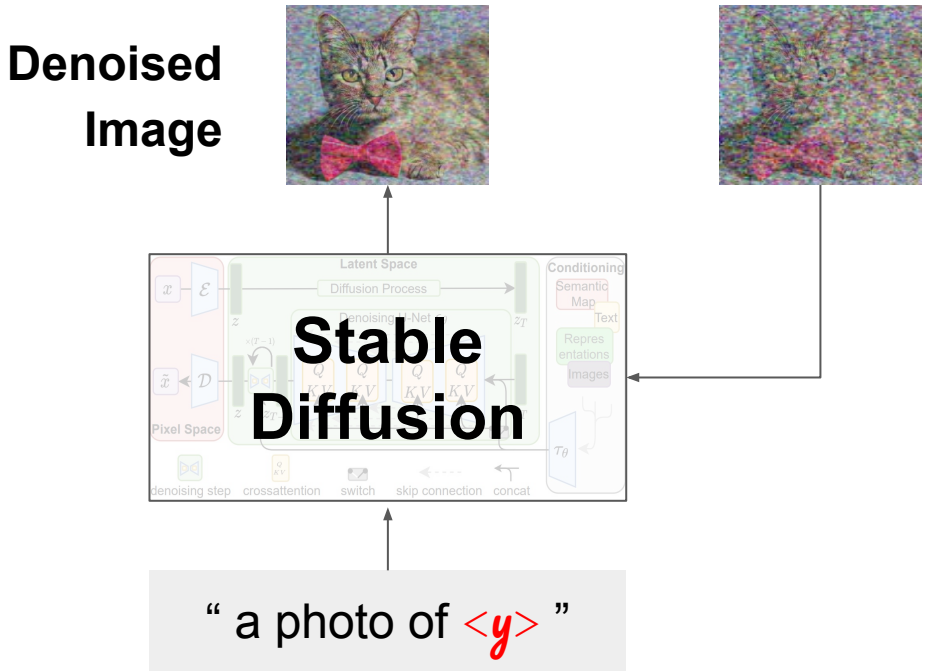
<y> ~ bow-tie

[3] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

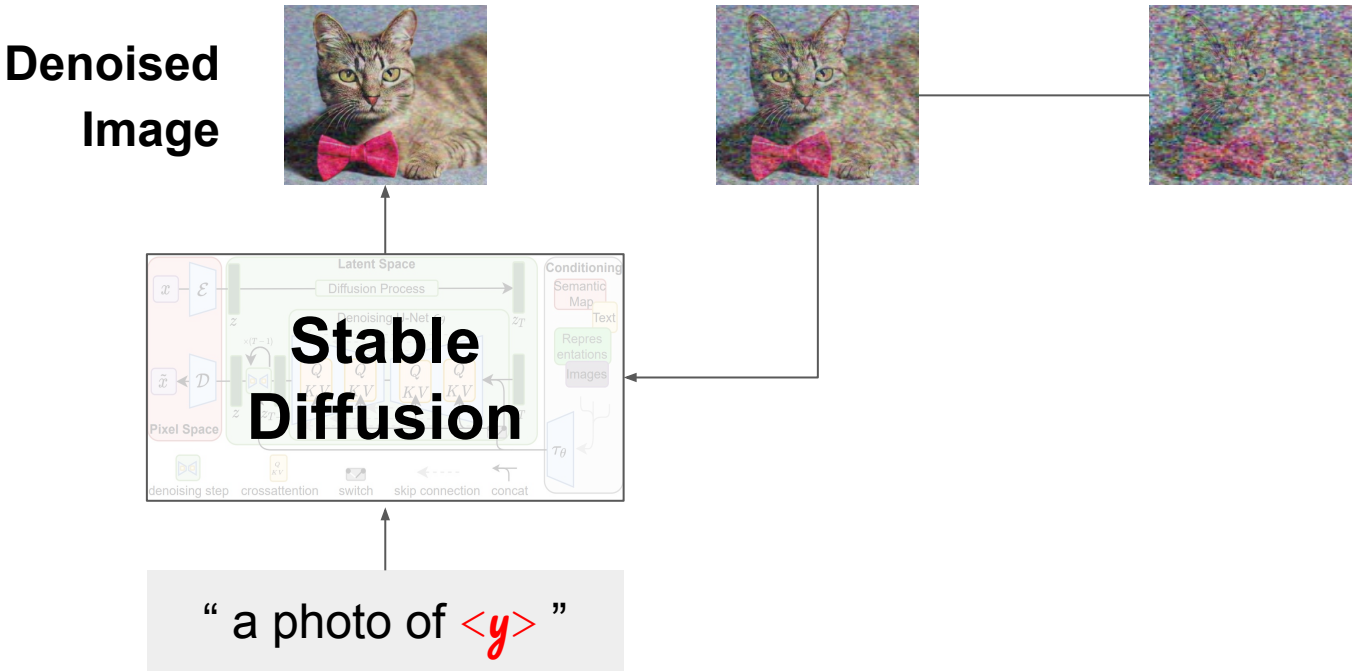
DA-Fusion: Data Augmentation With Diffusion

Source Image

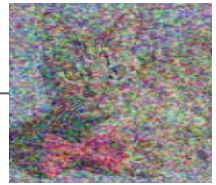
DA-Fusion: Data Augmentation With Diffusion

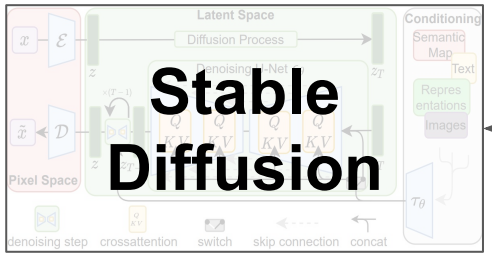


DA-Fusion: Data Augmentation With Diffusion



DA-Fusion: Data Augmentation With Diffusion

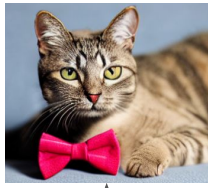


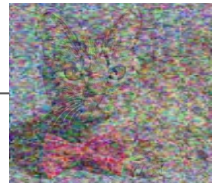


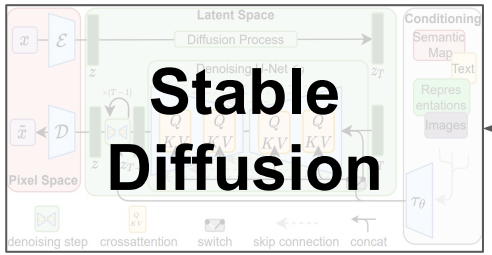
Stable Diffusion

“ a photo of <*y*> ”

DA-Fusion: Data Augmentation With Diffusion







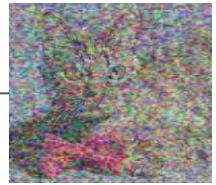
Stable Diffusion

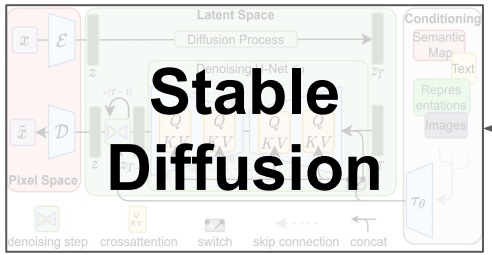
“ a photo of <y> ”

- Control the augmentation extent with the **insertion timestep**

Low

DA-Fusion: Data Augmentation With Diffusion





Stable Diffusion

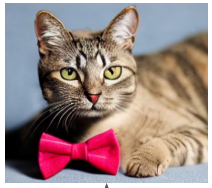
“ a photo of <y> ”

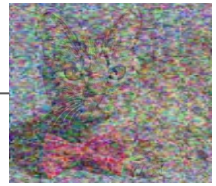
- Control the augmentation extent with the **insertion timestep**

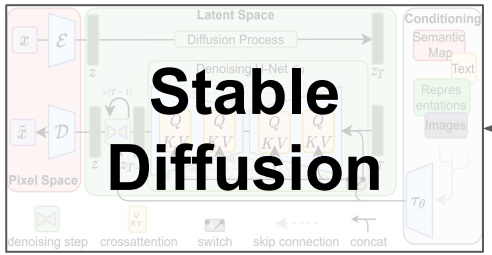
Low

Med

DA-Fusion: Data Augmentation With Diffusion







Stable Diffusion

“ a photo of <y> ”

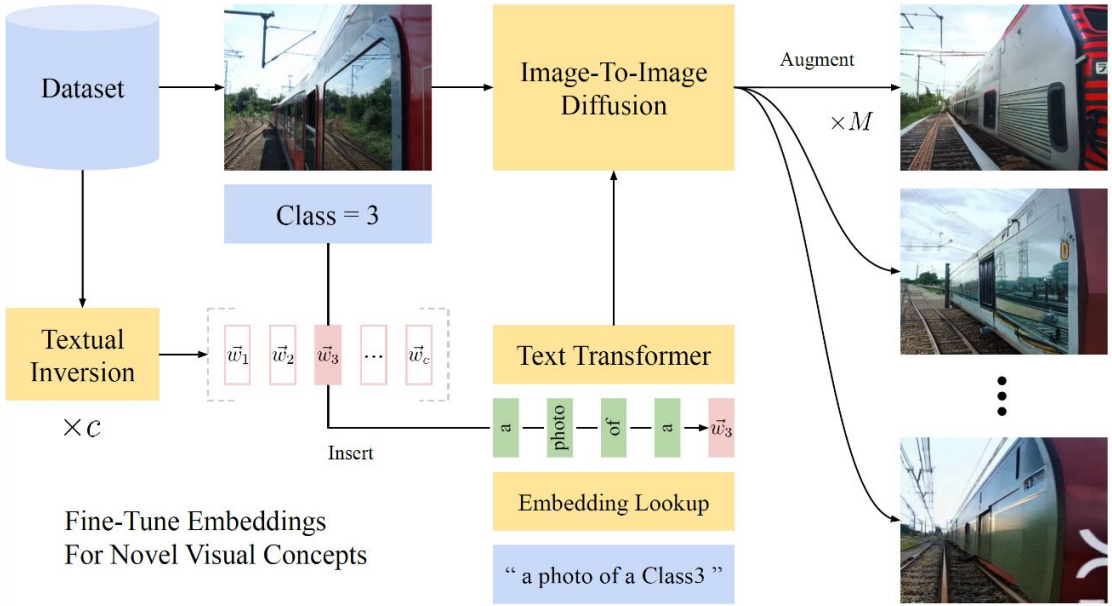
- Control the augmentation extent with the **insertion timestep**

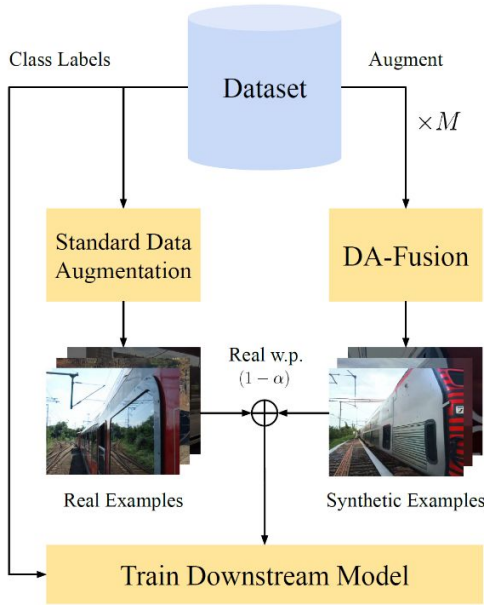
Low

Med

High

DA-Fusion: Data Augmentation With Diffusion

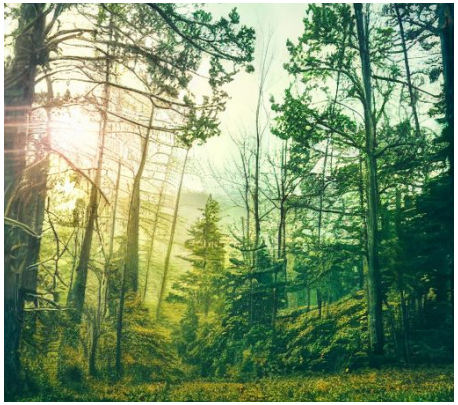




[7] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.
[8] Chenlin, Meng, et al., SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, ICLR 2022.








DA-Fusion Questions?

How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

- Six few-shot classification tasks from literature

Common Concepts

Fine-Grain Concepts

How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

- Six few-shot classification tasks from literature and **one we contribute**.

Common Concepts

Fine-Grain Concepts

Novel Concepts

How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

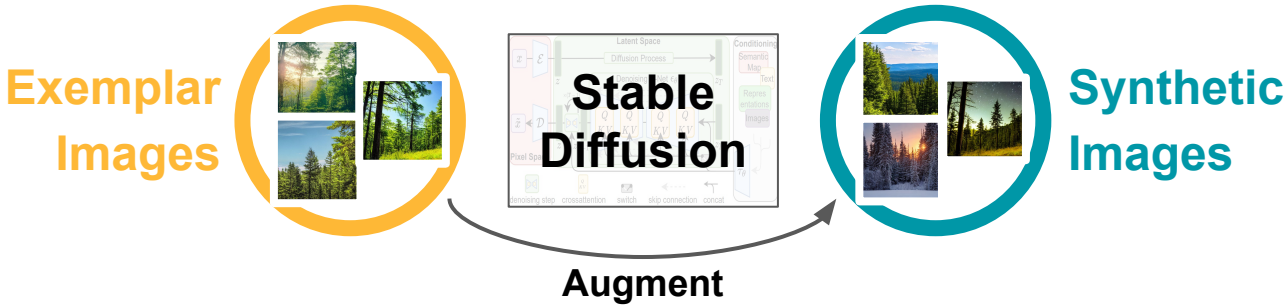
- Given a handful of **real images**

Exemplar Images

How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

- Given a handful of **real images**, generate **augmentations**

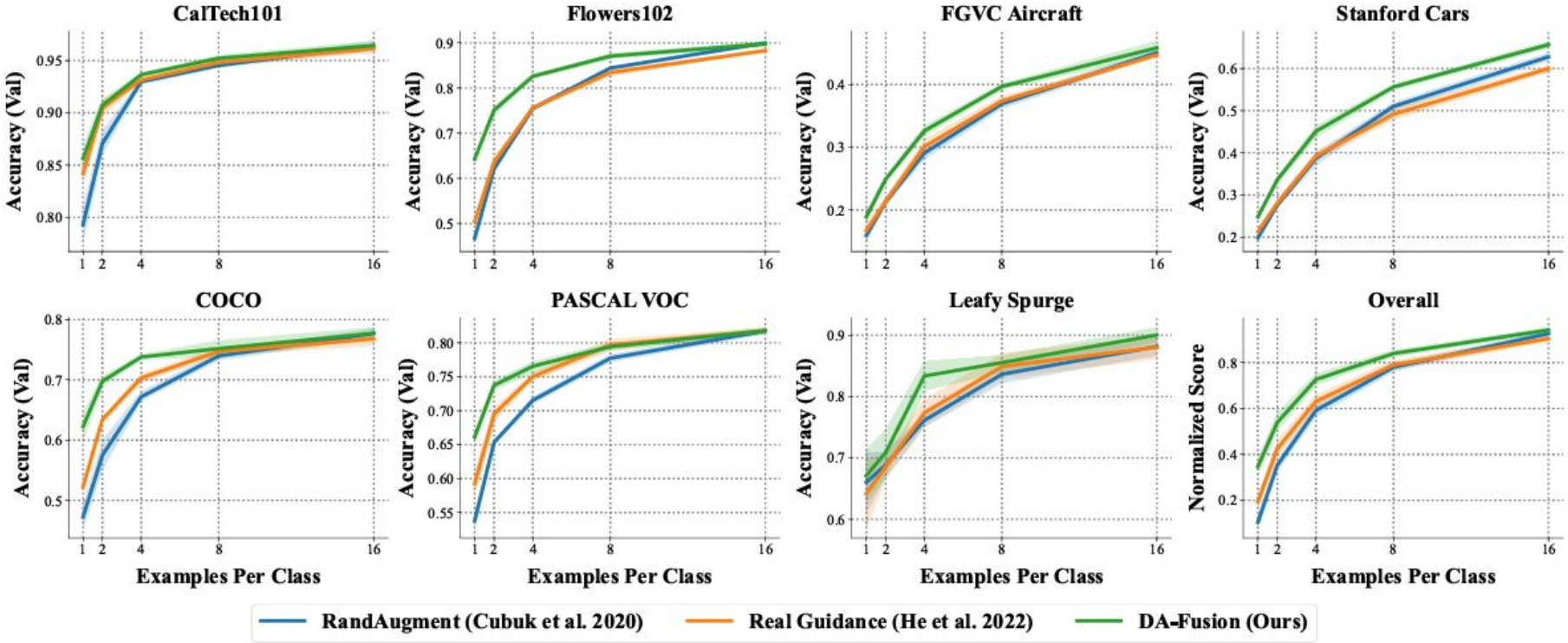


How Do We Evaluate DA-Fusion?

Question: How much do augmentations from DA-Fusion improve classification?

- Given a handful of **real images**, generate **augmentations**
- Train classifiers on a mix of real and synthetic data

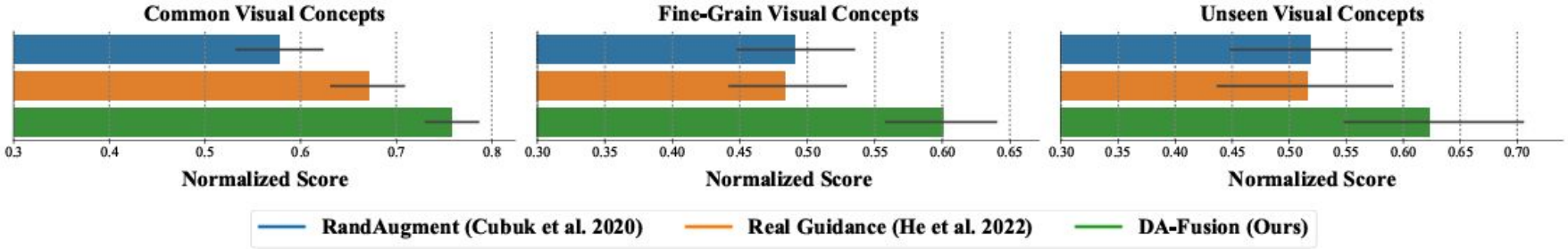
DA-Fusion Improves **Few-Shot Learning**



[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] He et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.

DA-Fusion Has **Consistent Performance**

Leafy Spurge



- DA-Fusion has **strong performance** for **all types of concepts**.

[1] Cubuk et al., RandAugment: Practical automated data augmentation with a reduced search space, NeurIPS 2020.
[2] He et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.

Strong Performance ✔

Strong Performance ✔

How Do You Control The Augmentation?

When Is Additional Control Necessary?

Real images are often cluttered with **distracting concepts**.

$$\max_{y} \log P(\quad | \text{ a photo of a } y \text{ })$$

When Is Additional Control Necessary?

Real images are often cluttered with **distracting concepts**.

$$\max_{y} \log P(\quad | \text{ a photo of a } y \text{)}$$

Which concept should DA-Fusion generate: cats and/or dogs?

How Do You Control What Concept Is Learned?

Implicit Solution: **select better images** without **distracting concepts**.

$$\max_{y} \log P(\quad | \text{ a photo of a } y \text{ })$$

How Do You Control What Concept Is Learned?

~~**Implicit Solution: select better images** without **distracting concepts**.~~

$$\max_{y} \log P(\quad | \text{ a photo of a } y \text{)}$$

This might be **costly**, what else can we do?

How Do You Control What Concept Is Learned?

Explicit Solution: **prompt with context** about the objects you want ignored.

How Do You Control What Concept Is Learned?

Explicit Solution: **prompt with context** about the objects you want ignored.

- **Why?** Prompts can supplement information **when images have ambiguity**.

How Do You Control What Concept Is Learned?

Explicit Solution: **prompt with context** about the objects you want ignored.

- **Why?** Prompts can supplement information **when images have ambiguity**.

DA-Fusion

How Do You Control What Concept Is Learned?

Explicit Solution: **prompt with context** about the objects you want ignored.

- **Why?** Prompts can supplement information **when images have ambiguity**.

DA-Fusion

\+ “ a <*y*> to the right of the dog ”

DA-Fusion

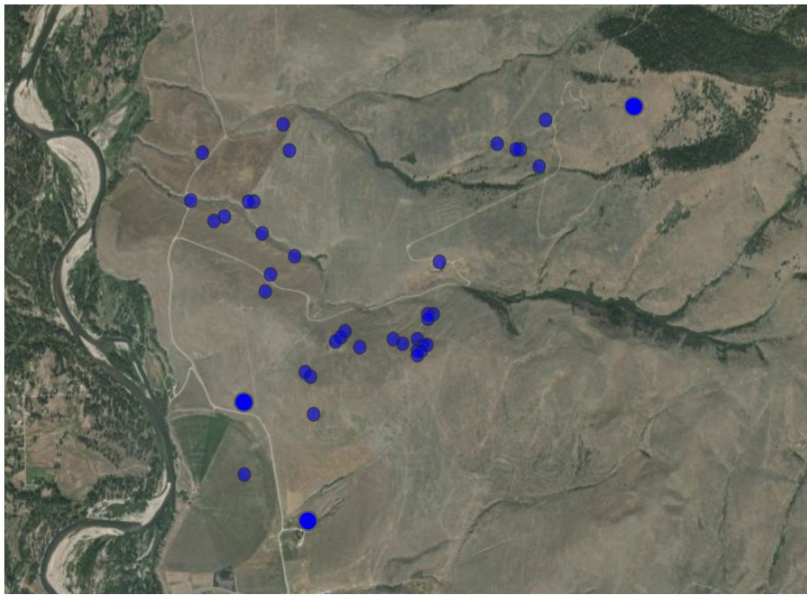
Results Questions?

Leafy spurge (Euphorbia esula): A Problematic Weed in N. America

Photo credit: Montana State University

Locations of spurge surveys in western Montana

- We surveyed 40 sites varied in land-use history and plant community composition



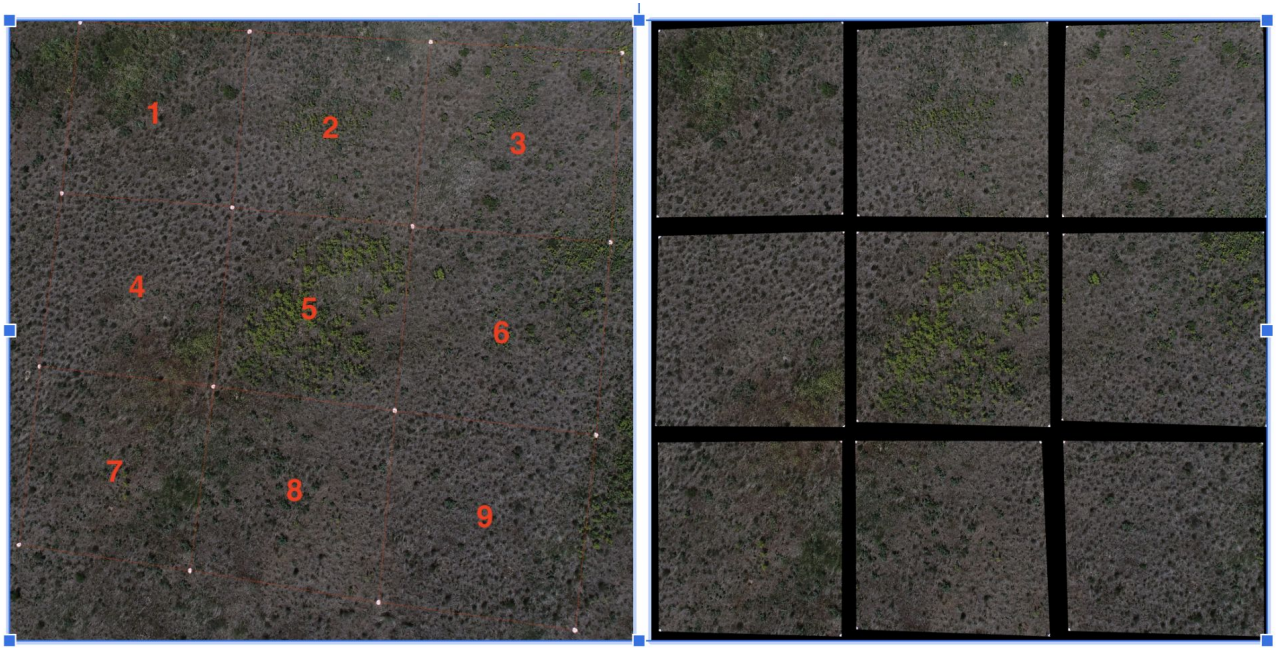
Botanists verified spurge presence at survey sites

- We searched for spurge within nine 10 x 10 meter plots

Drone then imaged surveyed areas

Post-processing: Identify plot boundaries

- Example: surveyed nine plots at each site. Markers visible in the image were used to crop plots.



Post-processing: Four-crop and verify

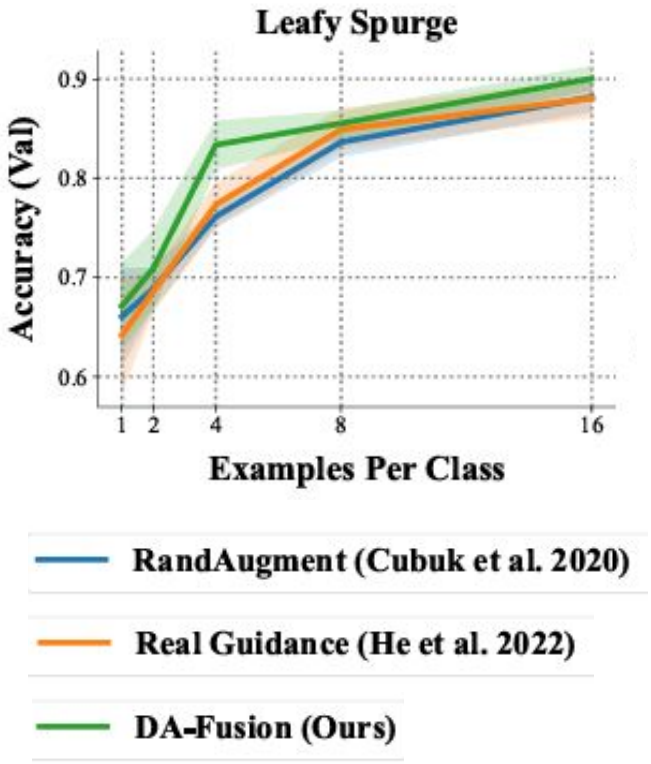
- We further subdivided imagery into quarters of 250 x 250 pixels in size (approximately 3.5 x 3.5 m).

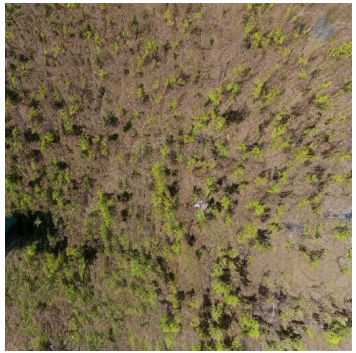
Aerial spurge images distinct from LAION dataset

- Top-down imagery of prairie ecosystems from 50m above the ground.
- Existing outside the domain of LAION and other foundation model training sets. ‘
- This makes these data well-suited for few-shot research.

LAION-5B Schuhmann et al. [2022]

Results + Synthetic Generations





original generation

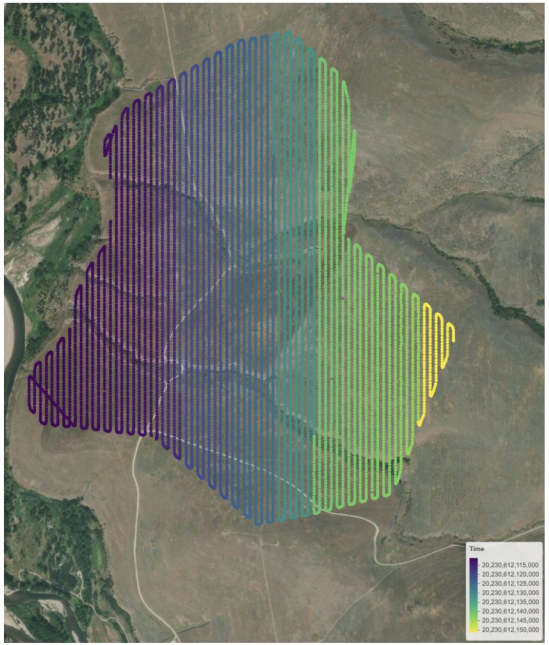
+ mountain

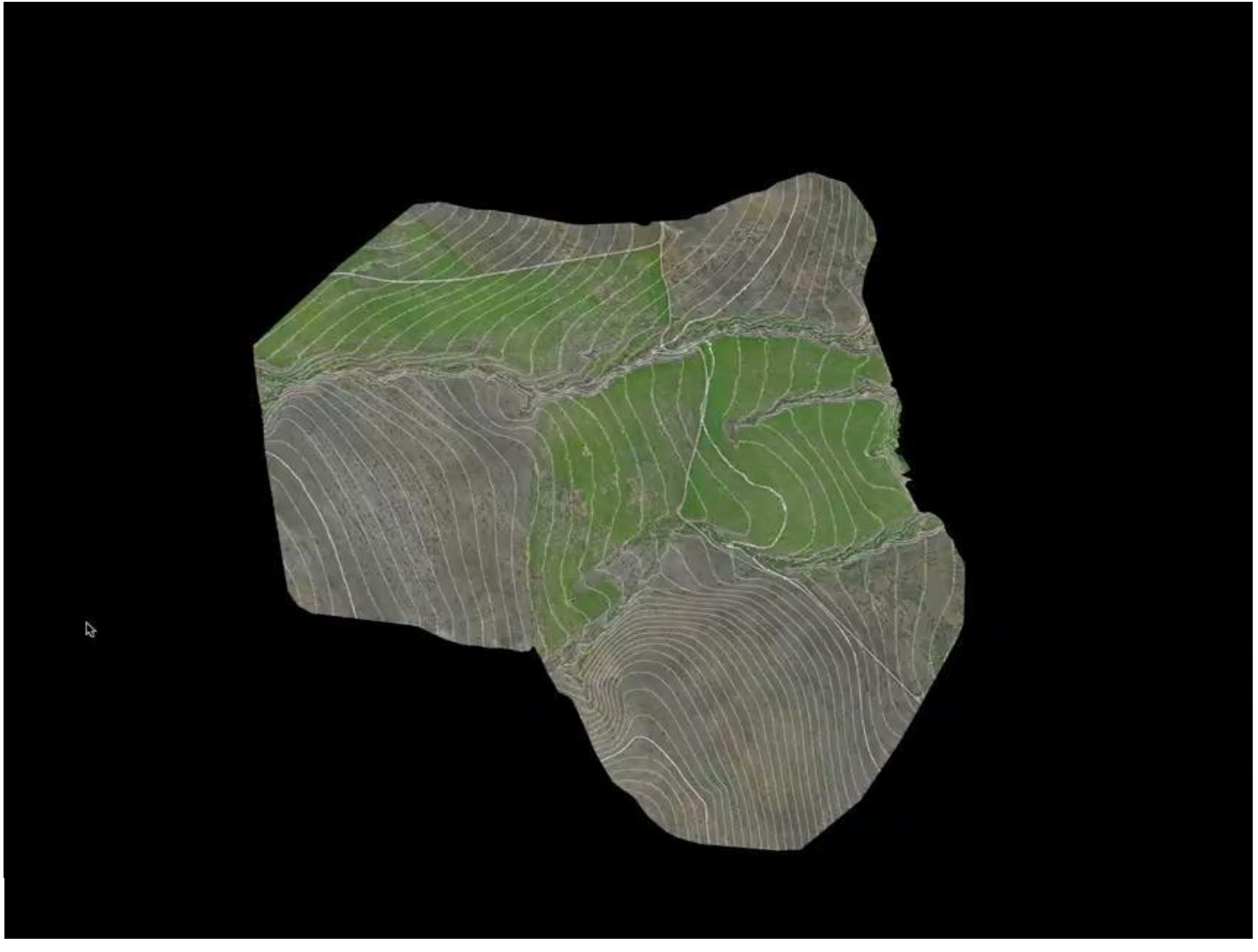
+ grassland

+ snow

Leafy Spurge Dataset V2

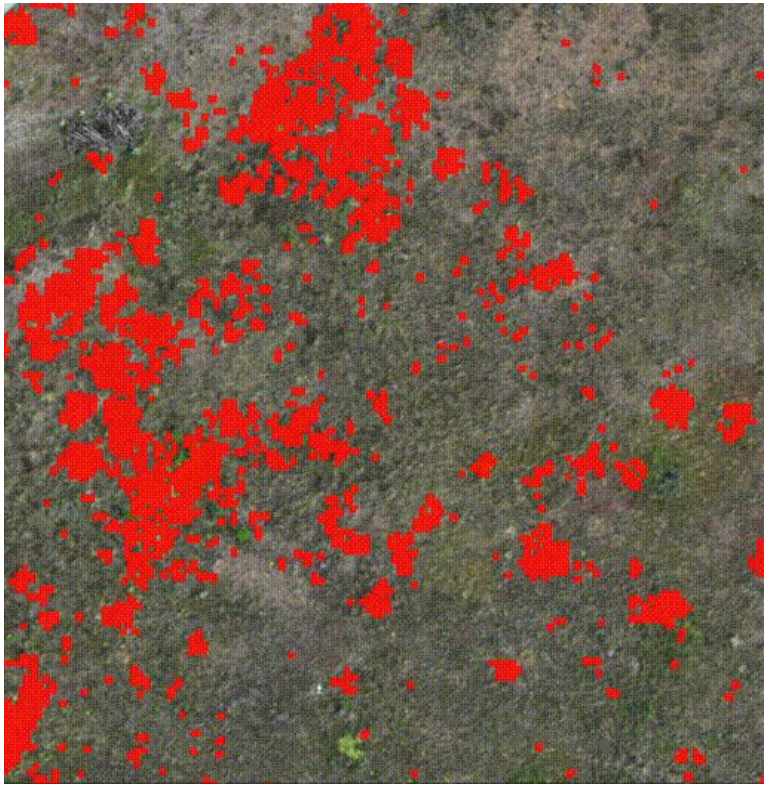
- We are working on the next version of the spurge dataset that is well-suited for mapping spurge presence at scale with a public release.



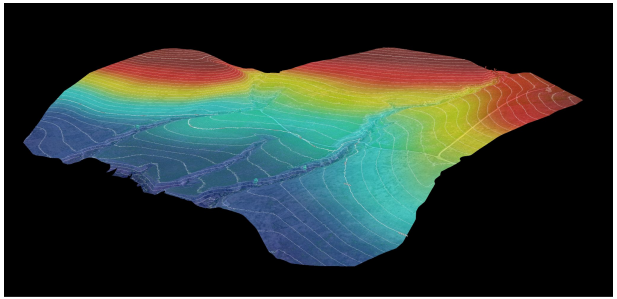


Fine-scale Classification Maps

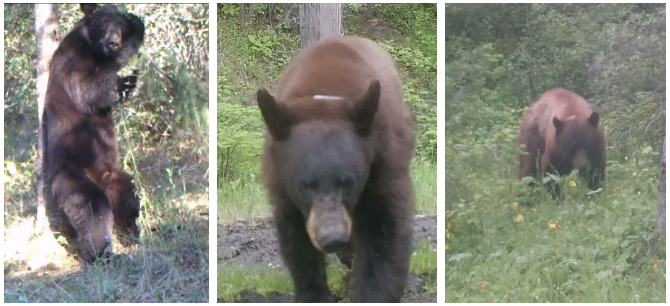
- Fine-scale classification will enable more effective management of leafy spurge and rapidly respond to these outbreaks.
- We hope to extend this approach to other species to monitor ecosystems at scale and rapidly respond to ecological change in near real-time.



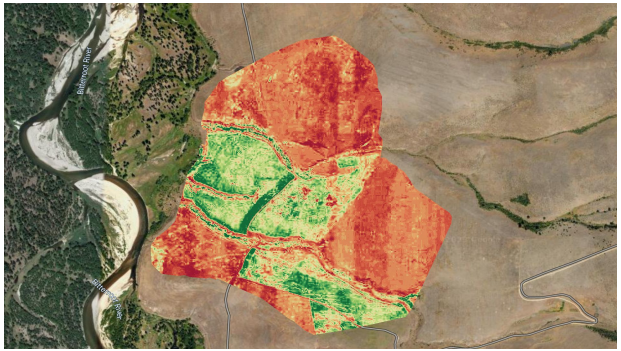
Ecologist Collaborators Offer a Rich Source of Unique Datasets



Elevation paired with RGB



Bear individual retrieval

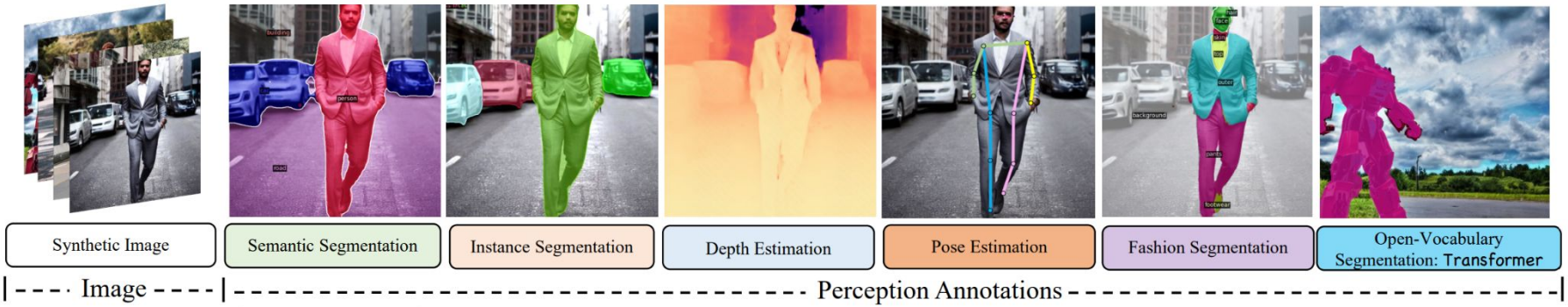


Non-visible spectra (plant health)

Landscape change detection

Leafy Spurge Questions?

Can We Also Obtain **Perceptual Annotations**?



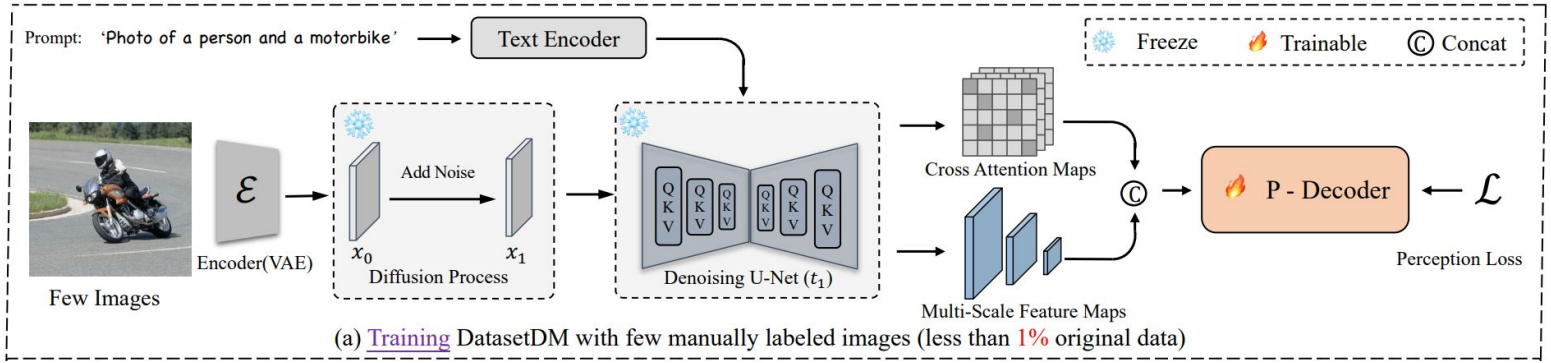
- In many data-limited settings, pixelwise labels matter, not just the class.
- Do generative models have an **inner representation** for these?
- And, can we **generate annotations** alongside the image?

[8] Weijia Wu, et al., DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models,NeurIPS 2023.

DatasetDM: Generating Images And Annotations

Key idea: the Diffusion Model has great features, just use these.

- Fine-tune a small **label decoder** on top of Diffusion Model features.

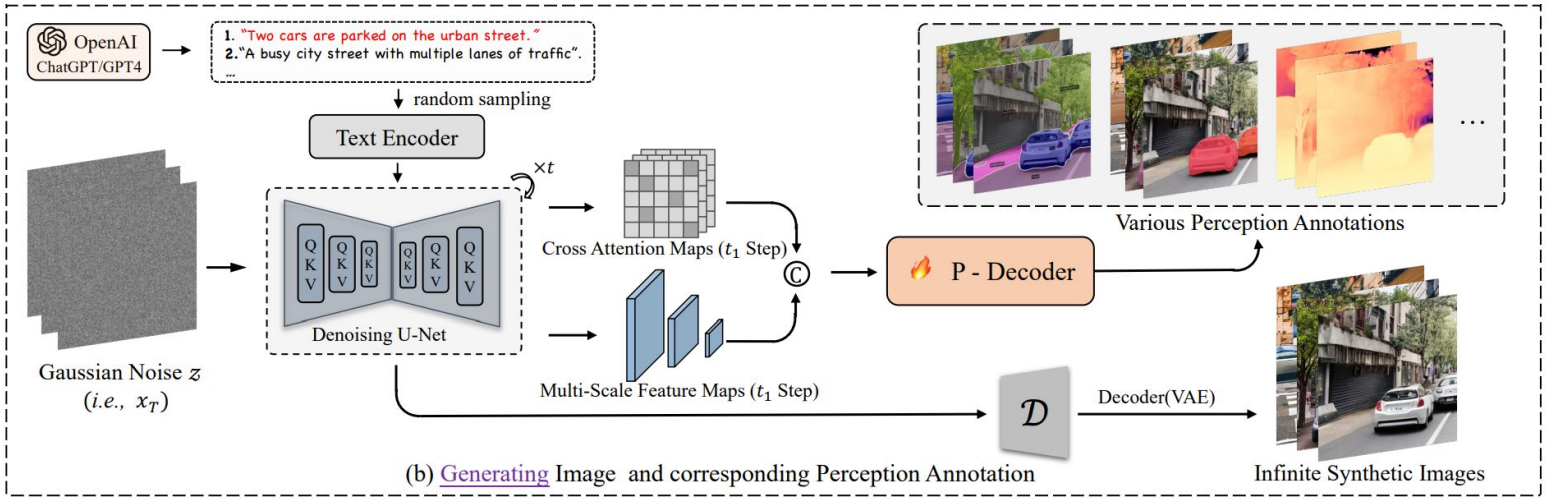


[8] Weijia Wu, et al., DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models,NeurIPS 2023.

DatasetDM: Generating Images And Annotations

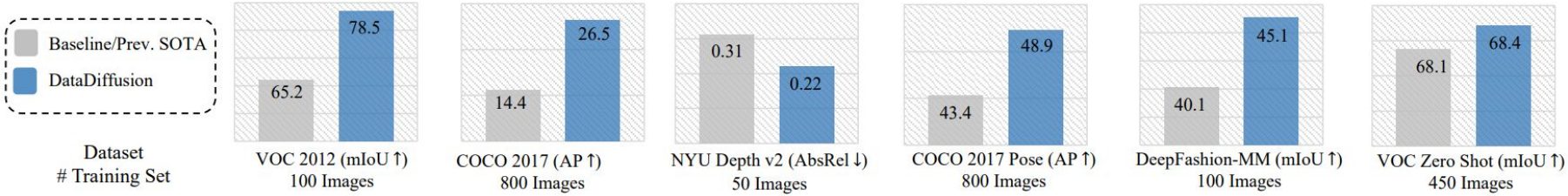
Key idea: the Diffusion Model has great features, just use these.

- During image generation, annotations are **very cheap to generate**.



[8] Weijia Wu, et al., DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models,NeurIPS 2023.

DatasetDM: Generating Images And Annotations



- Very efficient, can generate annotations from < 1000 example images.
- Synthetic data improves vision models in data-limited settings.

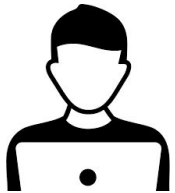
[8] Weijia Wu, et al., DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models,NeurIPS 2023.

DatasetDM Questions?

From Language Models To **Agent Models**

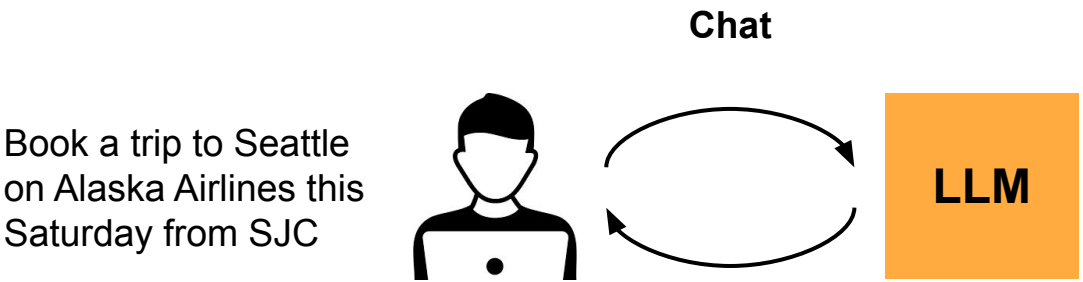
- A new paradigm is emerging using **Large Language Models** to automate **computer tasks** like Web Browsing, Software Dev, and more.

Book a trip to Seattle on Alaska Airlines this Saturday from SJC



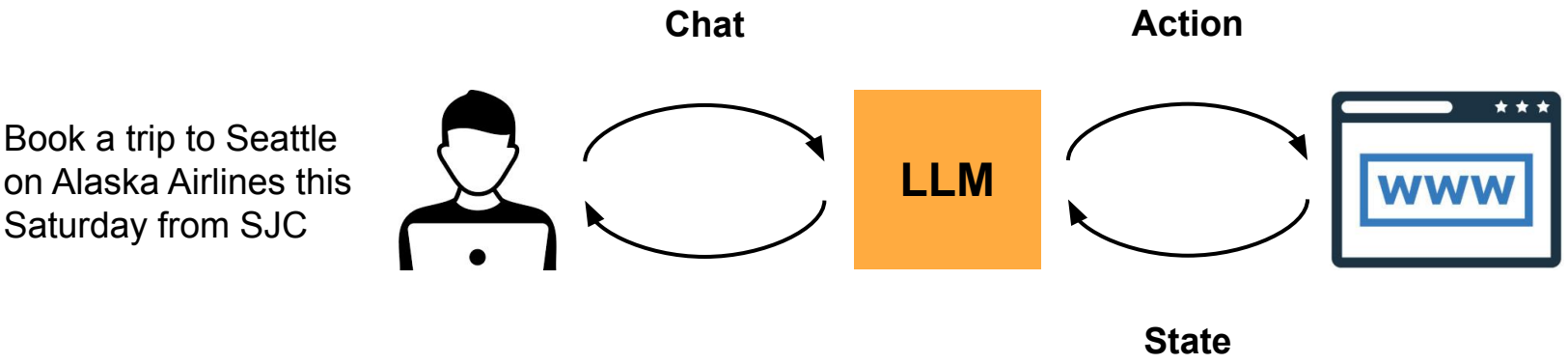
From Language Models To **Agent Models**

- A new paradigm is emerging using **Large Language Models** to automate **computer tasks** like Web Browsing, Software Dev, and more.



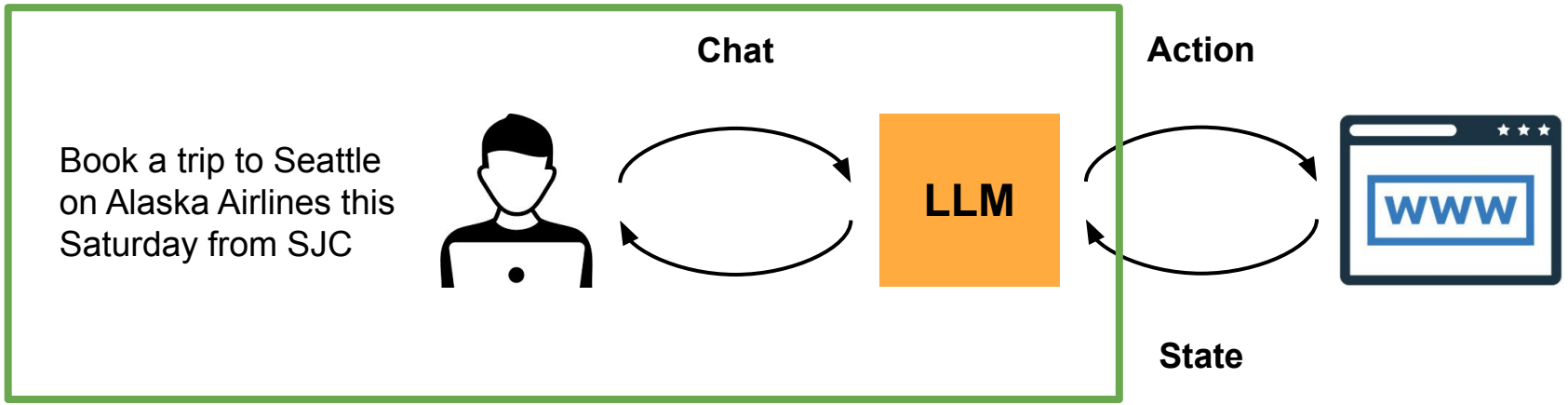
From Language Models To **Agent Models**

- A new paradigm is emerging using **Large Language Models** to automate **computer tasks** like Web Browsing, Software Dev, and more.



From Language Models To **Agent Models**

- A new paradigm is emerging using **Large Language Models** to automate **computer tasks** like Web Browsing, Software Dev, and more.



Our Goal: **Unlock A Critical Data Flywheel**

Task: Find a digitized historical map of the San Francisco Bay Area.


SCORE $100 OFF FOR EVERY $1000 YOU SPEND! Use Code SCORE At Checkout!
Choose from 40+ 4WP stores
Need help finding parts? Question on fitment? We can help! Give us a call 877-474-4821 or send us a message
Search...
FIND A LOCATION
SUPPORT
CART
STORES LOCATIONS
SHOP BY CATEGORY
SHOP BY BRAND
JOIN US!
SATURDAY MARCH 8TH
MADNESS IN MARCH
ALL LOCATIONS CELEBRATING
- SAVE STORE-WIDE
- OFF-ROAD VEHICLES ON DISPLAY
- WIN FREE RAFFLE PRIZES
- FREE FOOD & DRINKS
- MUSIC, GAMES & FAMILY FUN
- DOG & FAMILY FRIENDLY
9AM - 3PM AT ALL NATIONWIDE LOCATIONS. 3RD - 10TH STORE-WIDE DISCOUNTS.
CALL OR VISIT YOUR NEAREST STORE FOR MORE DETAILS.
SELECT VEHICLE
Select Year
SHOP BY CATEGORY
Select Main Category
SHOP BY BRAND
Select Brand
SHOP TOP CATEGORIES
Explore promotions
Task: Find the price of a Jeep Wrangler lift kit.

Towards Internet-Scale Training For Agents

- **Goal: unlock internet-scale data for training for agent models.**

Towards Internet-Scale Training For Agents

- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

Towards Internet-Scale Training For Agents

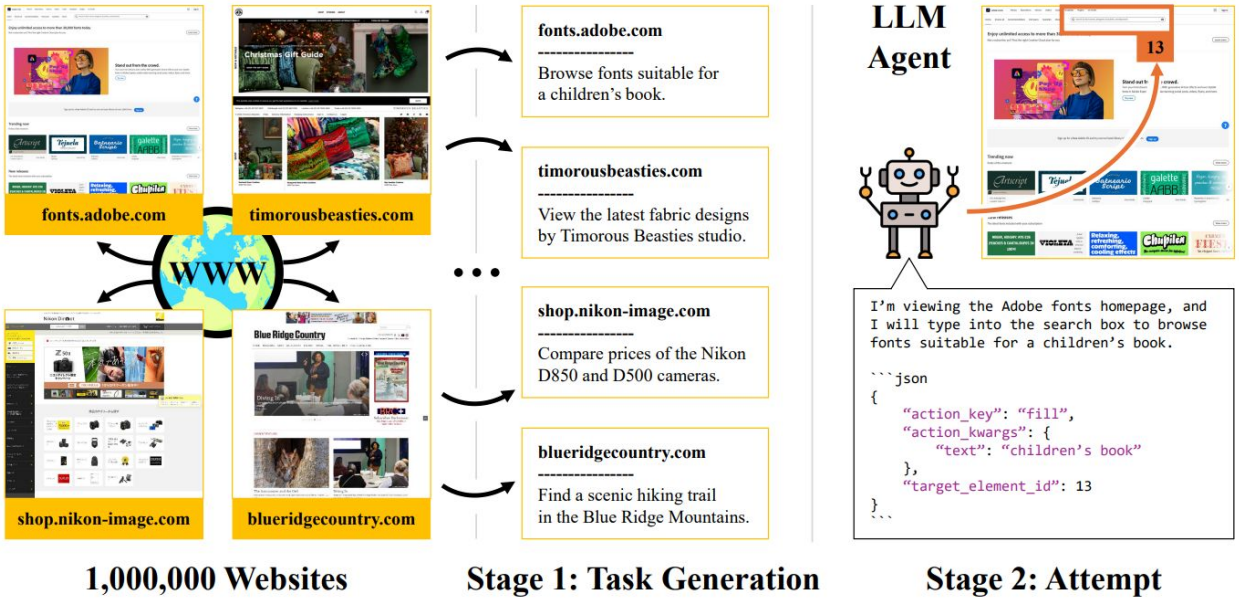
- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

Towards Internet-Scale Training For Agents

- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

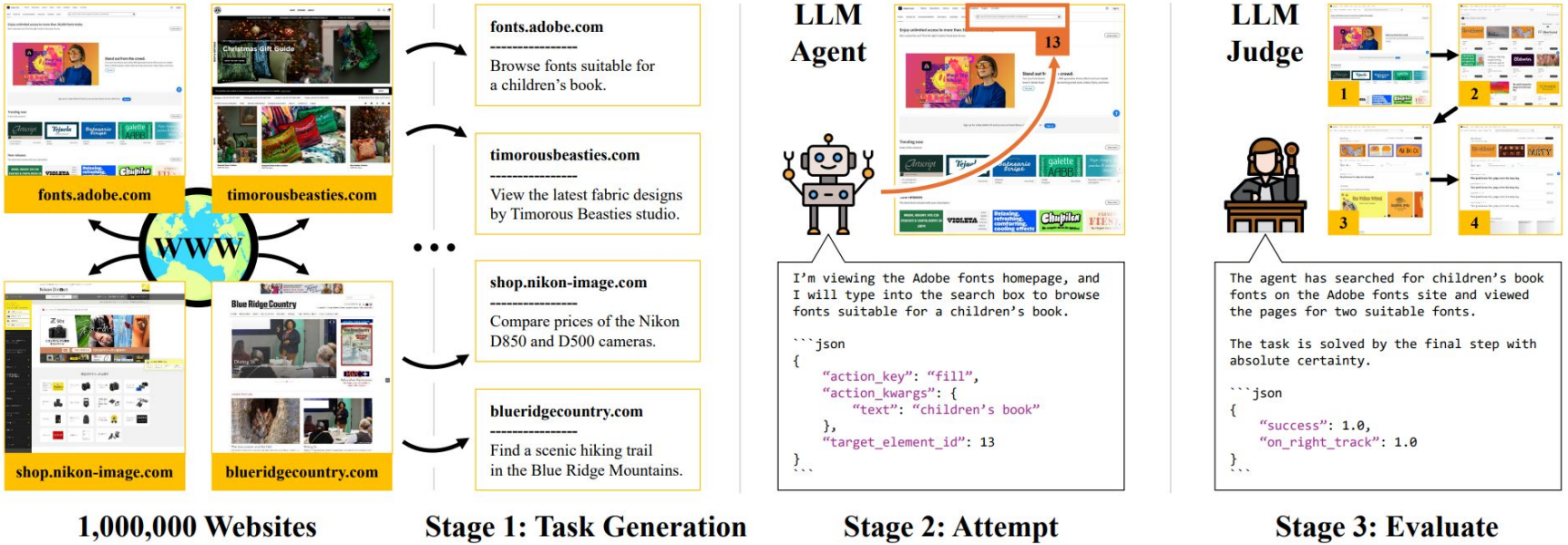
Towards Internet-Scale Training For Agents

- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.



Towards Internet-Scale Training For Agents

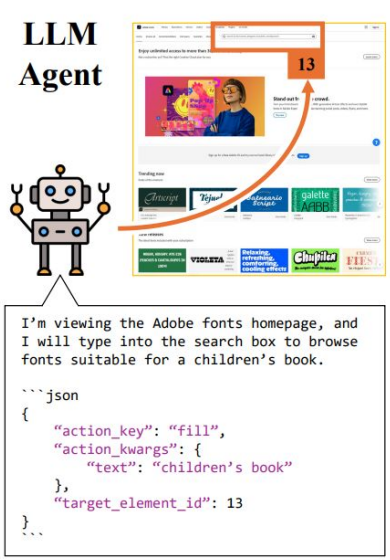
- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

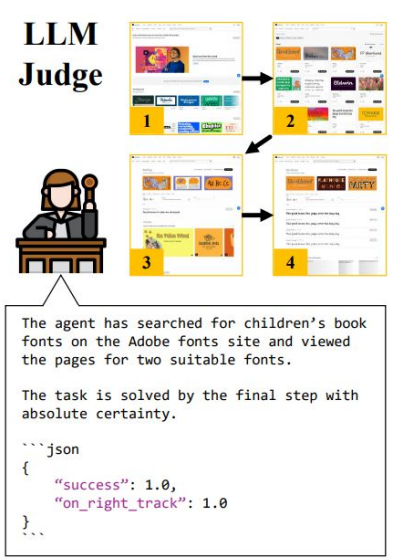


Towards Internet-Scale Training For Agents

- **Goal: unlock internet-scale data for training for agent models.**
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

Let's start here





LLMs Can Generate **Realistic Agent Tasks**

The Problem Statement:

Given a **web domain as text** (i.e. `www.uefi.org`).

Propose a realistic task that **an average user could complete** in one session.

LLMs Can Generate **Realistic Agent Tasks**

The Problem Statement:

Given a **web domain as text** (i.e. `www.uefi.org`).

Propose a realistic task that **an average user could complete** in one session.

This is simple.

LLMs Can Generate **Realistic Agent Tasks**

The Problem Statement:

Given a **web domain as text** (i.e. `www.uefi.org`).

Propose a realistic task that **an average user could complete** in one session.

This is simple. Why does it work so well?

LLMs Can Generate **Realistic Agent Tasks**

The Problem Statement:

Given a **web domain as text** (i.e. `www.uefi.org`).

Propose a realistic task that **an average user could complete** in one session.

This is simple. Why does it work so well?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Why Is A Simple Approach **So Effective**?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Why Is A Simple Approach **So Effective**?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Mersey Ferries (merseyferries.co.uk) is the **official website for the famous Mersey Ferry service** ?

Why Is A Simple Approach **So Effective**?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Mersey Ferries (merseyferries.co.uk) is the **official website for the famous Mersey Ferry service** ✅

Why Is A Simple Approach **So Effective**?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Mersey Ferries (merseyferries.co.uk) is the **official website for the famous Mersey Ferry service** in Liverpool, England. This website serves as a one-stop platform for users to access various services and information related to the ferry operation. Here are some things a user can do on this site: ✅

Why Is A Simple Approach **So Effective**?

Llama was pretrained on **broad data that contains references to web domains**, and has **broad prior knowledge** about websites and their uses.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Mersey Ferries (merseyferries.co.uk) is the **official website for the famous Mersey Ferry service** ✅ in Liverpool, England. This website serves as a one-stop platform for users to access various services and information related to the ferry operation. Here are some things a user can do on this site:

1. **Plan a Journey**: Users can plan their journey by checking the ferry timetable, which includes departure and arrival times for both commuter and cruise services.
2. **Book Tickets**: Users can book tickets online for both commuter and cruise services, including the popular Manchester Ship Canal Cruise and the Liverpool Bay Cruise. …

LLMs Are Competitive With **Human Annotators**

We compare LLMs to (non-expert) human annotators at **detecting harmful sites**, and creating **feasible agent tasks**, for a set of 100 sites.

LLMs Are Competitive With **Human Annotators**

We compare LLMs to (non-expert) human annotators at **detecting harmful sites**, and creating **feasible agent tasks**, for a set of 100 sites.

Method	Acc.	Prec.	Recall
Llama 3.1 70B	85%	0.77	**1.00**
GPT-4o	95%	0.91	**1.00**
Gemini 1.5 Pro	**97%**	**0.96**	0.98
Human Baseline	75%	0.71	0.84

Accuracy For Detecting Harmful Sites

Method	Feasibility Rate
Llama 3.1 70B	75%
GPT-4o	85%
Gemini 1.5 Pro	**89%**
Human Baseline	54%

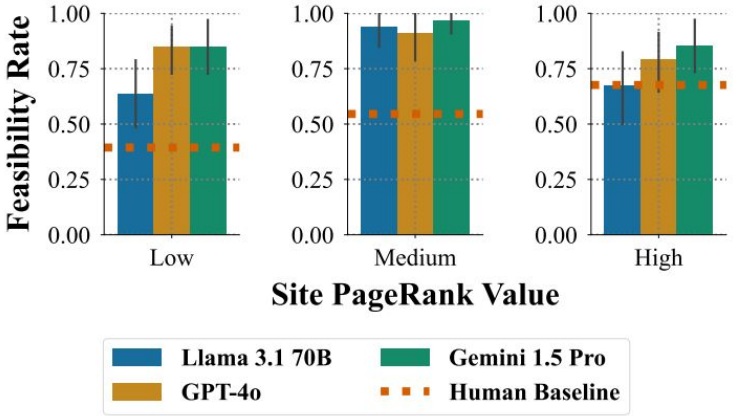
Expert Feasibility Of Proposed Tasks

LLMs Are Competitive With **Human Annotators**

Human annotators become **less reliable** as sites become **less popular**, and **LLMs perform consistently well** for diverse sites.

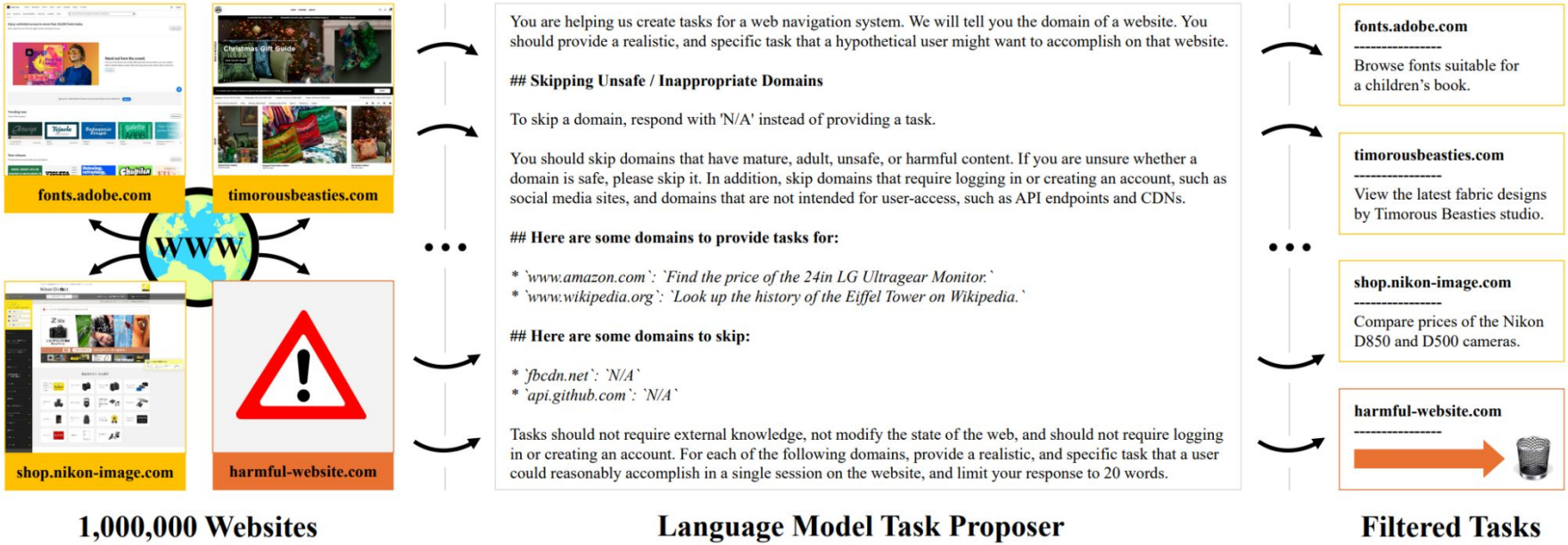
Method	Acc.	Prec.	Recall
Llama 3.1 70B	85%	0.77	**1.00**
GPT-4o	95%	0.91	**1.00**
Gemini 1.5 Pro	**97%**	**0.96**	0.98
Human Baseline	75%	0.71	0.84

Accuracy For Detecting Harmful Sites



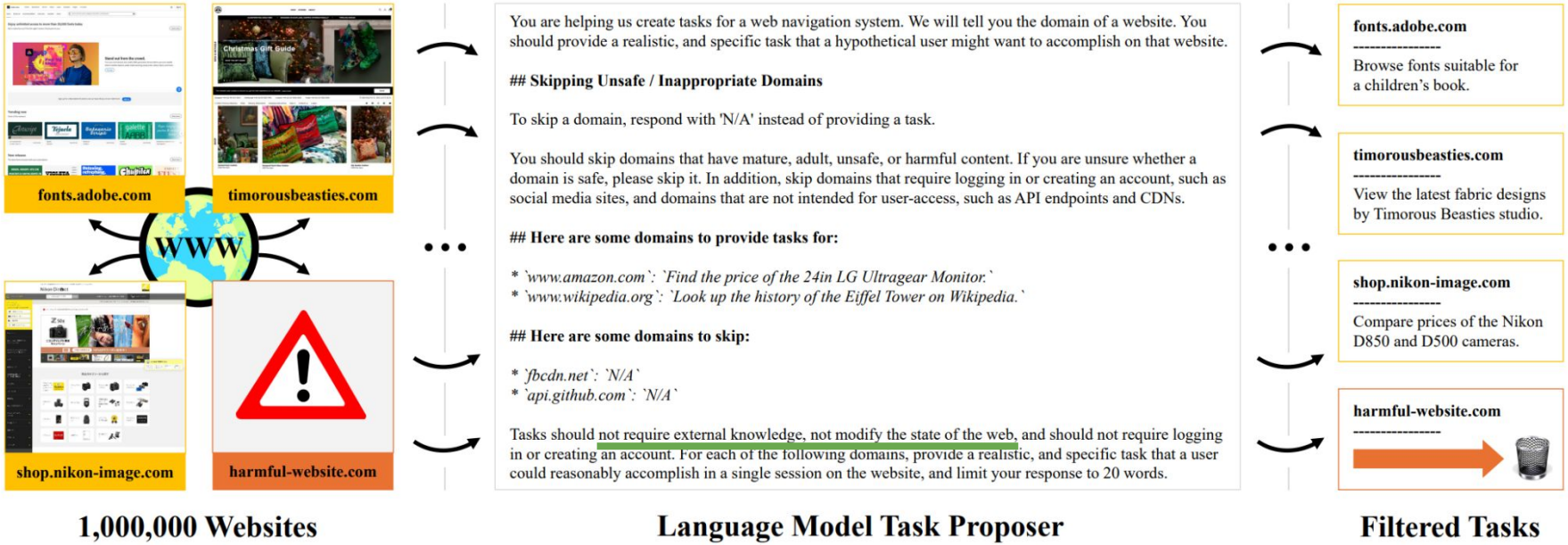
Scaling To **150,000 Sites** Using Llama

- Starting from the **top 1M websites**



Scaling To **150,000 Sites** Using Llama

- Starting from the **top 1M websites**, we generate **150k safe tasks**.



Exploring Tasks Generated By InSTA

- Tasks are **diverse**, and many require **multiple steps of reasoning** (i.e., what makes a font suitable for a children's book).

Domain	Task
wordpress.org	Find a free and popular theme for a personal blog.
policies.google.com	Read Google's terms of service for using YouTube.
ec.europa.eu	Retrieve a report on the EU's climate change policy.
vimeo.com	Find a short film on environmental conservation.
fonts.adobe.com	Browse fonts suitable for a children's book.
apps.apple.com	Find the top-rated free productivity app for iPhone.

Exploring Tasks Generated By InSTA

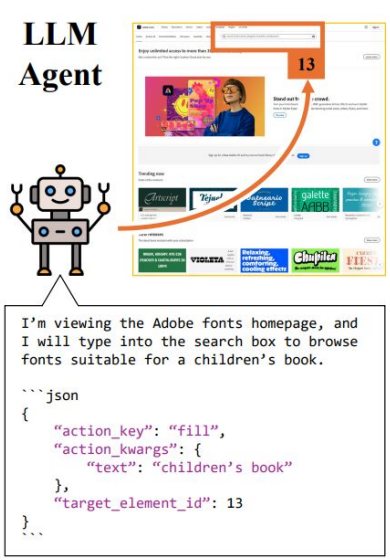
- There is an **emergent creativity** where Llama can recall facts that a site is likely to contain, such as the meaning of the Om symbol (ॐ).

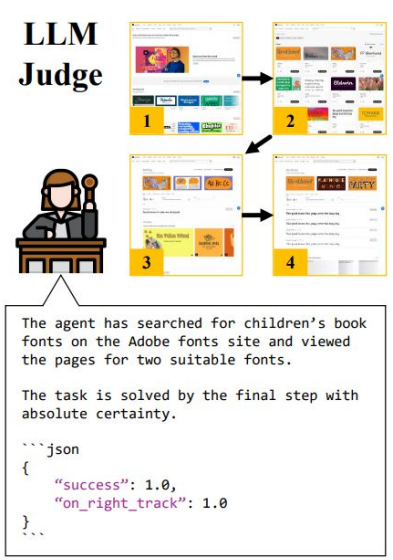
Domain	Task
ancient-symbols.com	Look up the meaning of the Om symbol in ancient cultures.
petsforhomes.com.au	Find a list of available dogs for adoption in New South Wales.
timorousbeasties.com	View the latest fabric designs by the Timorous Beasties studio.
shop.nikon-image.com	Compare prices of the Nikon D850 and D500 cameras.
blueridgecountry.com	Find a scenic hiking trail in the Blue Ridge Mountains.
awg-fittings.com	Find the dimensions of a 1/2\" NPT fitting.

Task Generation Questions

Revisiting The InSTA Pipeline

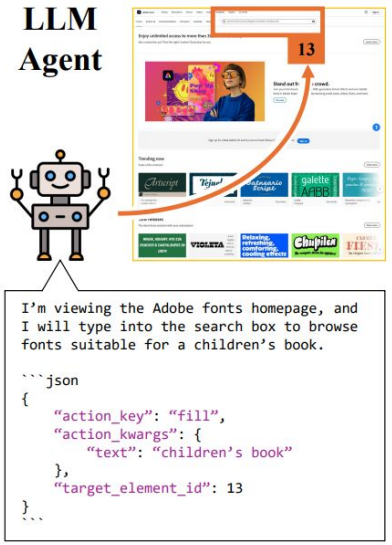
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.

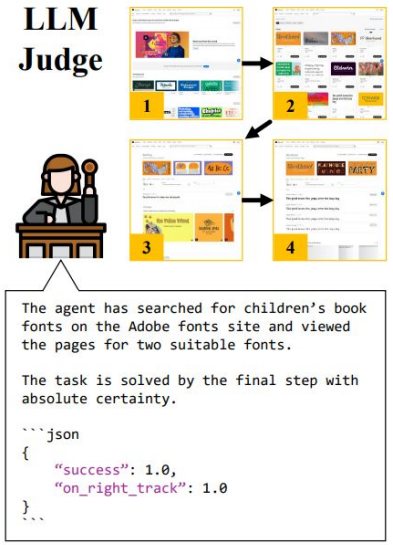




Revisiting The InSTA Pipeline

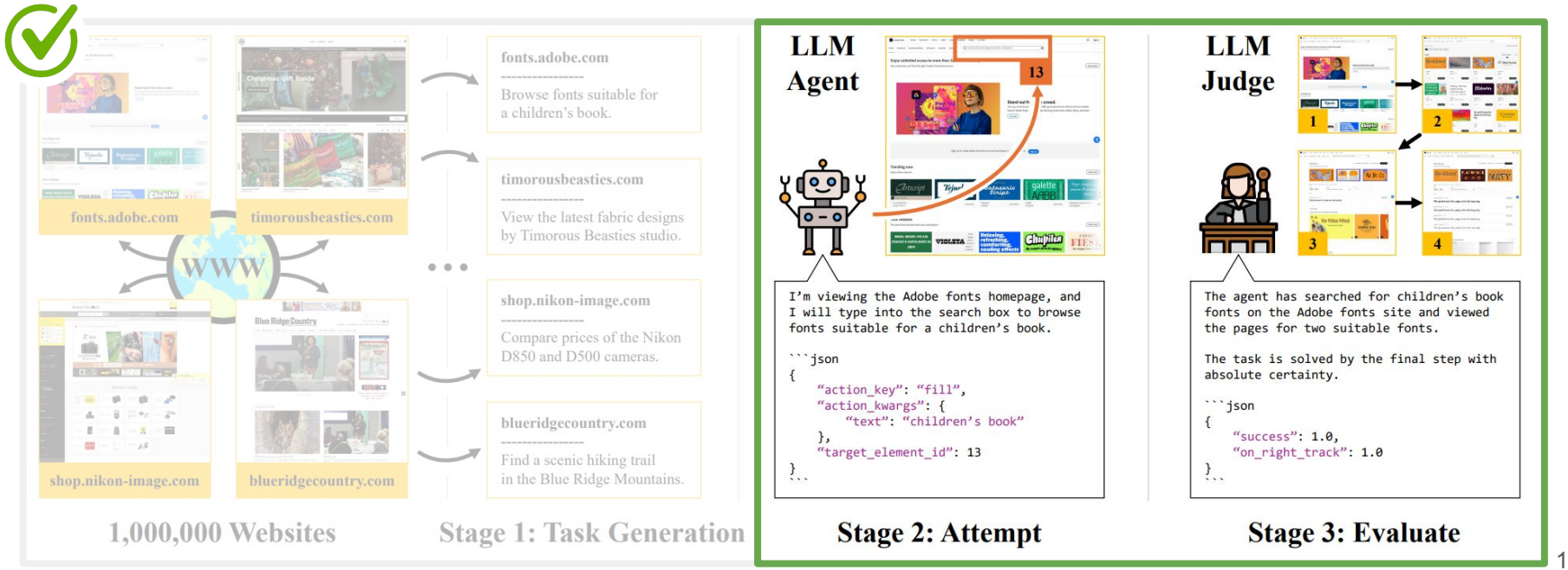
- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.
- We now have **broad and diverse** agent tasks.





Revisiting The InSTA Pipeline

- **Key Idea:** LLMs can **generate and verify** agent tasks on live websites.
- Now we can **evaluate agents** on these tasks.



How Do We Evaluate **Open-Ended Tasks**?

How Do We Evaluate **Open-Ended Tasks**?

A simple strategy works for evaluation:

$$\mathbf{r}_T = f^{\text{text}\to\text{val}}\Big(\text{LLM}([\, \mathbf{y}_{\text{sys}}, \mathbf{c}, \mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T) \,]) \Big)$$

How Do We Evaluate **Open-Ended Tasks**?

A simple strategy works for evaluation:

$$\mathbf{r}_T = f^{\text{text}\to\text{val}}\Big(\text{LLM}([\, \mathbf{y}_{\text{sys}}, \mathbf{c}, \mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T) \,]) \Big)$$

- Given a task $\mathbf{c}$, and a trajectory $\mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T)$, we can **directly estimate the probability of success** via **next-token prediction**.

How Do We Evaluate **Open-Ended Tasks**?

A simple strategy works for evaluation:

$$\mathbf{r}_T = f^{\text{text}\to\text{val}}\Big(\text{LLM}([\, \mathbf{y}_{\text{sys}}, \mathbf{c}, \mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T) \,]) \Big)$$

- Given a task $\mathbf{c}$, and a trajectory $\mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T)$, we can **directly estimate the probability of success** via **next-token prediction**.
- Recent work shows this **works well for math tasks** [3].

[3] Generative Verifiers: Reward Modeling as Next-Token Prediction, Lunjun Zhang, et al., Google, 2024.

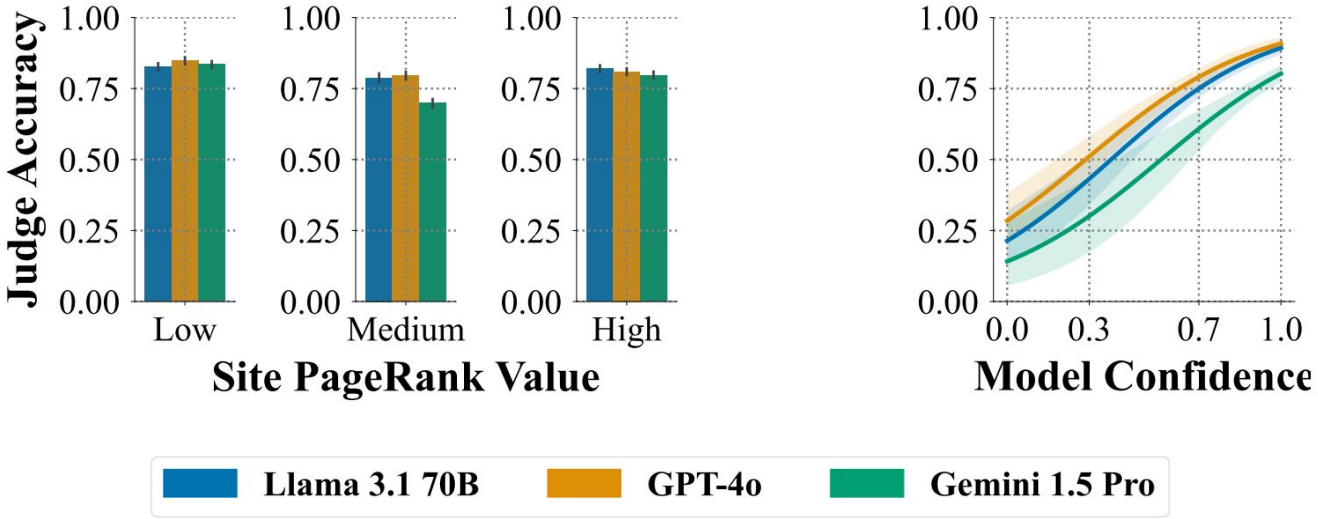
How Do We Evaluate **Open-Ended Tasks**?

A simple strategy works for evaluation:

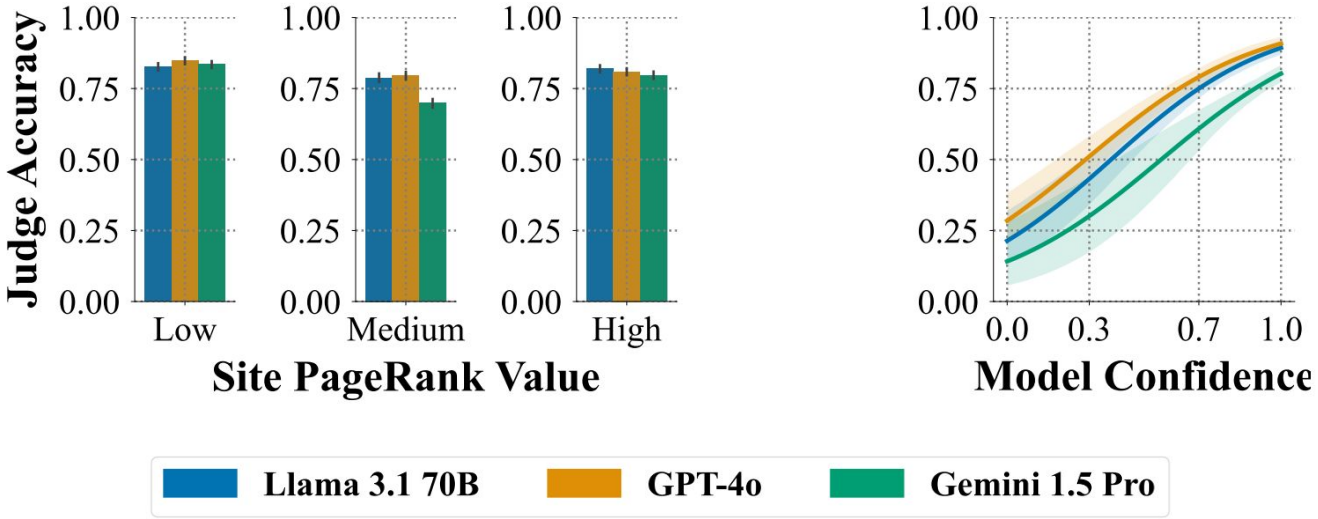
$$\mathbf{r}_T = f^{\text{text}\to\text{val}}\Big(\text{LLM}([\, \mathbf{y}_{\text{sys}}, \mathbf{c}, \mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T) \,]) \Big)$$

- Given a task $\mathbf{c}$, and a trajectory $\mathbf{a}_1, \cdots, \mathbf{a}_T, \text{Enc}(\mathbf{s}_T)$, we can **directly estimate the probability of success** via **next-token prediction**.
- How robust are LLMs for **evaluating agent tasks**?

Language Models Are **Robust Evaluators**

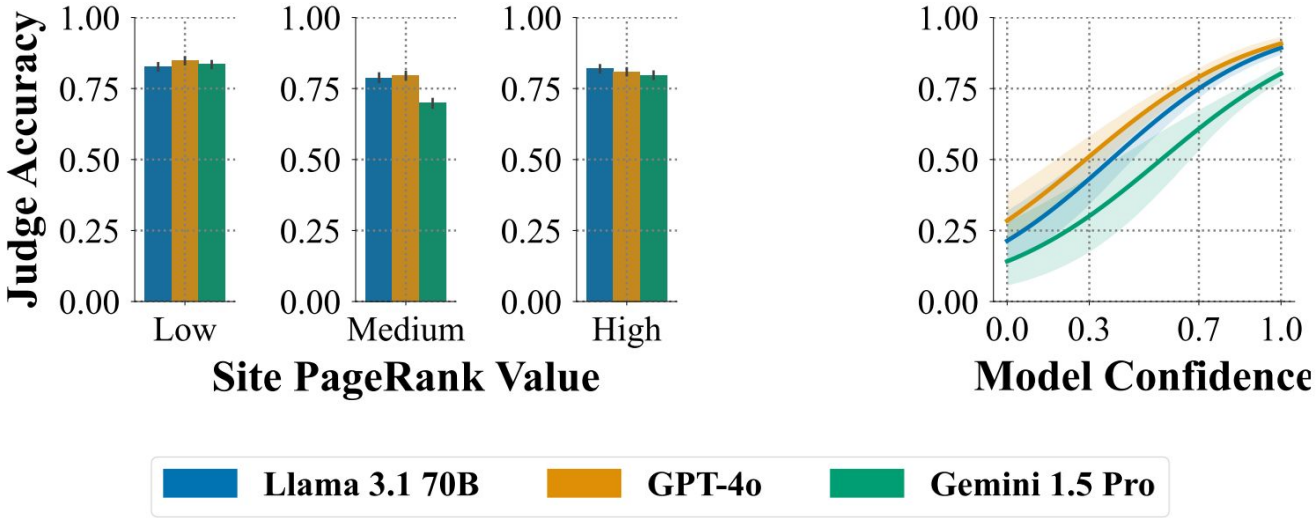


Language Models Are **Robust Evaluators**



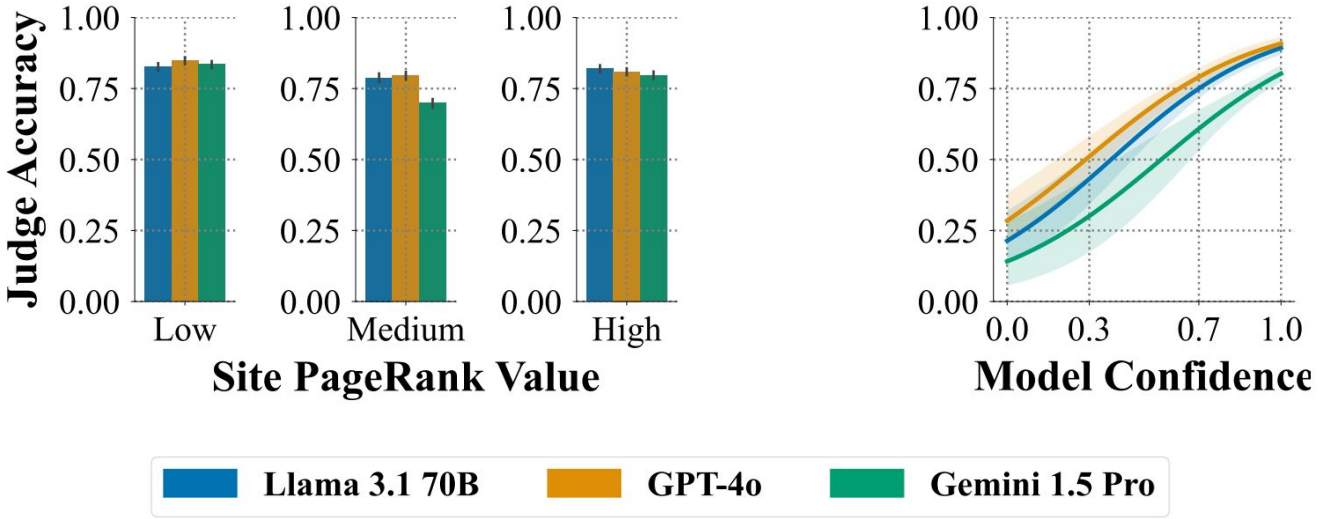
- LLMs show **high accuracy** for **detecting successful trajectories**.

Language Models Are **Robust Evaluators**



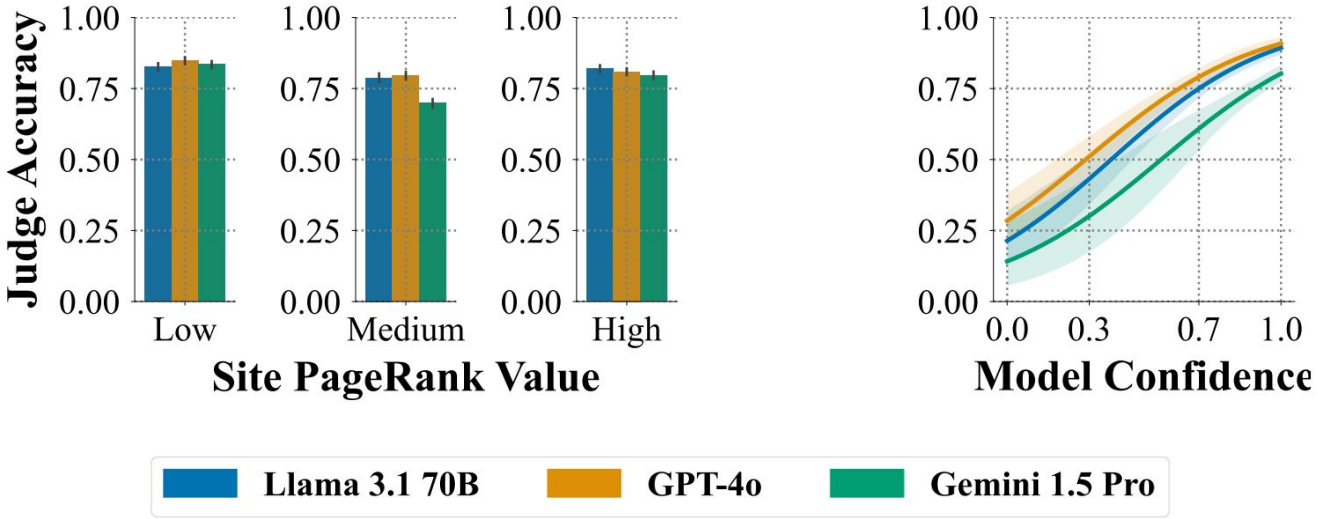
- LLMs show **high accuracy** for **detecting successful trajectories**.
- **GPT-4o □ 82.6%** , Llama 3.1 70B □ 81.7% , Gemini 1.5 Pro □ 78.0%

Language Models Are **Robust Evaluators**



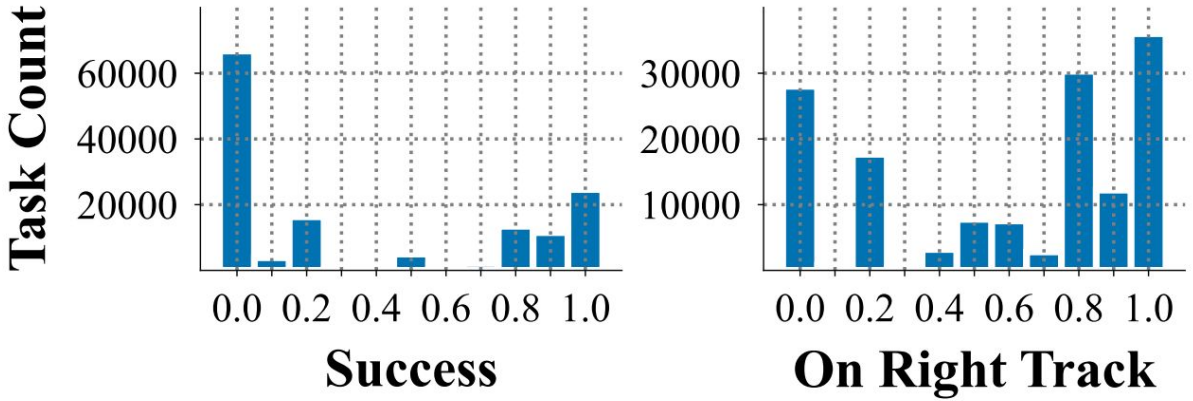
- **Accuracy improves** as LLMs become **more confident**.

Language Models Are **Robust Evaluators**

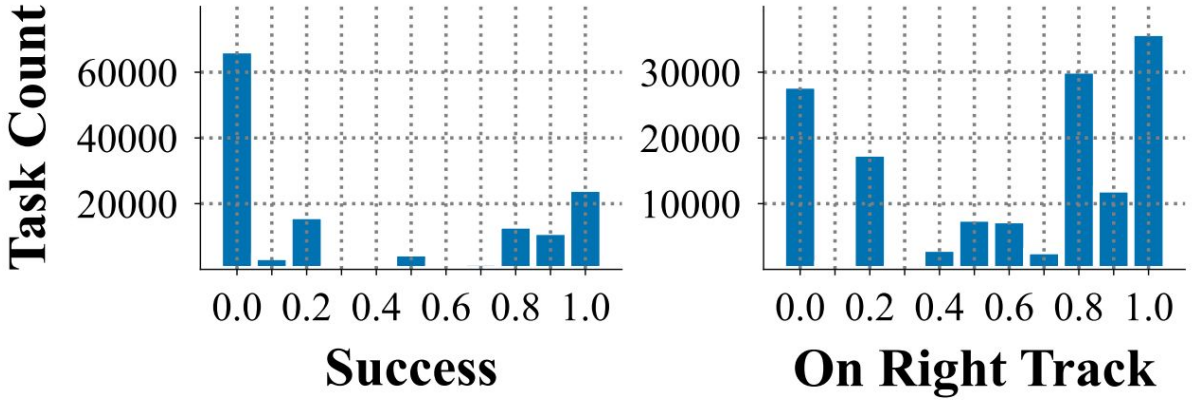


- **Accuracy improves** as LLMs become **more confident**.
- For sites with **confidence = 1.0**, **Llama 3.1 70B 93.1% Accuracy**.

Scaling To **150,000 Agents** Using Llama

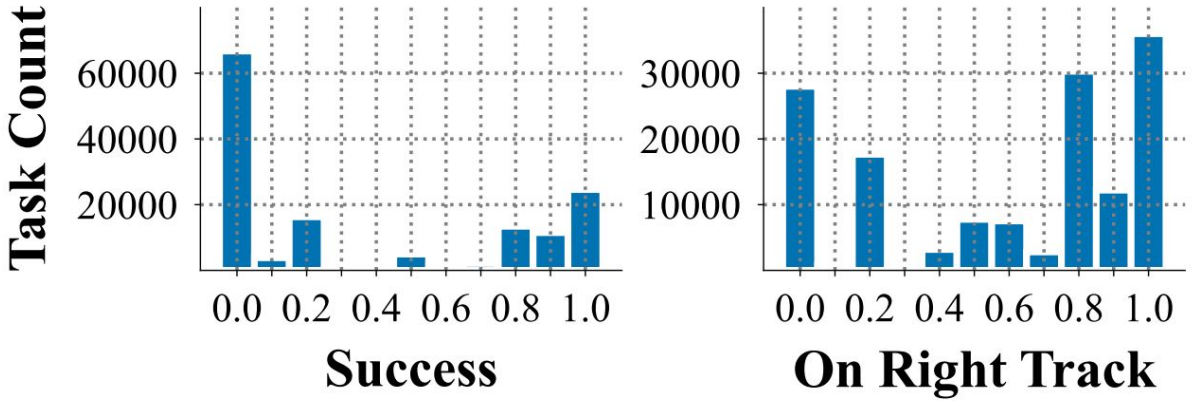


Scaling To **150,000 Agents** Using Llama



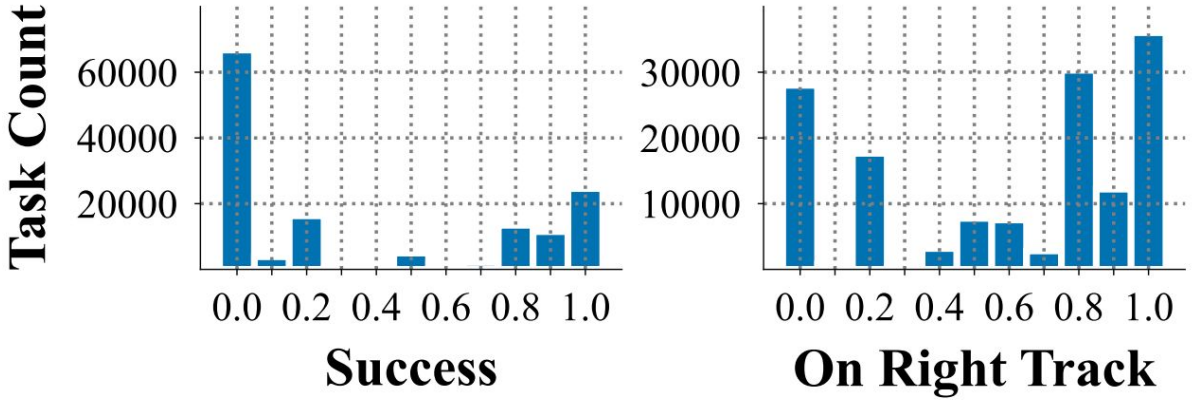
- **16.7% of tasks** are estimated to be **successful with confidence = 1.0.**

Scaling To **150,000 Agents** Using Llama



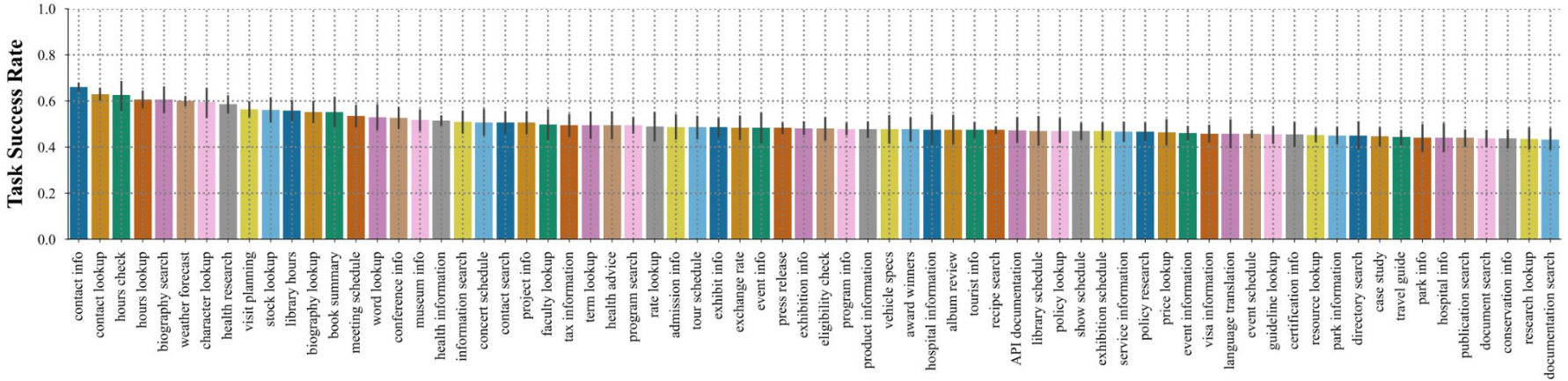
- **16.7% of tasks** are estimated to be **successful with confidence = 1.0.**
- Certain agents that failed are **on the right track**.

Scaling To **150,000 Agents** Using Llama

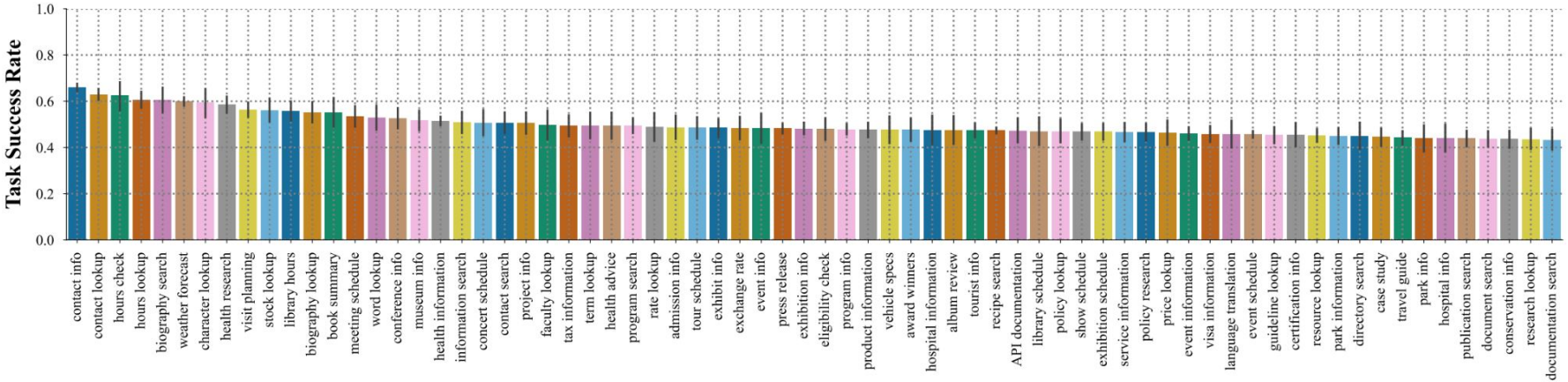


- **16.7% of tasks** are estimated to be **successful with confidence = 1.0.**
- Certain agents that failed are **on the right track**.
- Suggests **more agents can succeed** given more compute.

Most Successful Task Categories

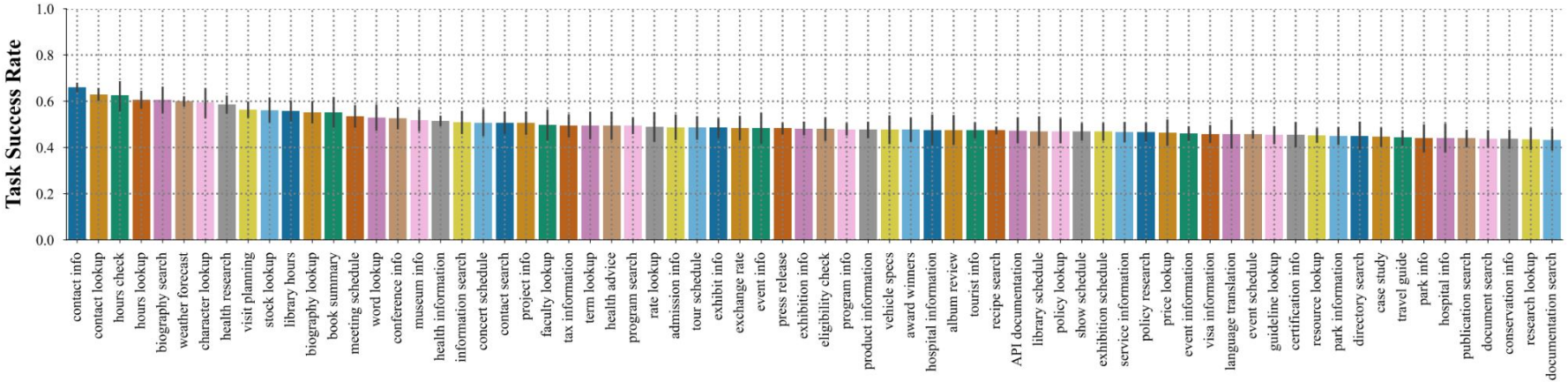


Most Successful Task Categories



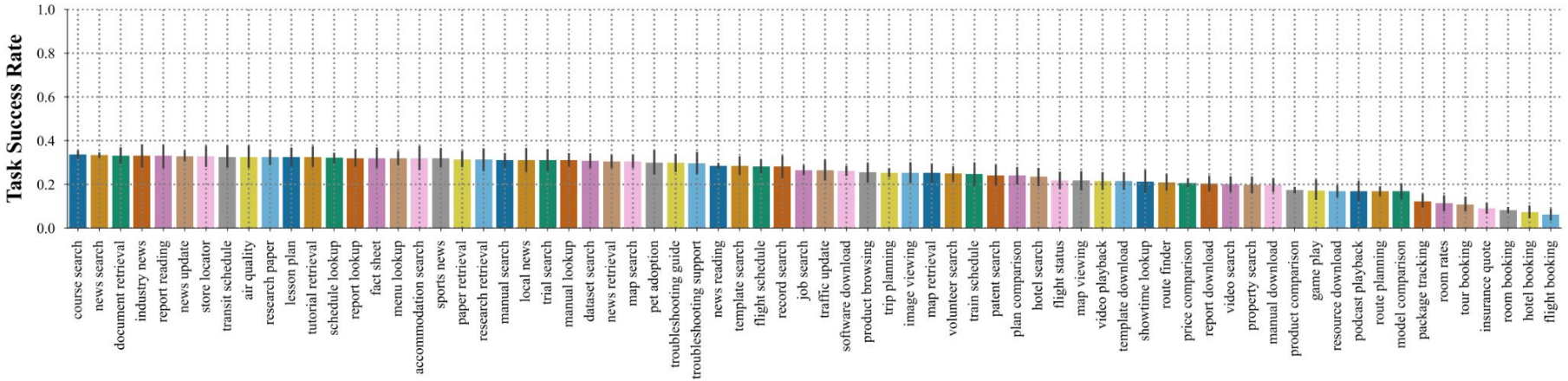
- Top 22 categories are solved with **> 50% success rate** with Llama 3.1 70B.

Most Successful Task Categories

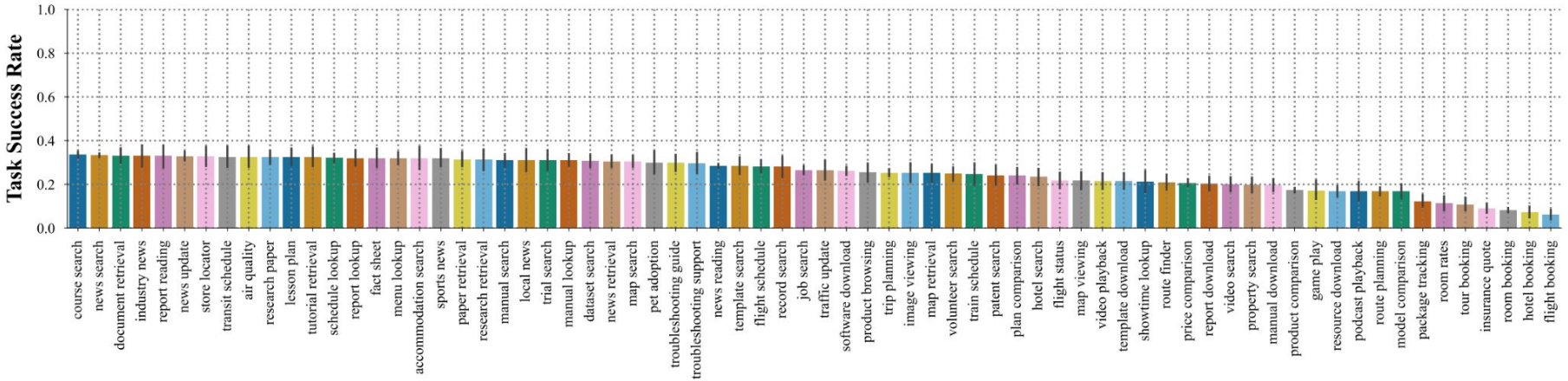


- Top 22 categories are solved with **> 50% success rate** with Llama 3.1 70B.
- Llama is most effective at **research-based tasks**.

Least Successful Task Categories

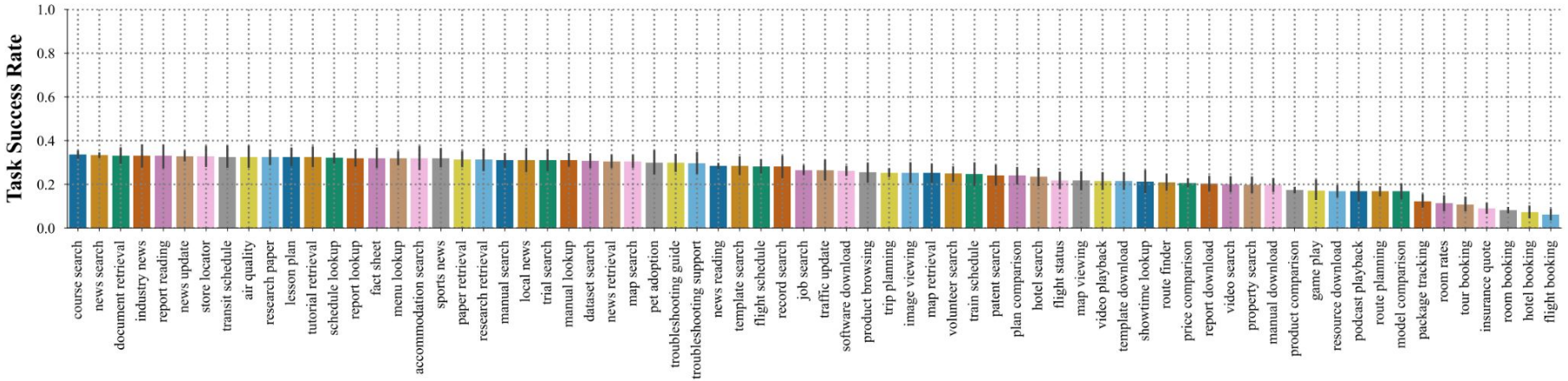


Least Successful Task Categories



- Agents often struggle to **remember and reason** about previous interactions.

Least Successful Task Categories

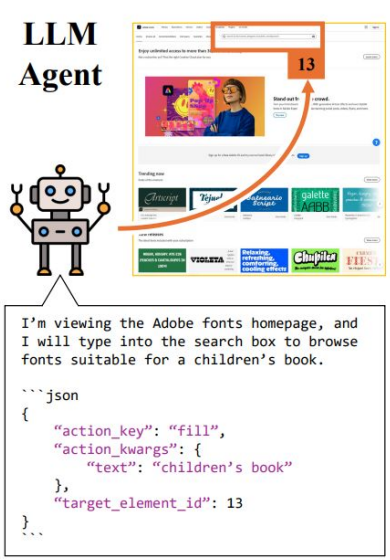


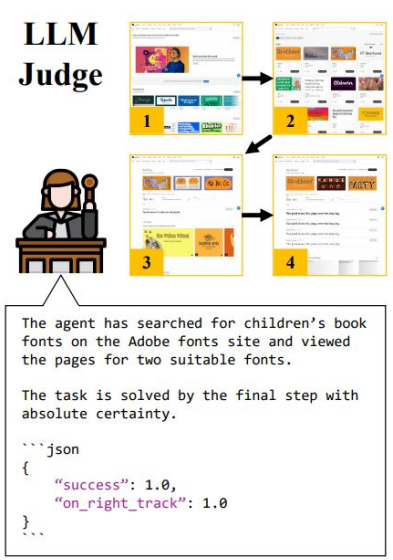
- Agents often struggle to **remember and reason** about previous interactions.
- Providing **tools like a note-taking app** could unlock these abilities.

Agent & Judge Questions

Wrapping Up The InSTA Pipeline

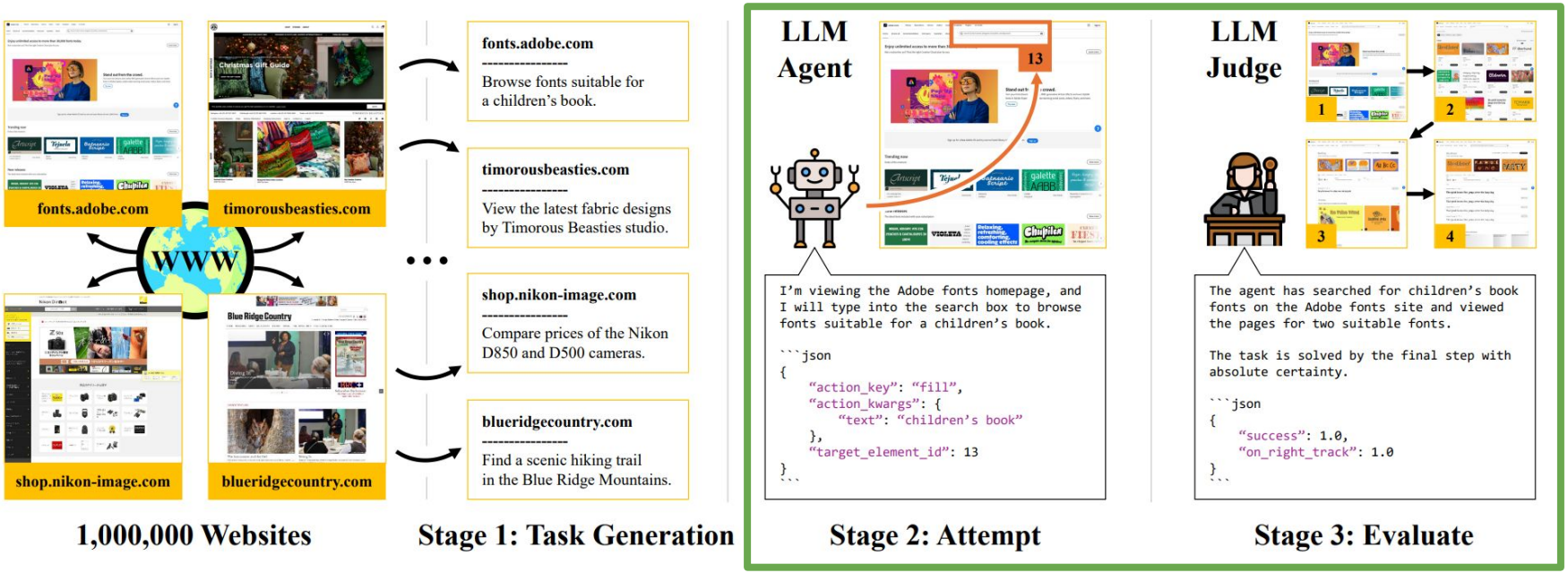
- We've covered **generation** of agent tasks.





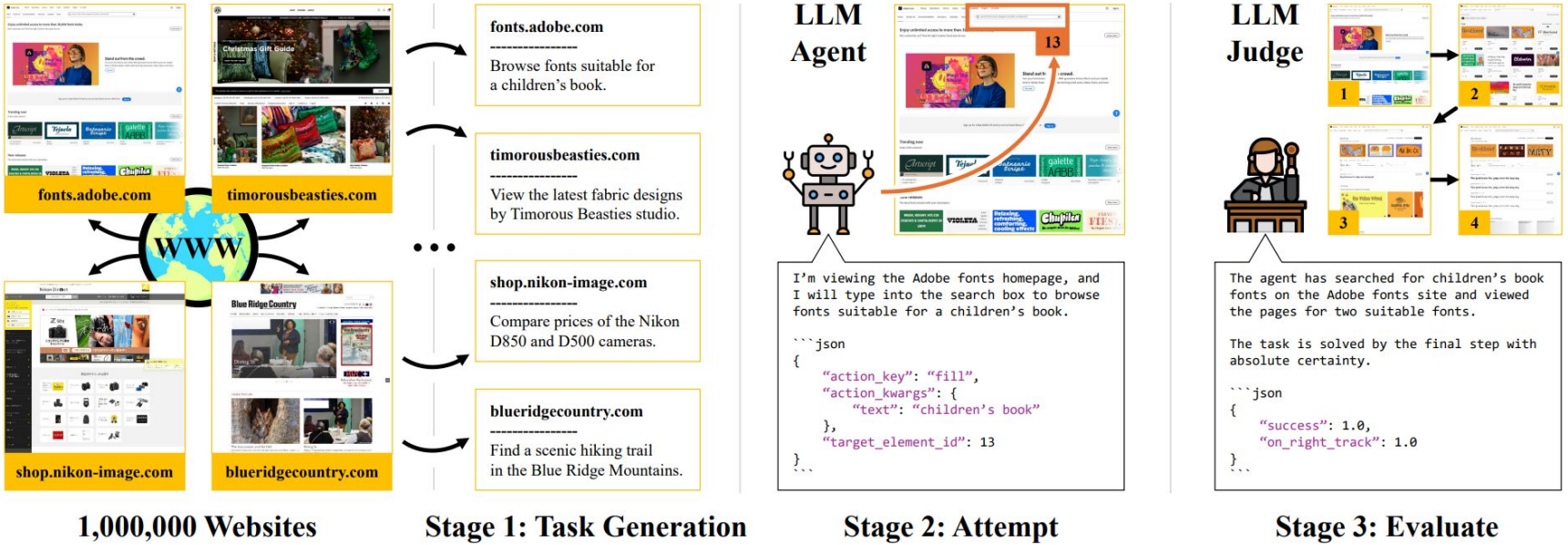
Wrapping Up The InSTA Pipeline

- We've covered **generation and verification** of agent tasks.

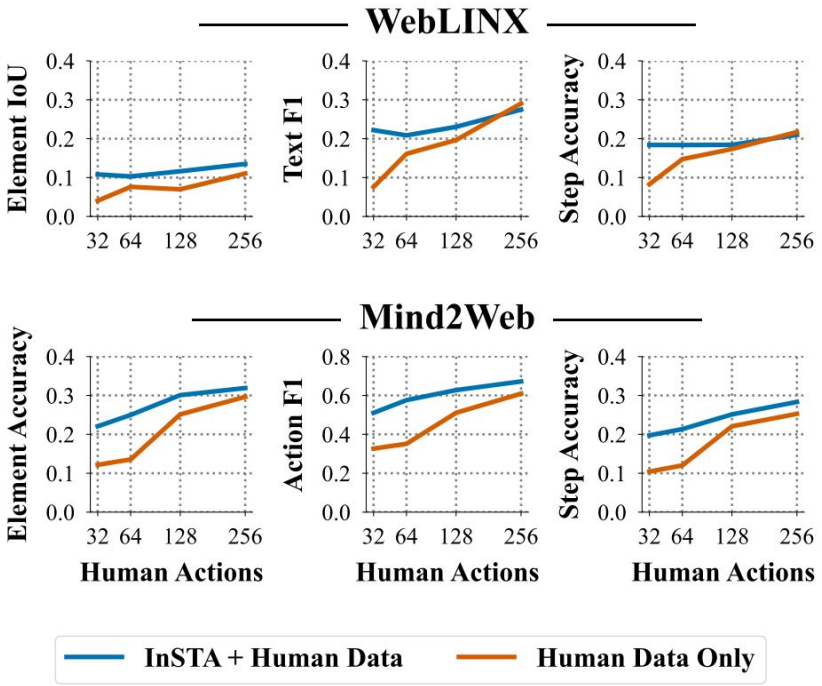


Wrapping Up The InSTA Pipeline

- We've covered **generation and verification** of agent tasks.
- How useful is this data for **training agents**?

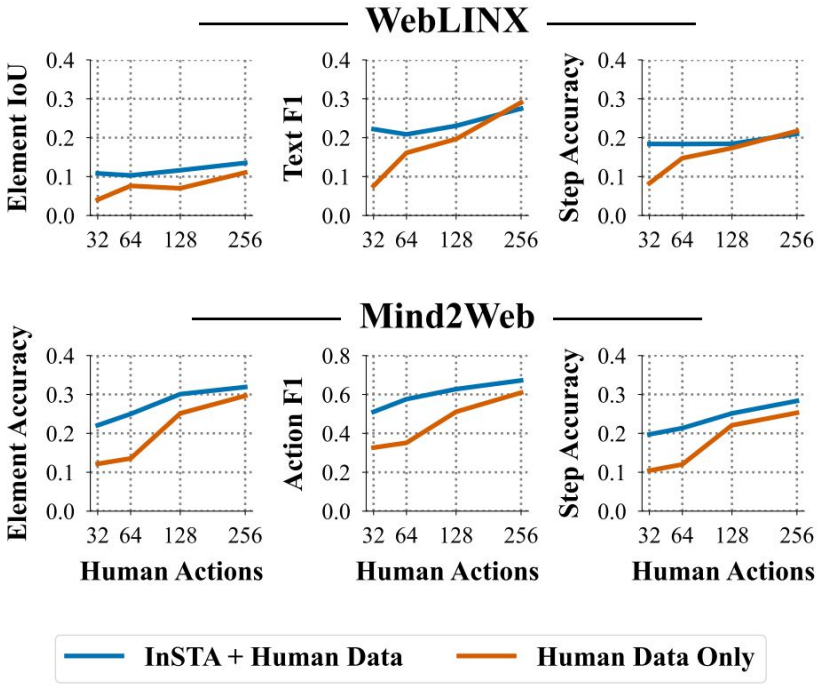


InSTA Improves **Data Efficiency**



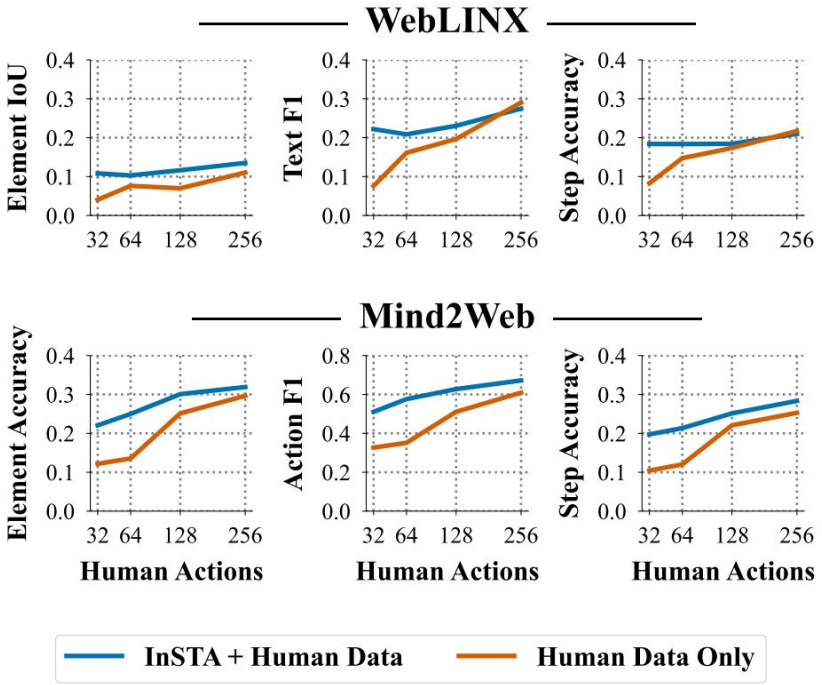
InSTA Improves **Data Efficiency**

- **In data-limited settings** derived from Mind2Web and WebLINX, **SFT on a mix** of our data and human data **improves efficiency**.

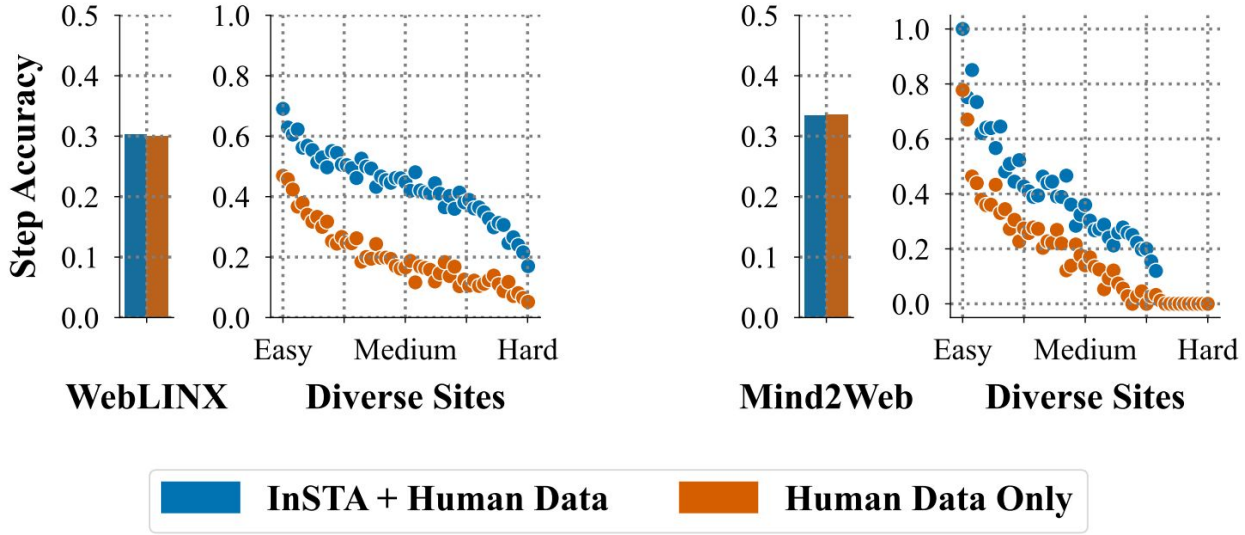


InSTA Improves **Data Efficiency**

- **In data-limited settings** derived from Mind2Web and WebLINX, **SFT on a mix** of our data and human data **improves efficiency**.
- Best gains for 32 human actions, performance **begins to converge** as human data size increases.

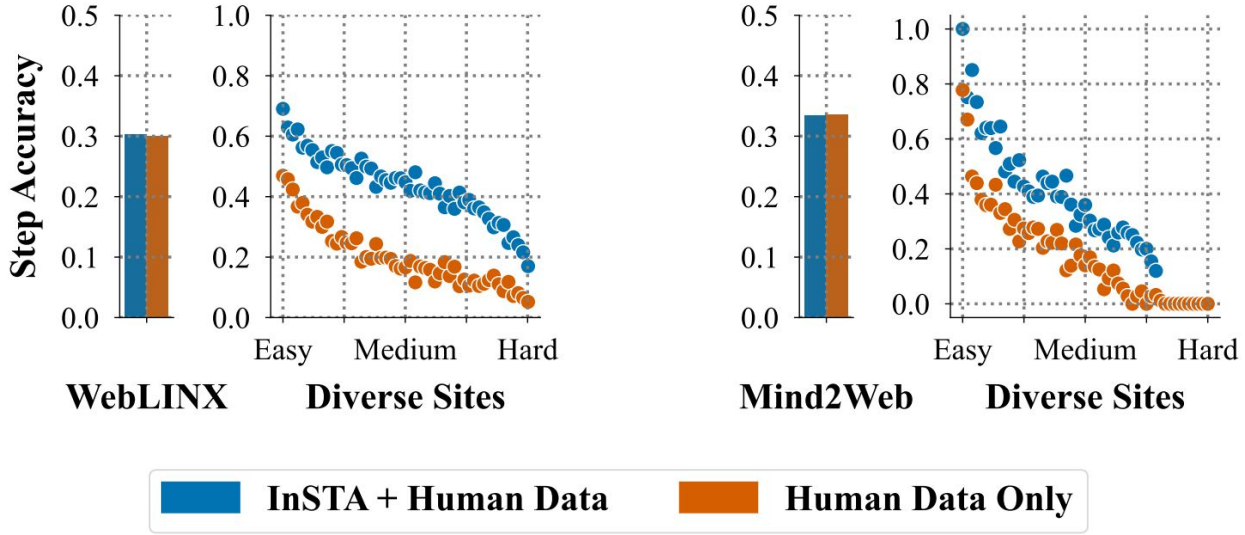


InSTA Improves **Generalization**



- These benchmarks test agents on a **limited set of popular sites**.

InSTA Improves **Generalization**



- These benchmarks test agents on a **limited set of popular sites**.
- Agents trained on **all human data** from these benchmarks **struggle to generalize** to a more diverse test set of 500 real sites.

Wrapping Up The Talk

- We developed InSTA, a pipeline **driven by language models** to unlock **internet-scale data** for training agent models.
- Scaled agents to 150k live web tasks.
- InSTA can improve **data-efficiency**, and **generalization**.

[4] CommonCrawl, The Common Crawl Foundation, 2024.

Wrapping Up The Talk

- We developed InSTA, a pipeline **driven by language models** to unlock **internet-scale data** for training agent models.
- Scaled agents to 150k live web tasks.
- InSTA can improve **data-efficiency**, and **generalization**.

What Comes Next?

There are **385M unique domains** in the Common Crawl PageRank **[4]**, so a significant amount of data is left to **scale further**.

[4] CommonCrawl, The Common Crawl Foundation, 2024.

Wrapping Up The Talk

- We developed InSTA, a pipeline **driven by language models** to unlock **internet-scale data** for training agent models.
- Scaled agents to 150k live web tasks.
- InSTA can improve **data-efficiency**, and **generalization**.

What Comes Next?

There are **385M unique domains** in the Common Crawl PageRank **[4]**, so a significant amount of data is left to **scale further**.

I'm excited about Reinforcement Learning with InSTA.

[4] CommonCrawl, The Common Crawl Foundation, 2024.

Final Questions?