NLP and Humans

CSE 517 Lecture

Sebastin Santy

Human-Computer Interaction

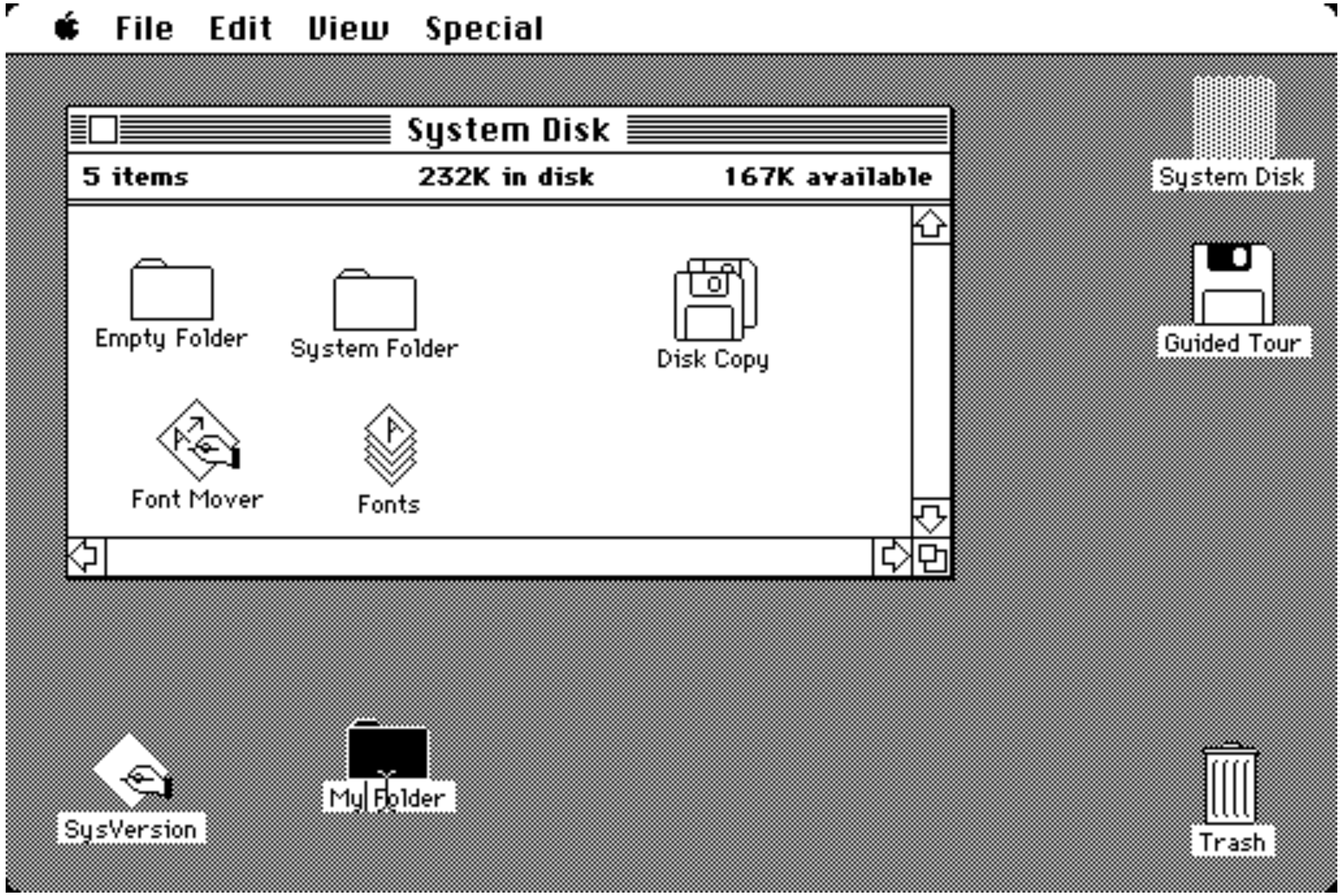
Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers.

Carroll, John M. "Human computer interaction (HCI). Interaction-Design. org." (2009).

Human-Computer Interaction

Carroll, John M. "Human computer interaction (HCI). Interaction-Design. org." (2009).

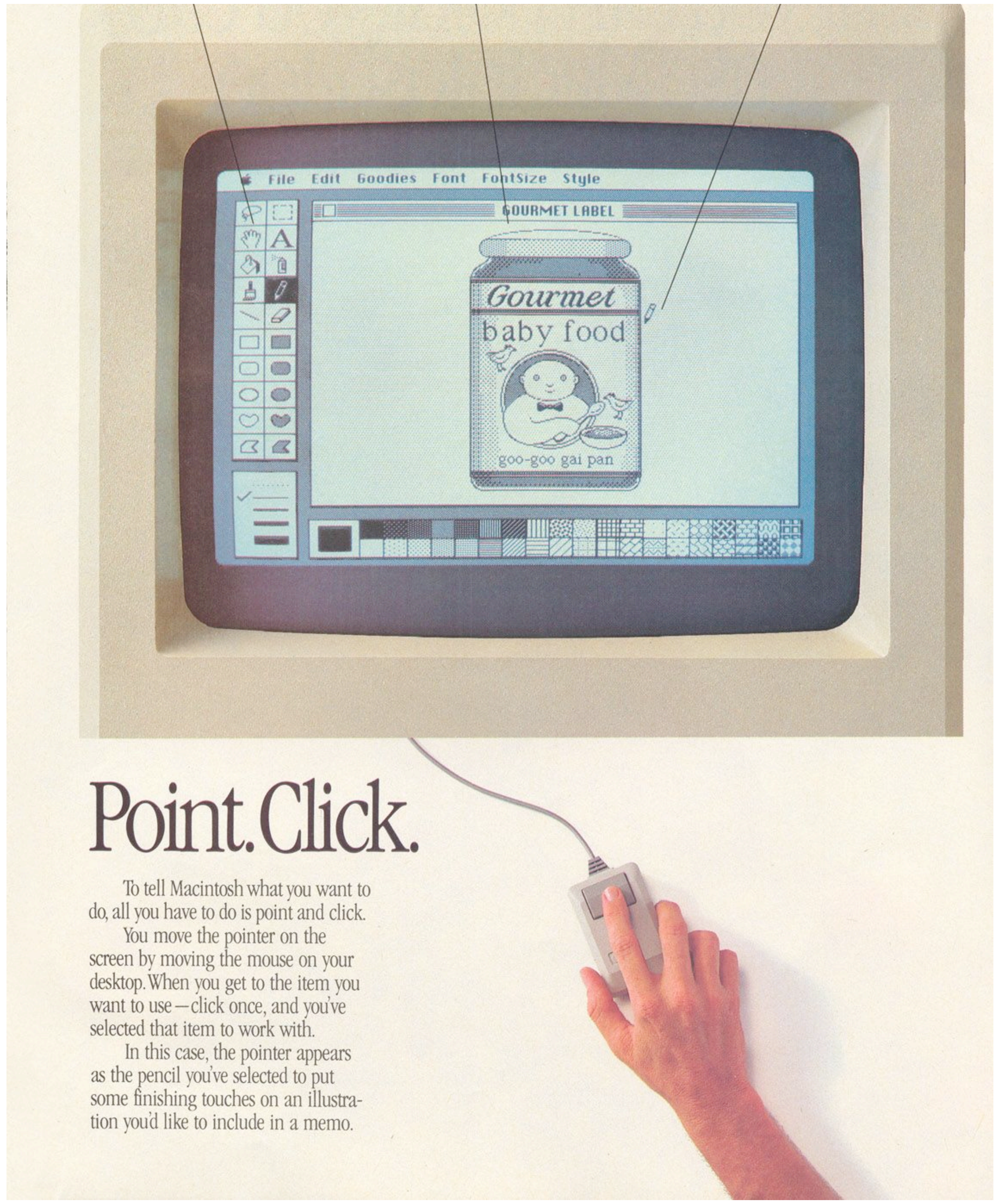
Human-Computer Interaction

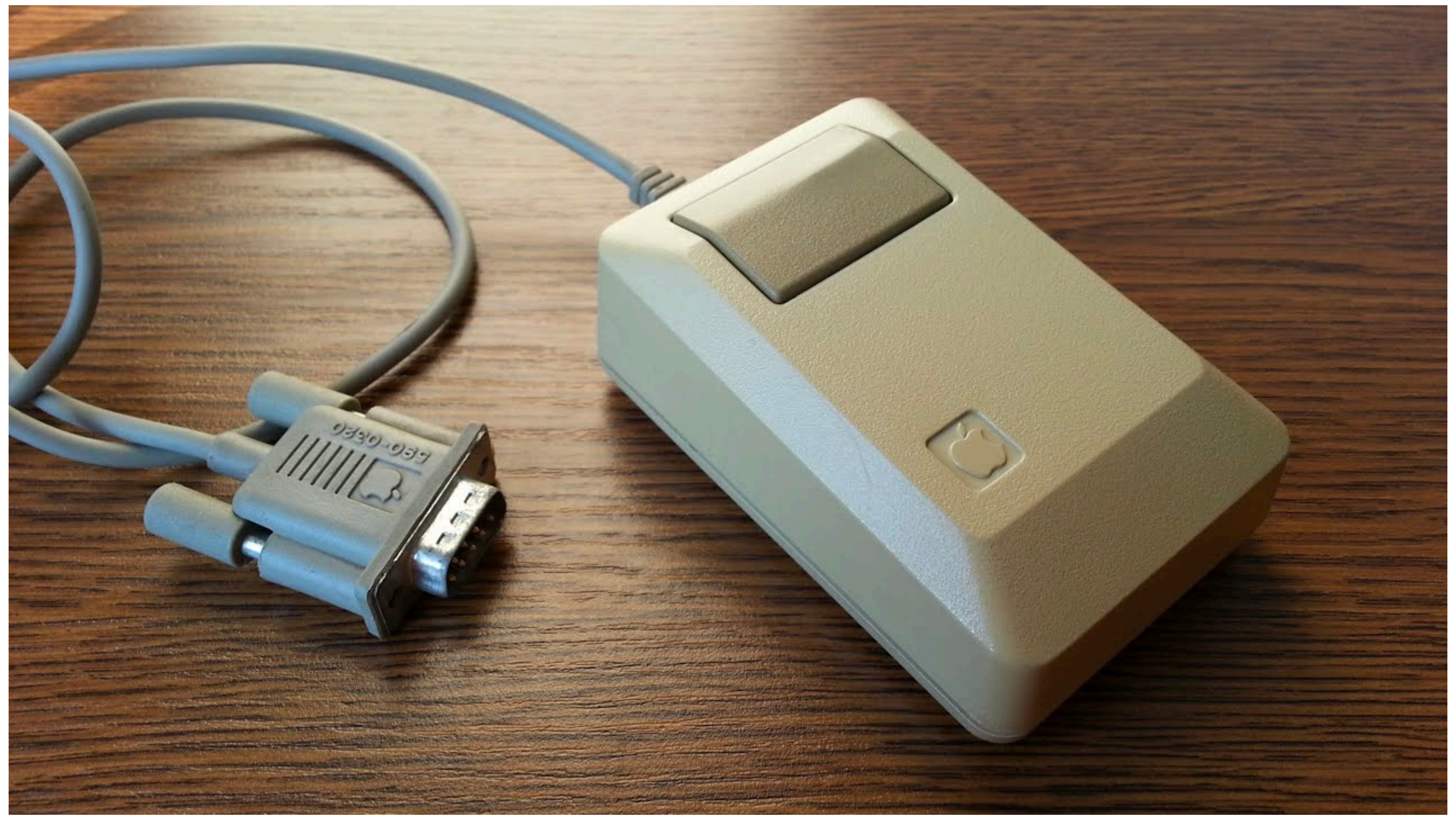


Carroll, John M. "Human computer interaction (HCI). Interaction-Design. org." (2009).

Human-Computer Interaction

Computer Mouse





Carroll, John M. "Human computer interaction (HCI). Interaction-Design. org." (2009).

Human-Computer Interaction

Brief History of HCI

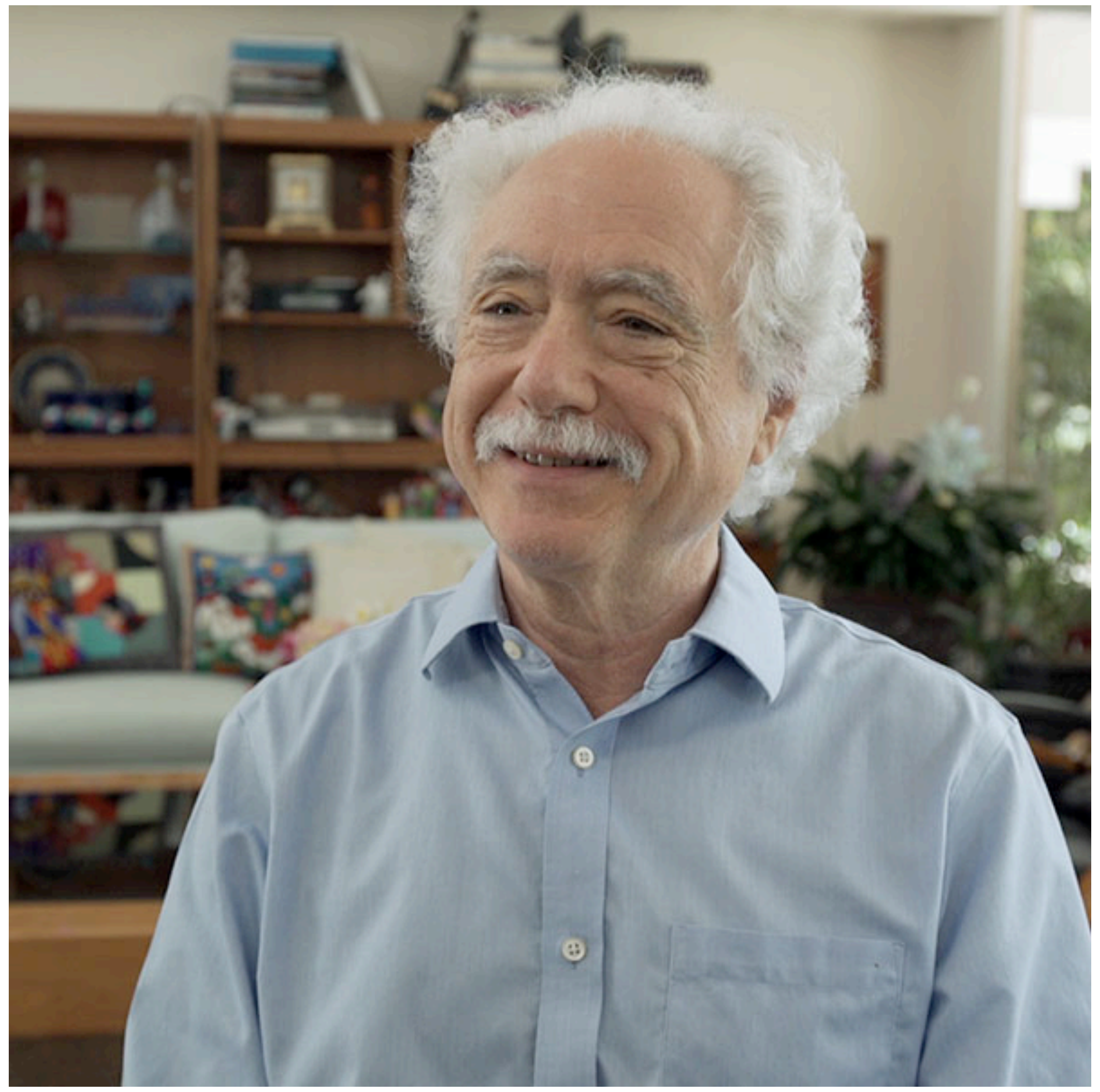
Alan Newell

Herb Simon

- HCI claims Alan Newell as the founding figure among others
- Alan Newell and Herb Simon were also pioneers of AI; first AI program called Logic Theorist to solve math theorems
- Turing award in 1975 for contributions to AI and human cognition

Human-Computer Interaction

Brief History of HCI

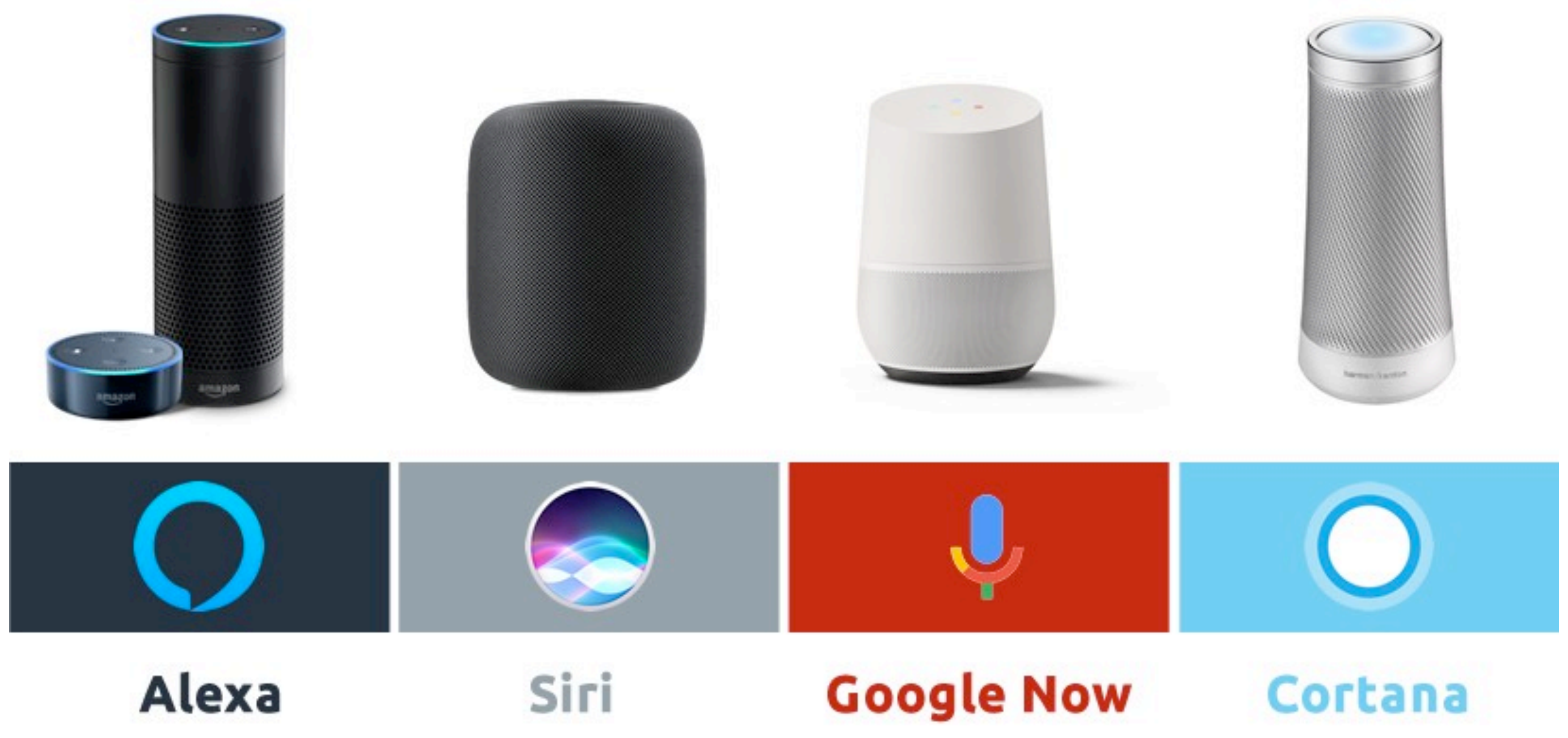


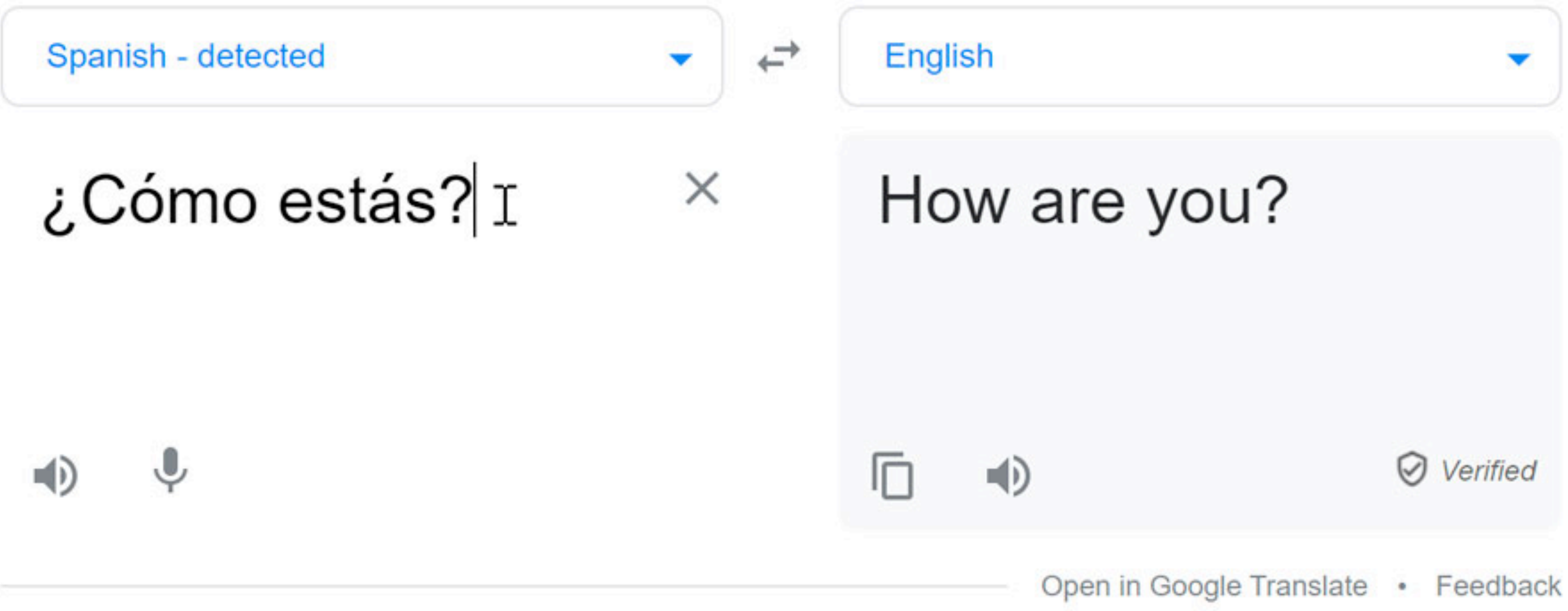
Terry Winograd

- Known in AI for work on natural language understanding; SHRDLU. Winograd Schema.
- Founded Stanford HCI group
- Advisor to Larry Page, Sergey Brin

NLP and Humans

Why should we care?





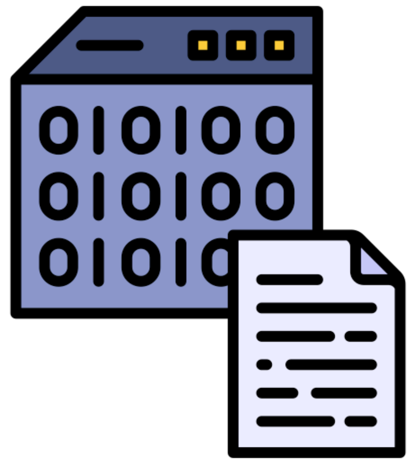
NLP and Humans

- We've mostly talked about NLP in isolation
- But at the end of the day NLP is about engineering tools
  - to be used by humans
    - for achieving their tasks
- This lecture is about:
  - Covering topics when humans and NLP models interact
  - Specifically, highlighting Issues that arise during interaction

NLP and Humans

Evaluation

Interaction



Data Curation

Social Biases

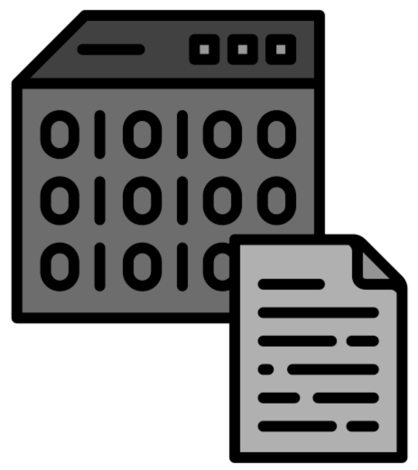
Downstream Impacts

Comp. Social Science

NLP and Humans

Evaluation

Interaction



Data Curation

Social Biases

Downstream Impacts

Comp. Social Science

Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation
  - What makes a good translation?
    - The translation is grammatical and fluent? (Fluency, Grammaticality)
    - The translation preserves the meaning? (Adequacy)
    - The translation sounds natural? (Naturalness)
    - The translation uses local phrases and idioms? (Contextualness)

Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation (Fluency, Adequacy)

Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation (Fluency, Adequacy)
- Summarization (Meaning Preservation, Coherence)

Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation (Fluency, Adequacy)
- Summarization (Meaning Preservation, Coherence)
- Story Generation (Relevance, Naturalness)

Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation (Fluency, Adequacy)
- Summarization (Meaning Preservation, Coherence)
- Story Generation (Relevance, Naturalness)
- Dialog Agents (Appropriateness, Answerability)

Evaluation

Criterion	Total	Criterion	Total
Fluency	40 (27%)	Readability	9 (6%)
Overall quality	29 (20%)	Appropriateness	7 (5%)
Informativeness	15 (10%)	Meaning preservation	6 (4%)
Relevance	15 (10%)	Clarity	5 (3%)
Grammaticality	14 (10%)	Non-reduncancy	4 (3%)
Naturalness	12 (8%)	Sentiment	4 (3%)
Coherence	10 (7%)	Consistency	4 (3%)
Accuracy	10 (7%)	Answerability	4 (3%)
Correctness	9 (6%)	Other criteria	124 (48%)*

van der Lee, Chris, et al. "Human evaluation of automatically generated text: Current trends and best practice guidelines." *Computer Speech & Language* 67 (2021)

Evaluation

- How to evaluate Natural Language Generation systems?

- In a previous lecture on Machine Translation:

BLEU Score

N-gram overlap between machine translation output and reference translation

Evaluation

- How to evaluate Natural Language Generation systems?

- In a previous lecture on Machine Translation:

BLEU Score

N-gram overlap between machine translation output and reference translation

What does it capture?

Evaluation

- n-gram precision -> BLEU
- n-gram w/ synonym match -> METEOR
- tf-idf weighted n-gram -> CIDER
- n-gram recall -> ROUGE
- % of insert,delete, replace -> WER

n-gram match

- EDIT-DISTANCE

distance-based

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- n-gram precision -> BLEU
- n-gram w/ synonym match -> METEOR
- tf-idf weighted n-gram -> CIDER
- n-gram recall -> ROUGE
- % of insert,delete, replace -> WER

n-gram match

- EDIT-DISTANCE

distance-based

Untrained Automatic Metrics

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in several ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT

Any flaws?

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT
  - Human-centric Evaluation

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT
  - Human-centric Evaluation Most Preferred

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT
  - Human-centric Evaluation

Harder to do

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT
  - Human-centric Evaluation

Harder to do

but

More **accurate** evaluation

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Human Evaluation

- Human ratings are considered gold-standard in NLG evaluation
- Given a generated text how does a human rate it?

Ethayarajh, Kawin, and Dan Jurafsky. "The Authenticity Gap in Human Evaluation." *arXiv preprint arXiv:2205.11930* (2022).

Human Evaluation

- Let's try a sample human evaluation

On a scale of 1-5, rate the naturalness of the sentence

"Time flies like an arrow; fruit flies like a banana"

	1	2	3	4	5	
Very unnatural	○	○	○	○	○	Very natural

Amidei et al. "The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations." (2019).

Human Evaluation

- Let’s try a sample human evaluation

How easy or difficult is the following sentence?

“Katie sipped on her cappuccino”

Very difficult	Difficult	Ok	Easy	Very easy
○	○	○	○	○

Amidei et al. "The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations." (2019).

Human Evaluation

- Rating Scale popularly known as Likert scale
- Evaluation is **outcome-level absolute assessment (OAA)**

- What are some issues with this approach?

Ethayarajh, Kawin, and Dan Jurafsky. "The Authenticity Gap in Human Evaluation." *arXiv preprint arXiv:2205.11930* (2022).

Human Evaluation

- Rating Scale popularly known as Likert scale
- Evaluation is **outcome-level absolute assessment (OAA)**

- What are some issues with this approach?
  - **Interpretation**: What is meant by 'naturalness' or 'difficulty'? How do you instruct annotators?
  - **Upper Bounds**: What does 1 and 5 mean?
  - **Interval width**: Is a jump from 3-4 same as 4-5?

Ethayarajh, Kawin, and Dan Jurafsky. "The Authenticity Gap in Human Evaluation." *arXiv preprint arXiv:2205.11930* (2022).

Human Evaluation

- Another form: Comparative Ratings

Express preference for one of the following sentences S1 or S2

S1: “Hello world, I am Alexa”
S2: “Hey there, I am Alexa”

Prefer S1 Strongly	Prefer S1	Both S1 and S2	Prefer S2	Prefer S2 Strongly
○	○	○	○	○

Amidei et al. "The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations." (2019).

Human Evaluation

- Ranking system
- Evaluation is **outcome-level relative assessment (ORA)**

- What are some issues with this approach?

Ethayarajh, Kawin, and Dan Jurafsky. "The Authenticity Gap in Human Evaluation." *arXiv preprint arXiv:2205.11930* (2022).

Human Evaluation

- Ranking system
- Evaluation is **outcome-level relative assessment (ORA)**

- What are some issues with this approach?
  - **Absolute Numbers**: What is the absolute performance of the model?
  - **Head-to-head Comparisons**: Massive number of comparisons

Ethayarajh, Kawin, and Dan Jurafsky. "The Authenticity Gap in Human Evaluation." *arXiv preprint arXiv:2205.11930* (2022).

Human Evaluation

Issue: Language Subjectivity

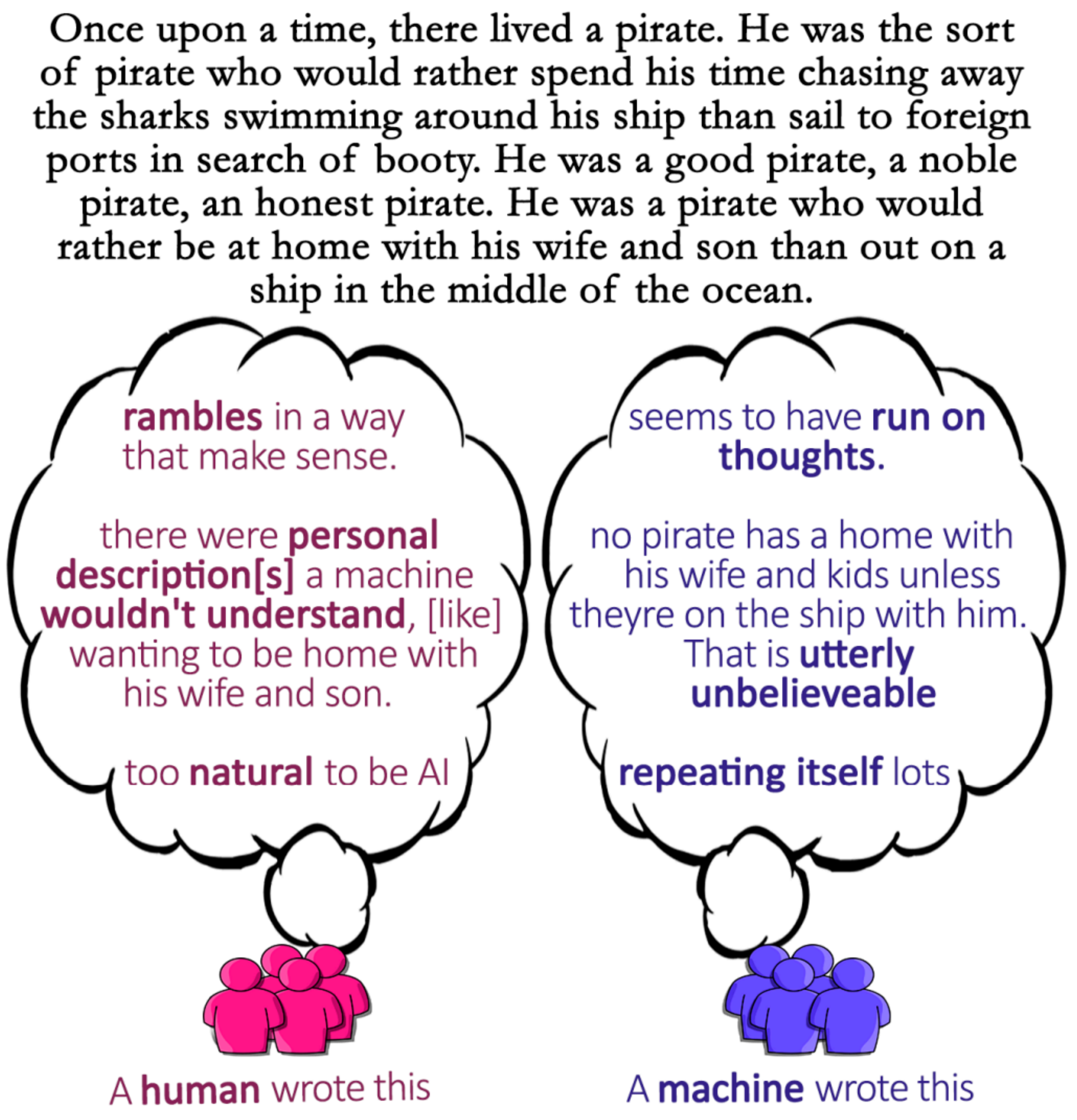
- People find a particular tone to be better than the other
  - “Hey there” vs. “Hello World”
- What is toxic?
  - Depends on the person and their demographic group
    Sap, Maarten, et al. "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.” NAACL 2022
  - Given much of research happens (and data is collected) in West, these annotations can make NLP systems unworkable
    Ghosh, Sayan, et al. "Detecting cross-geographic biases in toxicity modeling on social media.” WNUT 2021

Human Evaluation

Issue: Human Ratings

- For a long time, human ratings were gold standard

- However, with recent advances, humans find it difficult to distinguish between human-generated and model-generated text



Clark, Elizabeth, et al. "All that's' human' is not gold: Evaluating human evaluation of generated text." ACL 2021

Human Evaluation

Issue: Human Ratings

Model	Overall Acc.	Domain	Acc.	F_1	Prec.	Recall	Kripp. α	% human	% confident
GPT2	*0.58	Stories	*0.62	0.60	0.64	0.56	0.10	55.23	52.00
		News	*0.57	0.52	0.60	0.47	0.09	60.46	51.38
		Recipes	0.55	0.48	0.59	0.40	0.03	65.08	50.31
GPT3	0.50	Stories	0.48	0.40	0.47	0.36	0.03	62.15	47.69
		News	0.51	0.44	0.54	0.37	0.05	65.54	52.46
		Recipes	0.50	0.41	0.50	0.34	0.00	66.15	50.62

Table 1: §2 results, broken down by domain and model, along with the F_1, precision, and recall at identifying machine-generated text, Krippendorff's α, % human-written guesses, and % confident guesses (i.e., *Definitely* machine- or human-authored). * indicates the accuracies significantly better than random (two-sided t-test, for Bonferroni-corrected $p < 0.00333$).

Clark, Elizabeth, et al. "All that's' human' is not gold: Evaluating human evaluation of generated text." ACL 2021

Human Evaluation

- Types of human-involved evaluation
  - Intrinsic Evaluation (OAA, ORA)
    - Measure the text in itself

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Human Evaluation

- Types of human-involved evaluation
  - Intrinsic Evaluation (OAA, ORA)
    - Measure the text in itself
  - Extrinsic Evaluation
    - Measure whether the system is able to help humans achieve a task

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Human Evaluation

- Extrinsic Evaluation
  - Summarization -> Did the user get an idea of what a document was talking about?
  - Dialog Agents -> Was the user able to efficiently navigate through a website based on the outputs of a dialog agent?
  - Machine Translation -> Did the translation help user to achieve a task e.g., understanding directions and navigating in a foreign country?

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Human Evaluation

- Extrinsic Evaluation
  - How?

    Evaluate at the system level and comparing systems that differ only in the NLG module

Hastie, Helen F., and Anja Belz. "A Comparative Evaluation Methodology for NLG in Interactive Systems." *LREC*. 2014.

Human Evaluation

- Extrinsic Evaluation
  - Evaluate at the system level and comparing systems that differ only in the NLG module
  - Examples -
    - Reiter et al. (2003) generate personalized smoking cessation letters and report how many recipients actually gave up smoking.
    - Post-editing (Denkowski et al., 2014) can be used to measure a system's success by measuring how many changes a person makes to a machine-generated text.

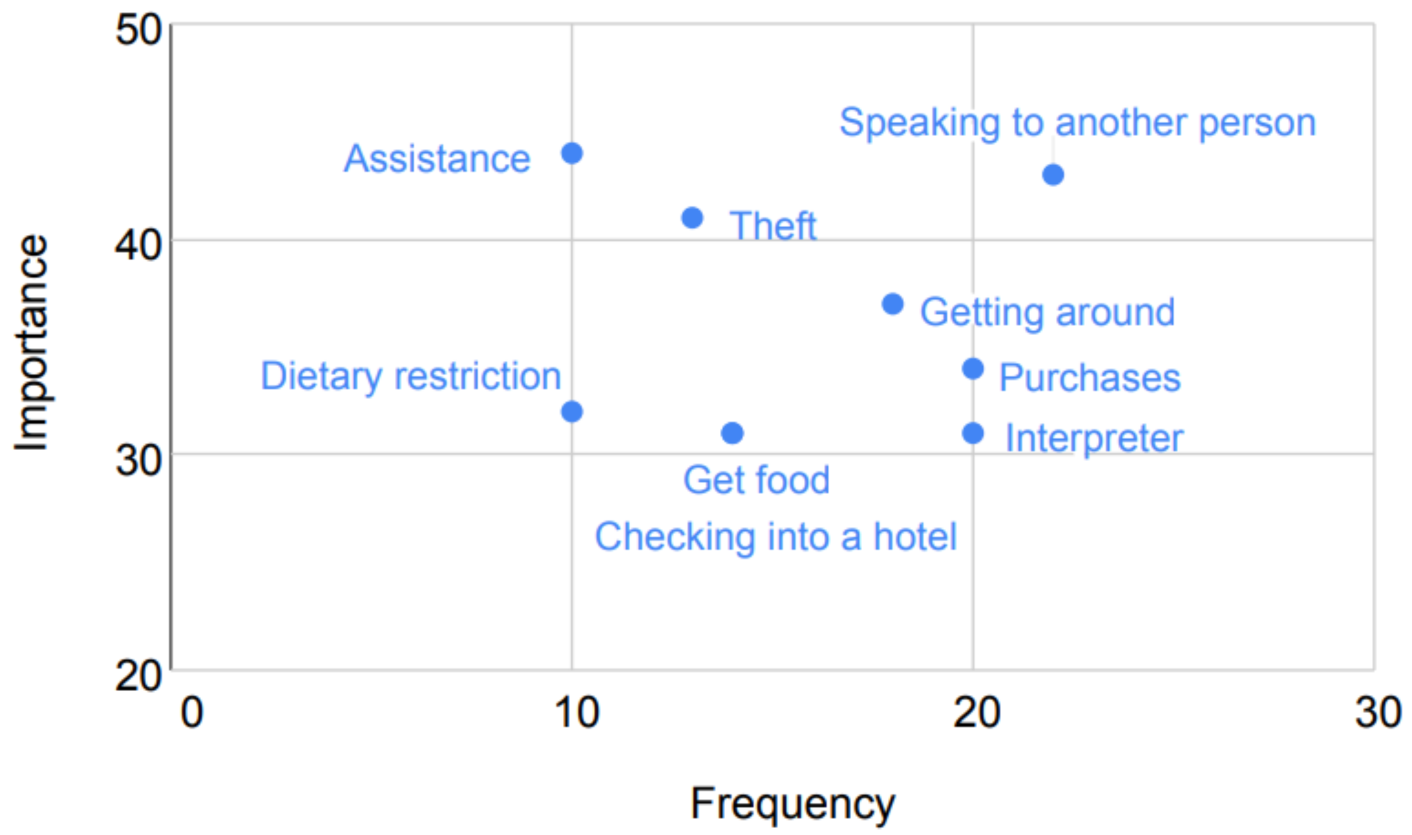
Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." *arXiv preprint arXiv:2006.14799* (2020).

Human Evaluation

- Extrinsic Evaluation
  - **Most important**, as at the end of the day, it matters whether the end-user systems are usable
  - However,
    - Difficult to operationalize in NLP research
      - Systems are expensive to build and difficult to evaluate
    - Difficult to make progress within text generation
      - Systems used in varied context; other confounders in evaluation of systems other than just generated text

Human Evaluation

- Extrinsic Evaluation
  - HCI Research has several work that takes it to people and test it
    - Liebling et al. "Unmet needs and opportunities for mobile translation AI." CHI 2020.



Scenario	Prompt
Speaking with people	I need to speak to someone who speaks another language than I do.
Getting around	I need to ask for directions but I don't speak the local language.
Purchases	I need to buy something but I don't speak the local language.
Checking into a hotel	I need to check-in to my hotel but I don't speak the local language.
Get food	I need to buy food but I don't speak the local language.
Dietary restrictions	I have a food allergy or preference I need to tell someone about but I don't speak the local language.
Assistance	I need medical assistance but I don't speak the local language.
Theft	I need help from the police but I don't speak the local language.
Interpreter	I need a language interpreter or guide to help me communicate.

Table 1. Scenarios rated by respondents on dimensions of importance and frequency.

Interaction

- Human-Teacher, Machine-Learner
  - Learning from human feedback
- Machine-leading
  - Machines initiate interactions with their optimal competence, then humans respond with suggestions
- Human-leading
  - Humans initiate the task, then machines give suggestions based on their expertise
- Human-machine collaborators
  - Either can initiate. No explicit benefit for humans or machines

Wan, Ruyuan, et al. "User or Labor: An Interaction Framework for Human-Machine Relationships in NLP." *DASH* 2022

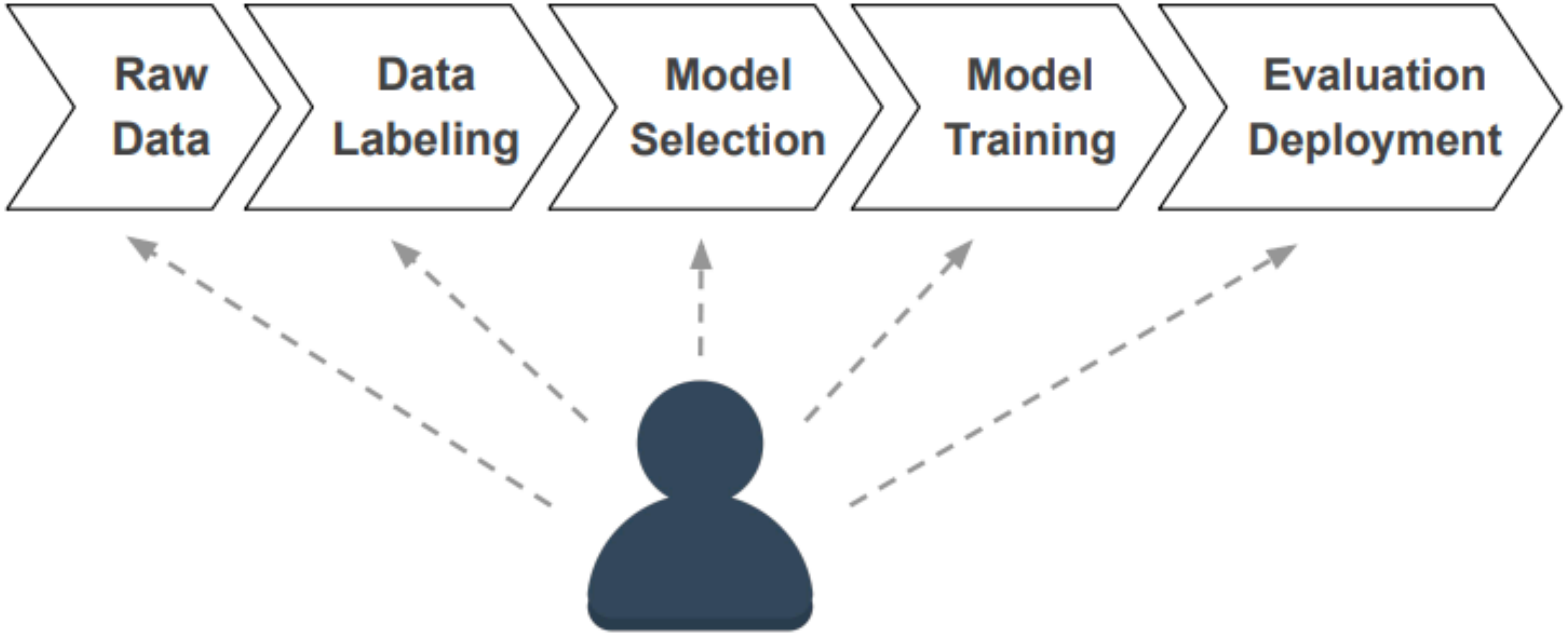
Interaction
Learning from human feedback

- Users generate rich signals that reveal model incorrectness and point to future model improvements (Krishna et. al., PNAS 2022)
  - Clickstream / Post-editing may implicitly reflect their expectations on a model like when they revise a model-generated text after accepting the suggestions
- How to integrate human feedback to improve the model itself?
- Also, called Human-in-the-loop (HITL)

Wang, Zijie J., et al. "Putting humans in the natural language processing loop: A survey." *arXiv preprint arXiv:2103.04044* (2021).

Interaction

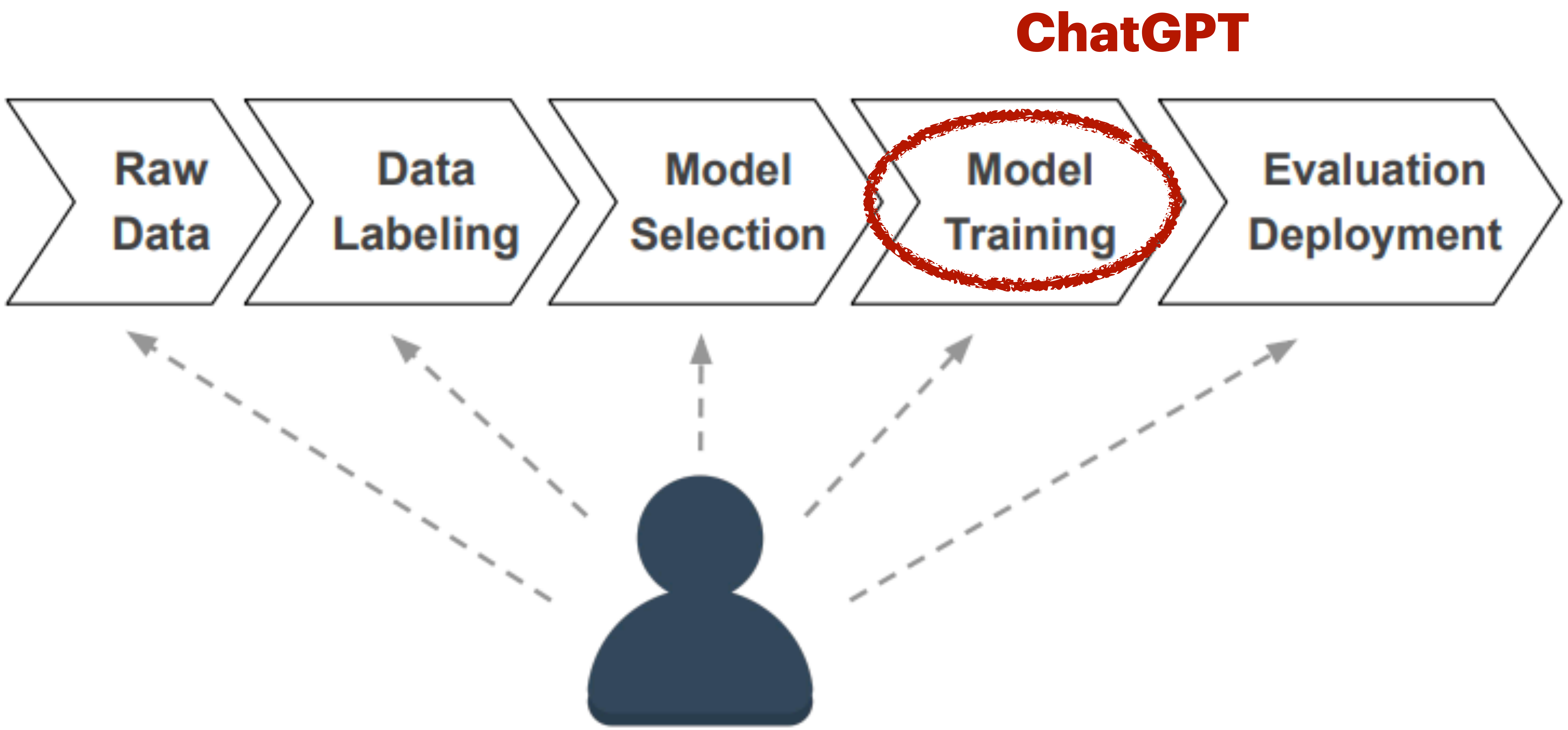
Learning from human feedback



Wang, Zijie J., et al. "Putting humans in the natural language processing loop: A survey." *arXiv preprint arXiv:2103.04044* (2021).

Interaction

Learning from human feedback



Wang, Zijie J., et al. "Putting humans in the natural language processing loop: A survey." *arXiv preprint arXiv:2103.04044* (2021).

Interaction

	TASK					GOAL			INTERACTION							UPDATE		
Work	Text Classification	Parsing and Entity Linking	Topic Modeling	Summarization and Machine Translation	Dialogue and Question Answering System	Model Performance	Model Interpretability	Usability	Mediums – Graphical User Interface	Mediums – Natural Language Interface	User Feedback Type – Binary	User Feedback Type – Scaled	User Feedback Type – Natural Language	User Feedback Type – Counterfactual Example	Intelligent Interaction	Data Augmentation – Offline Model Update	Data Augmentation – Online Model Update	Model Direct Manipulation
Godbole et al. (2004)	●					●			●		●				●	●		
Settles (2011)	●					●			●		●				●	●		
Simard et al. (2014)	●					●			●		●	●				●		
Karmakharm et al. (2019)	●					●		●	●		●					●		
Jandot et al. (2016)	●						●		●		●					●		
Kaushik et al. (2019)	●					●				●			●	●		●		
He et al. (2016)		●				●			●		●					●		
Klie et al. (2020)		●				●		●	●		●					●		
Lo and Lim (2020)		●				●			●		●				●	●		
Trivedi et al. (2019)		●				●			●		●		●			●		
Lawrence and Riezler (2018)		●				●				●	●		●	●		●		
Kim et al. (2019)			●			●		●	●		●						●	
Kumar et al. (2019)			●			●			●		●	●					●	
Smith et al. (2018)			●			●	●	●	●		●	●					●	
Stiennon et al. (2020)				●		●			●			●			●	●		
Kreutzer et al. (2018)				●		●			●			●			●			●
Hancock et al. (2019)					●	●				●			●				●	
Liu et al. (2018)					●	●				●	●		●		●		●	●
Li et al. (2017)					●	●				●	●		●		●	●	●	●
Wallace et al. (2019)					●	●				●	●		●	●		●		

Wang, Zijie J., et al. "Putting humans in the natural language processing loop: A survey." *arXiv preprint arXiv:2103.04044* (2021).

Interaction

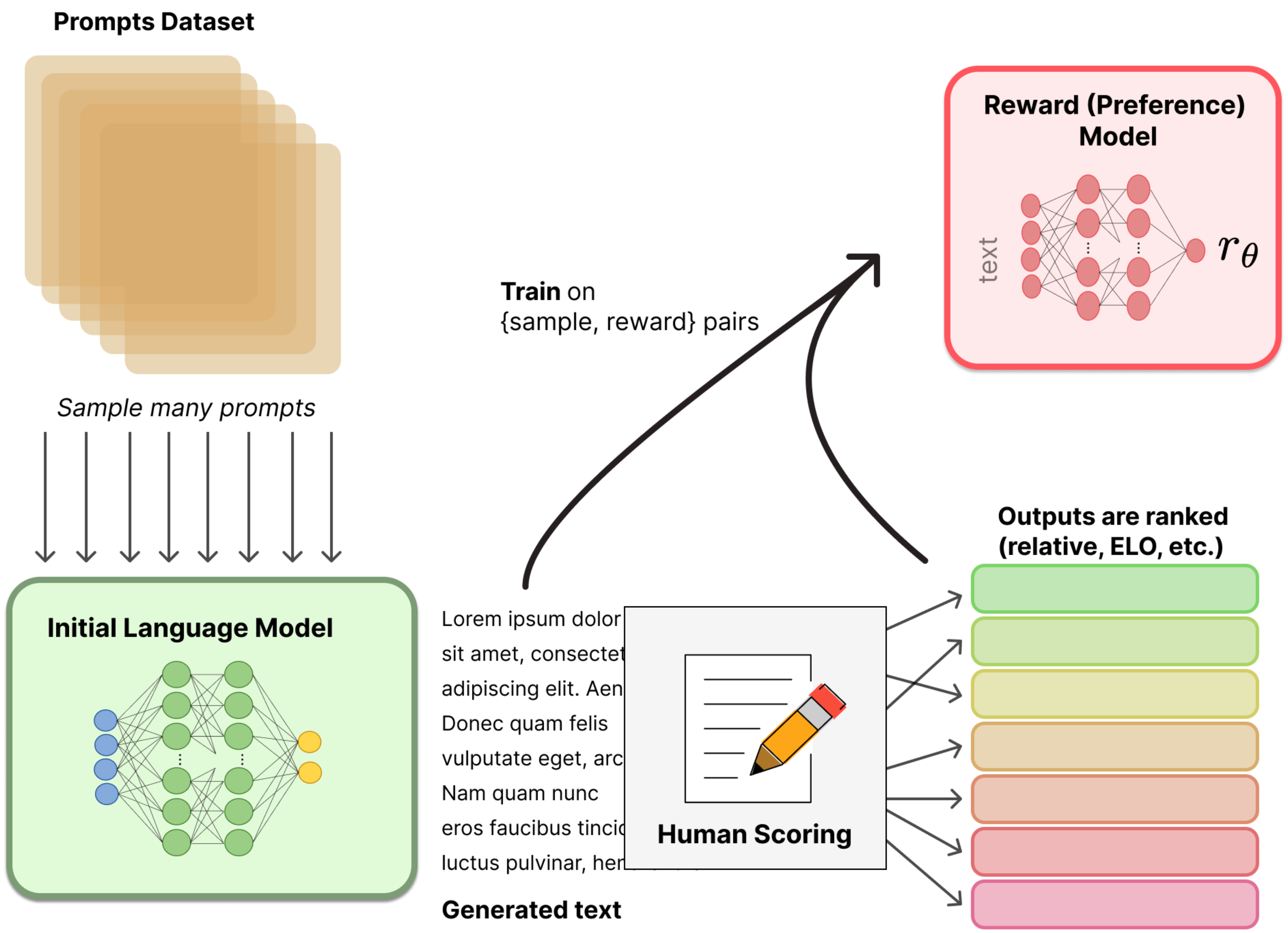
Learning from human feedback

- Making LMs bigger does not inherently make them better at following a user's intent.
  - untruthful, toxic, or simply not helpful to the user?
- InstructGPT
  - Fine-tune GPT-3 with labeler demonstrations of the desired model behavior (supervised learning)
  - Further fine-tune GPT-3 with dataset of rankings of model outputs (reinforcement learning with human feedback)

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Interaction

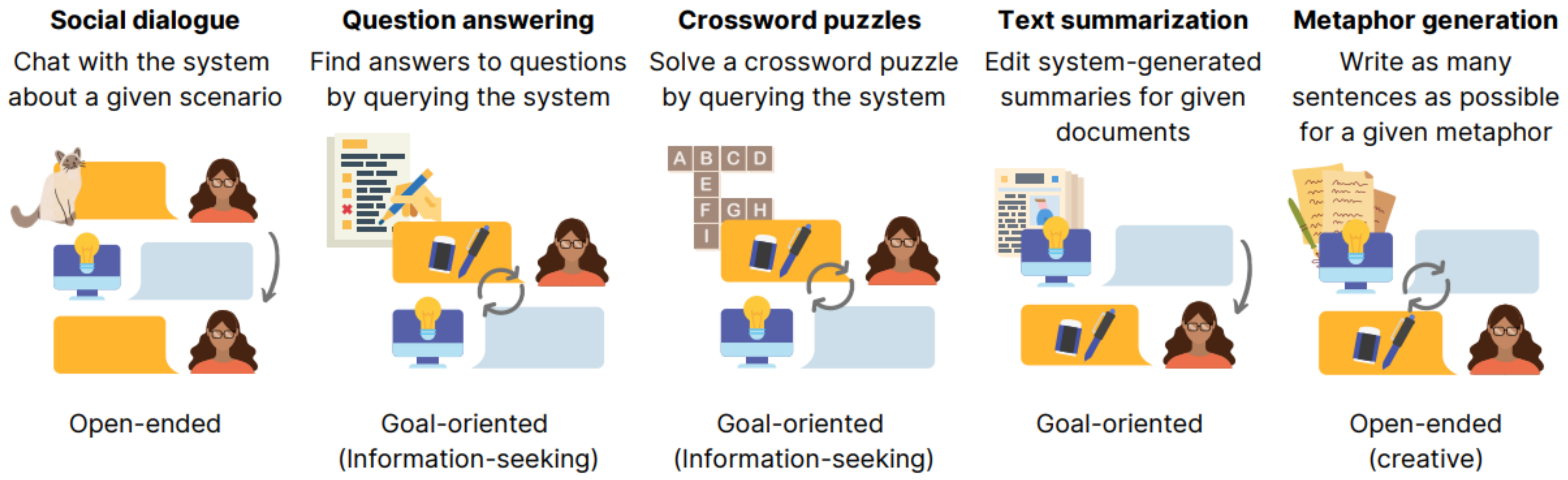
Learning from human feedback



Lambert, et al., "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.

Human Evaluation

Evaluating Interaction



Dimensions			Tasks				
Targets	Perspectives	Criteria	Social dialogue	Question answering	Crossword puzzles	Text summarization	Metaphor generation
Process	First-person	Preference	Reuse	Ease	Enjoyment		Enjoyment
Process	First-person	Quality		Helpfulness	Helpfulness	Improvement	Helpfulness
Process	Third-party	Preference			Queries		
Process	Third-party	Quality		Queries		Edit distance	Queries
Output	First-person	Preference	Interestingness				Satisfaction
Output	First-person	Quality	Specificity	Fluency	Fluency	Consistency	Helpfulness
Output	Third-party	Preference					Interestingness
Output	Third-party	Quality		Accuracy	Accuracy	Consistency	Aptness

Lee, Mina, et al. "Evaluating Human-Language Model Interaction." arXiv preprint arXiv:2212.09746 (2022)

Interaction
Collaboration and Design

- (Natural Language) Interfaces
- Communication of inputs / intermediate / outputs, their visualization
- Model Explanations
- Design choices:
  - name of the model (“GPT-3” vs. “Galactica”) (Khadpe et. al. CSCW)
  - preferences (what is an effective communication? politeness?)

Interaction

Bonus: Conceptual Metaphors

- Khadpe, Pranav, et al. "Conceptual metaphors impact perceptions of human-AI collaboration." *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020): 1-26.

- Stereotype-content model: Warmth vs. Competence
- Warmth: Follows assimilation theory
  - More warmth results in humans responding favorably
- Competence: Follows contrast theory
  - More competence results makes humans not respond favorably

Thank you

Questions?

Icons from https://www.flaticon.com/