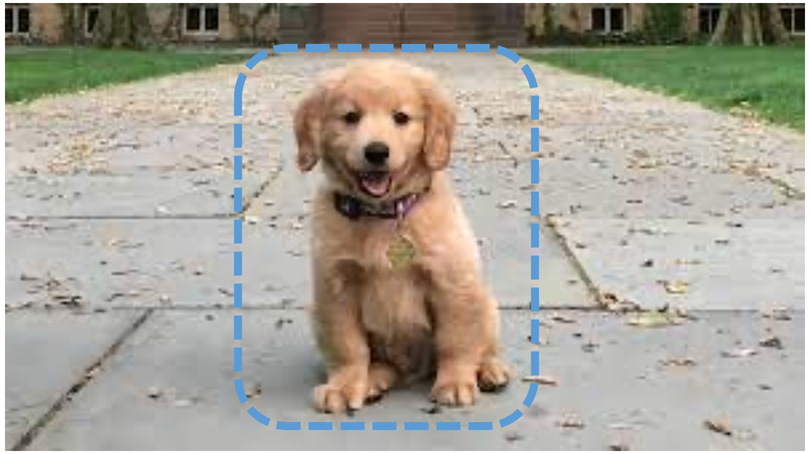
From Perceptron to Cognition: Towards Next Generation Personalized Intelligent Agent

Xiaoyu Shen

Machine Learning Scientist@Amazon Alexa AI

Earlier Intelligent Agent

Image classification



Dog

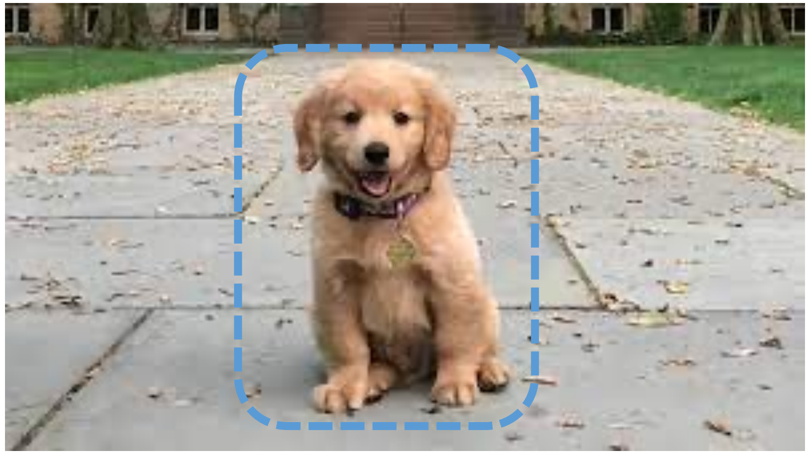
Sentiment analysis

"I love this movie.
I've seen it many times
and it's still awesome."

"This movie is bad.
I don't like it it all.
It's terrible."

Earlier Intelligent Agent: Perceptron

Image classification



Dog

Sentiment analysis

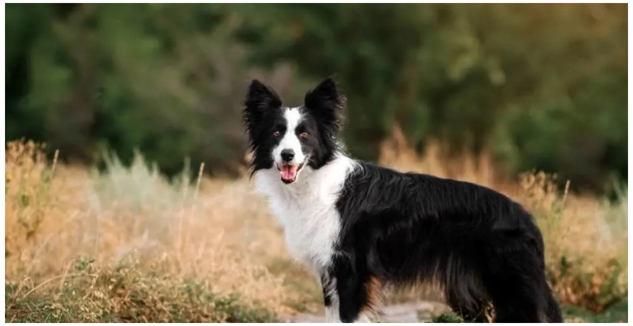
"I love this movie.
I've seen it many times
and it's still awesome."

"This movie is bad.
I don't like it it all.
It's terrible."

- Works only on strictly-defined input-output space
- Narrow expert. Sensitive to domain shifts
- Cannot really “think” or communicate thoughts

Next Generation Intelligent Agent

Here is an image, tell me which animal it is.



It is a dog.

Why do you think so?

Because it has a distinct snout and a prominent nose. The ears are wide and…

Which breed is it?

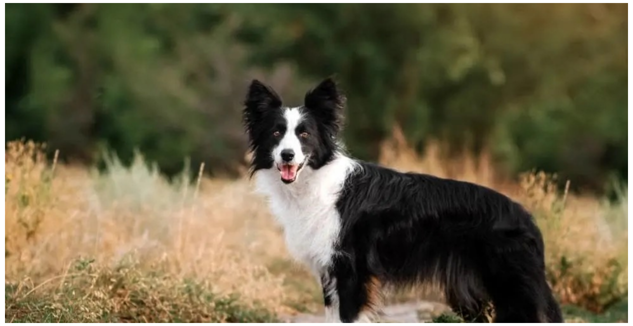
It is a Border Collie, a British breed of herding …

Do you think the dog is happy or not?

The dog seems to be happy because it has a loose, relaxed body posture. The mouth is…

Next Generation Intelligent Agent: Cognition

Here is an image, tell me which animal it is.



It is a dog.

Why do you think so?

Because it has a distinct snout and a prominent nose. The ears are wide and…

Which breed is it?

It is a Border Collie, a British breed of herding …

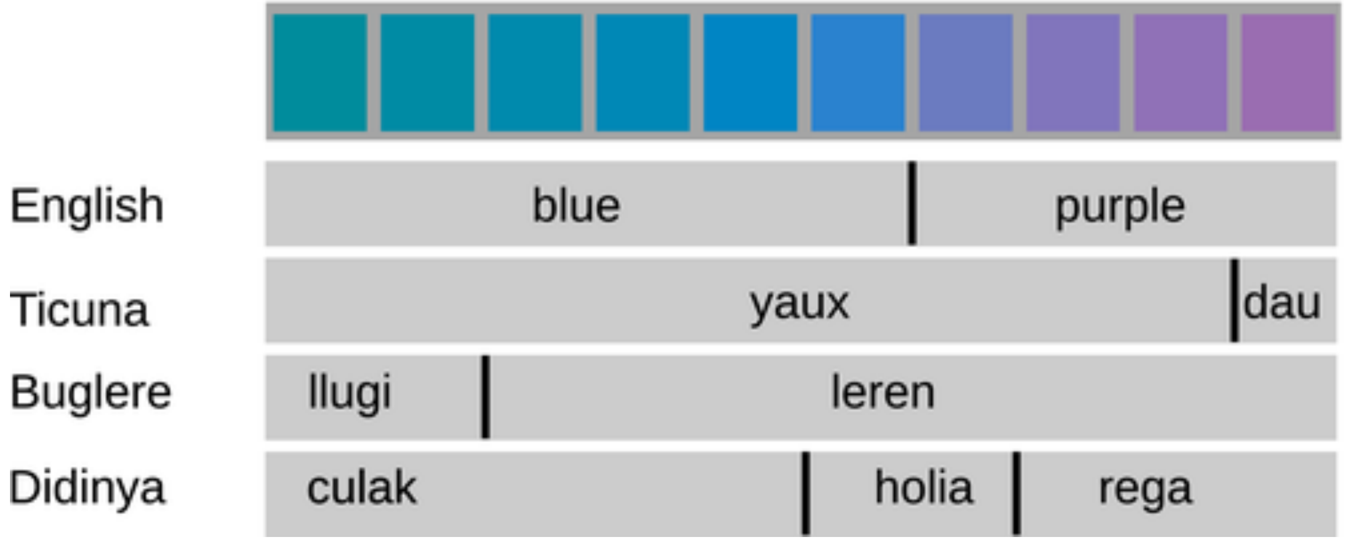
Do you think the dog is happy or not?

The dog seems to be happy because it has a loose, relaxed body posture. The mouth is…

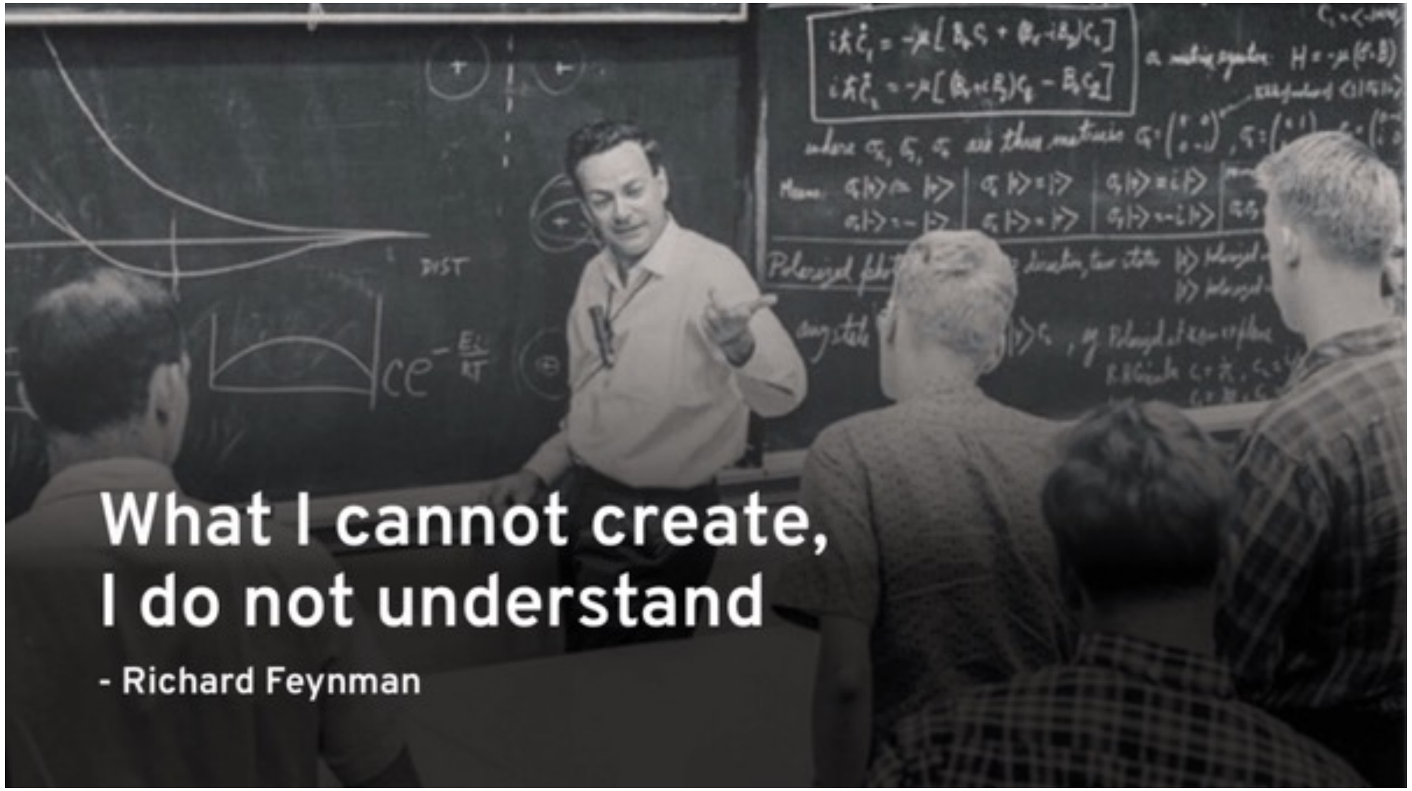
- Can follow various types of human requests
- General expert. Understand the grounded world knowledge
- Can "think" and communicate thoughts

Center For Cognition: Language

- Languages, being the cornerstone of human intelligence, are a foundational tool to formulate and communicate thoughts.
- Linguistic relativity: The structure of a language influences its speakers' worldview or cognition.



Language Model



Language models assign probability distributions to strings. They can be used as language generators to predict next words.

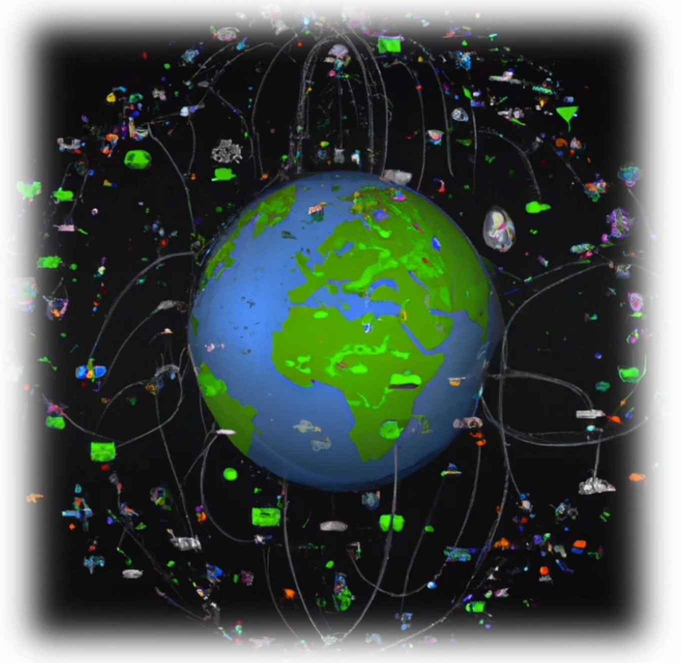
Can you please come here ?

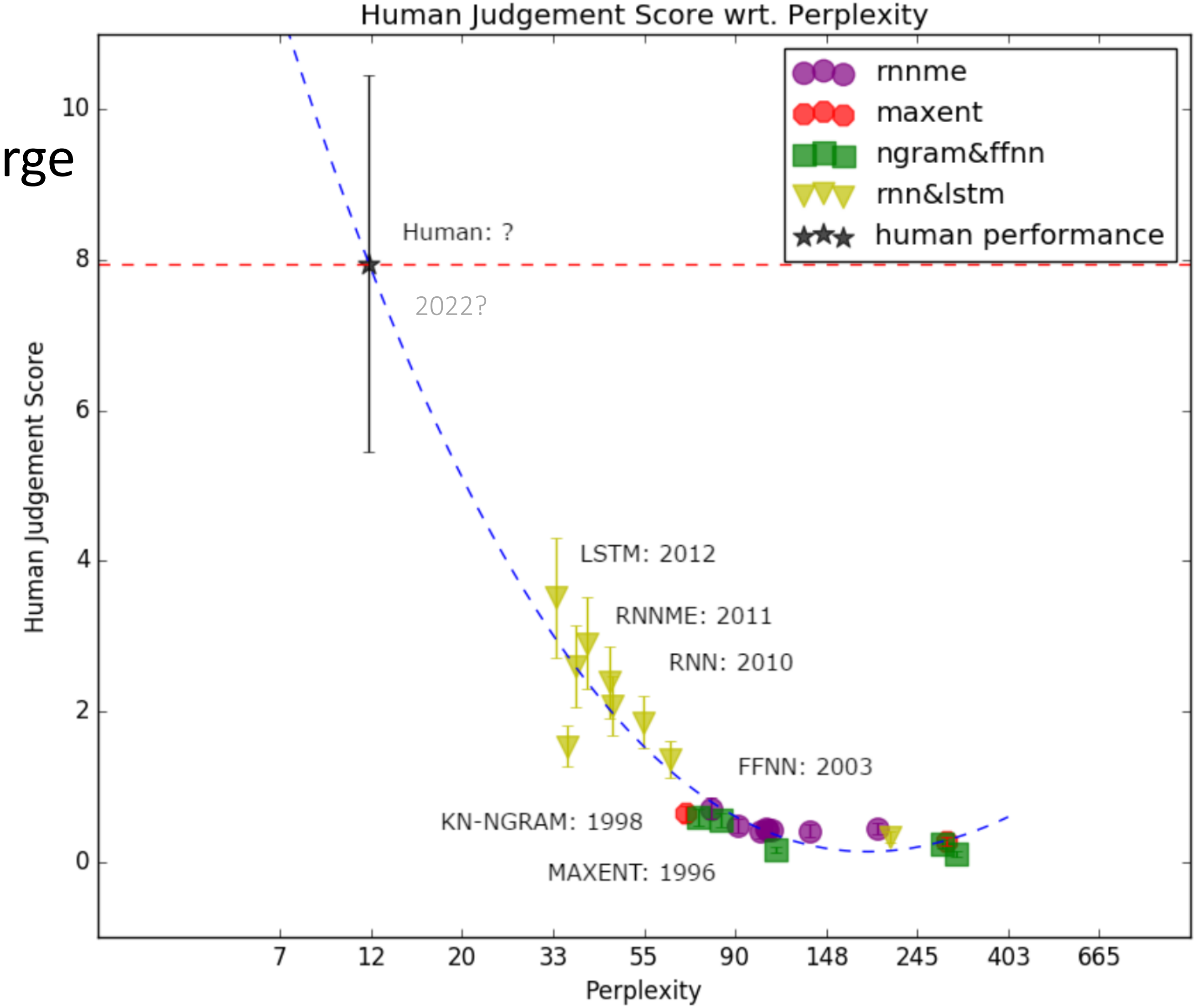
History

Word being predicted

Large Language Model

Large language models are built on large model and large data

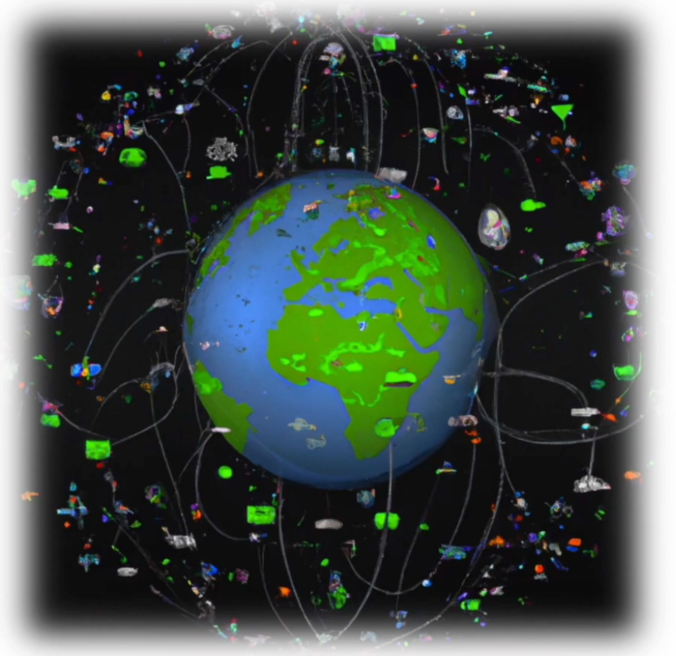


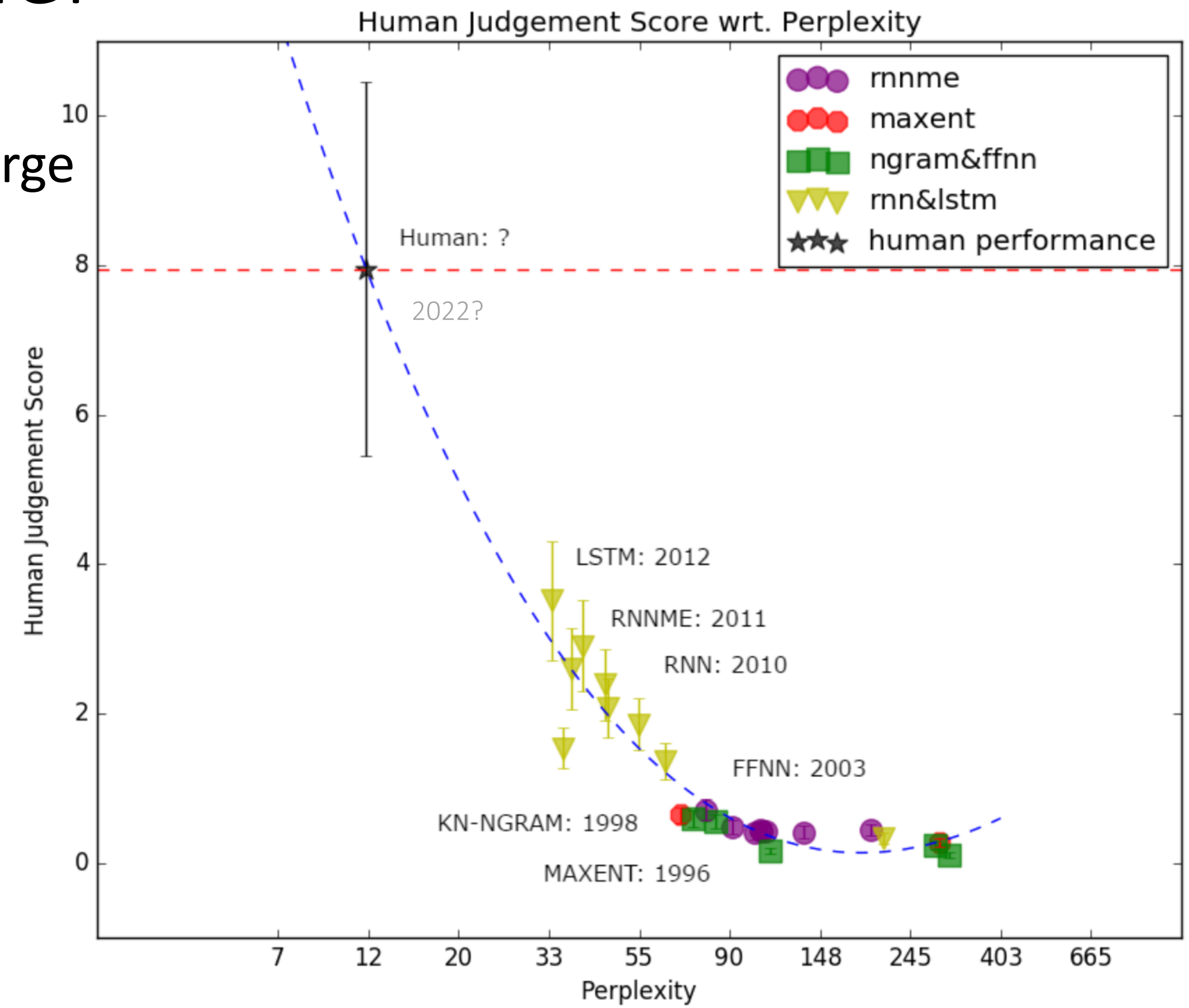


Shen et al. (2017). Estimation of gap between current language models and human performance. In Interspeech 2017

Large Language Model

Large language models are built on large model and large data: Good at mastering language at **linguistic** level but not **cognitive** level



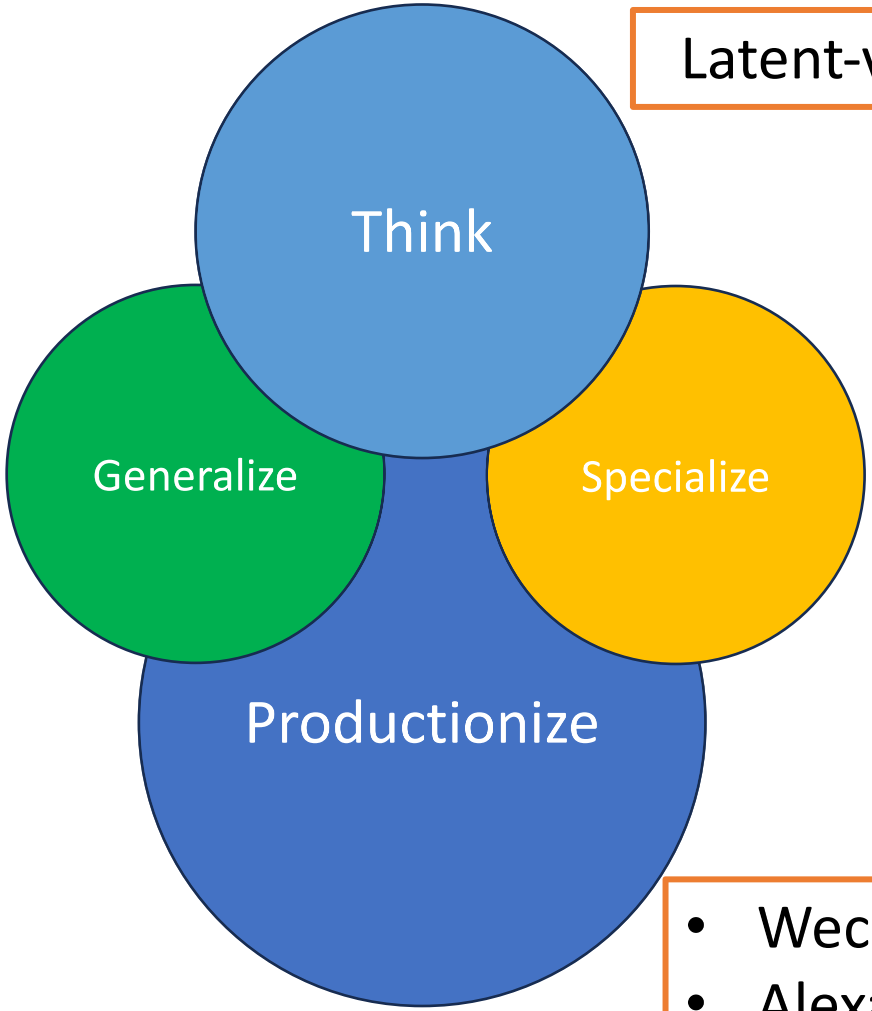


Shen et al. (2017). Estimation of gap between current language models and human performance. In Interspeech 2017

Prior Research: Enable Language Model to

Latent-variable Model

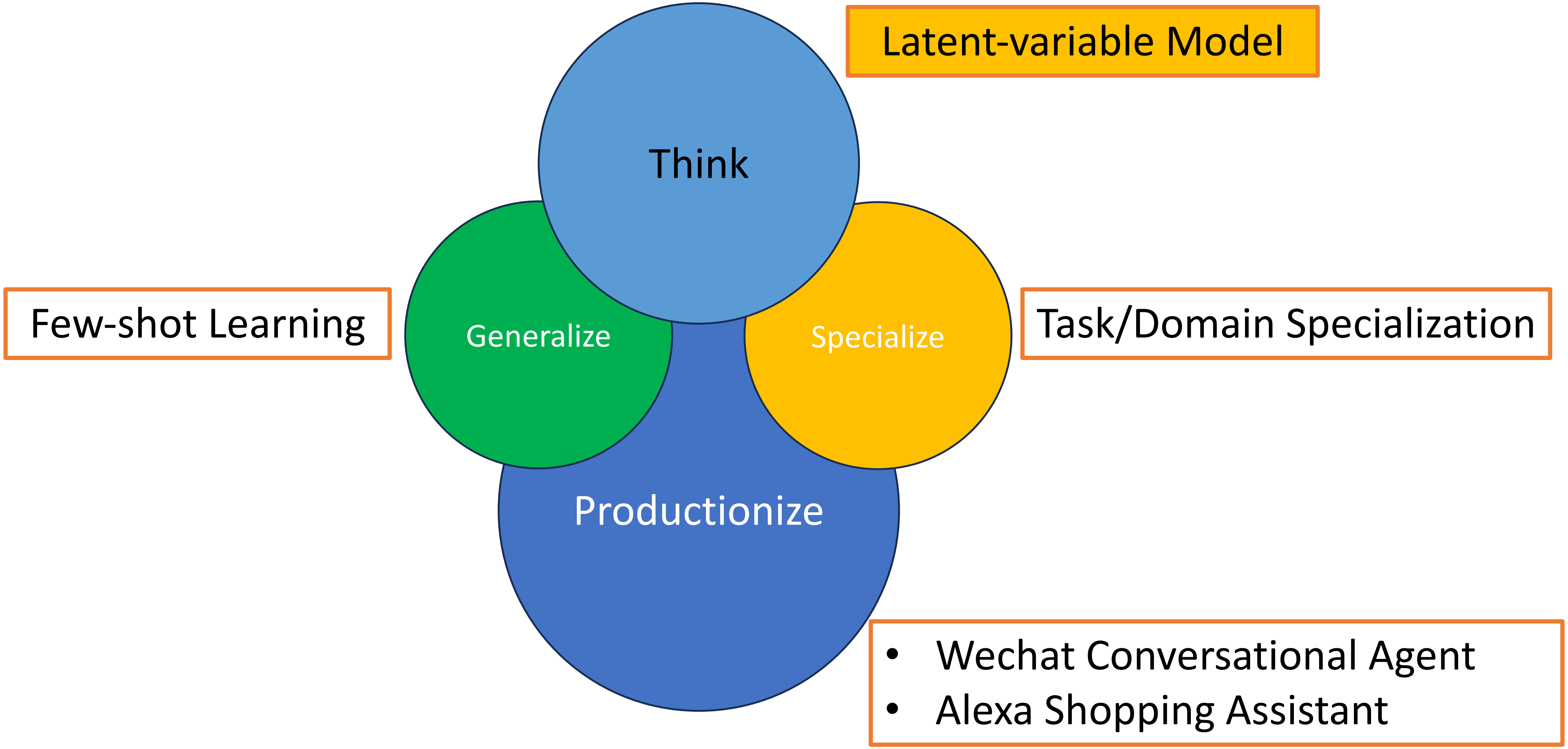
Few-shot Learning



Task/Domain Specialization

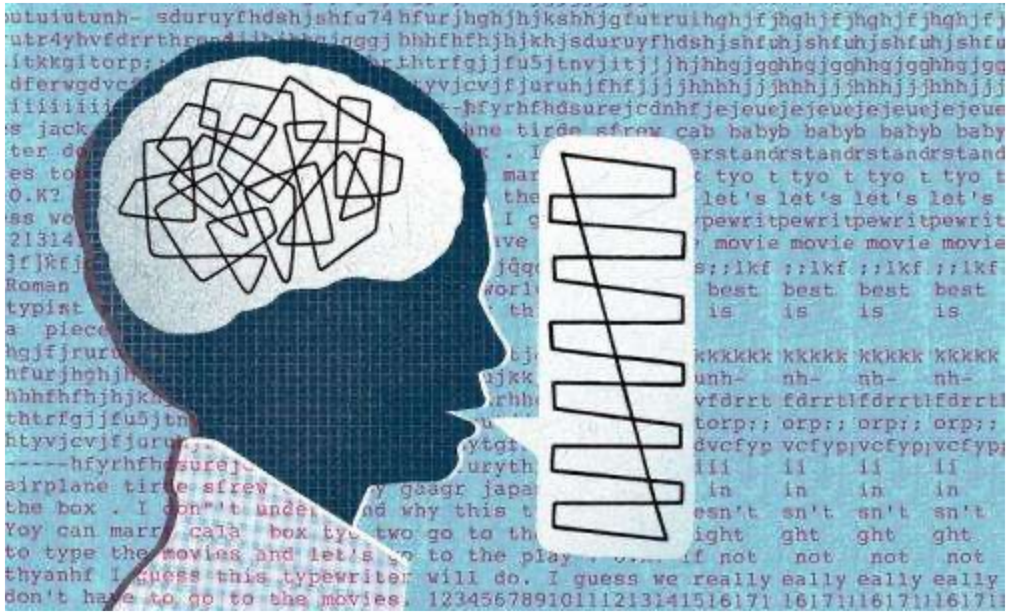
- Wechat Conversational Agent
- Alexa Shopping Assistant

Prior Research: Enable Language Model to



Teach Model to Think

Here is an image, tell me which animal it is.



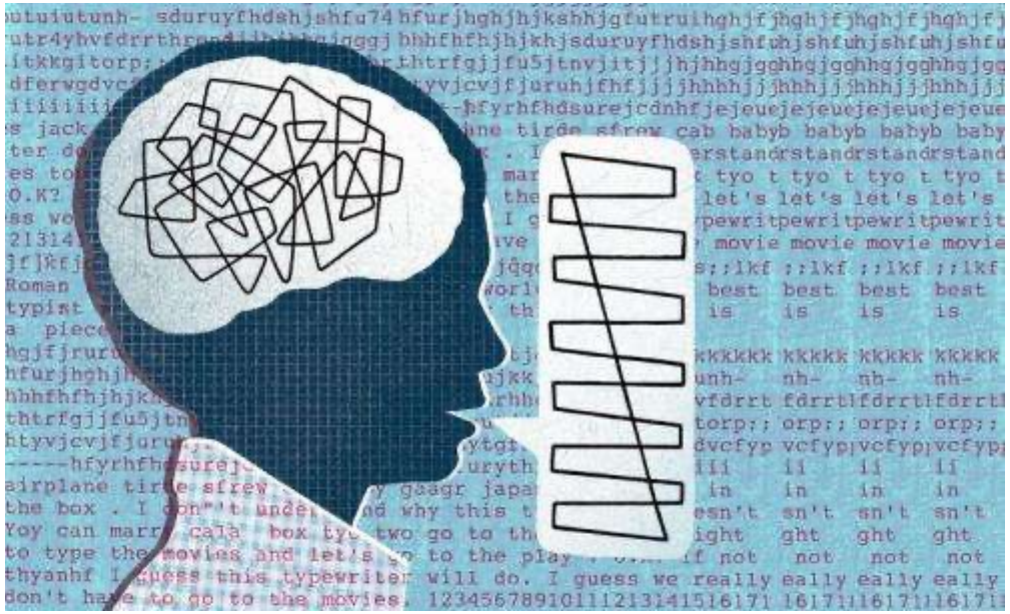
The ears are wide.
The nose is big.
It has yellow furs.
It is opening its mouth.
…

→ It is a dog.

Teach Model to Think

Here is an image, tell me which animal it is.

Latent-variable model is used to infer the latent thinking process without explicitly label these variables



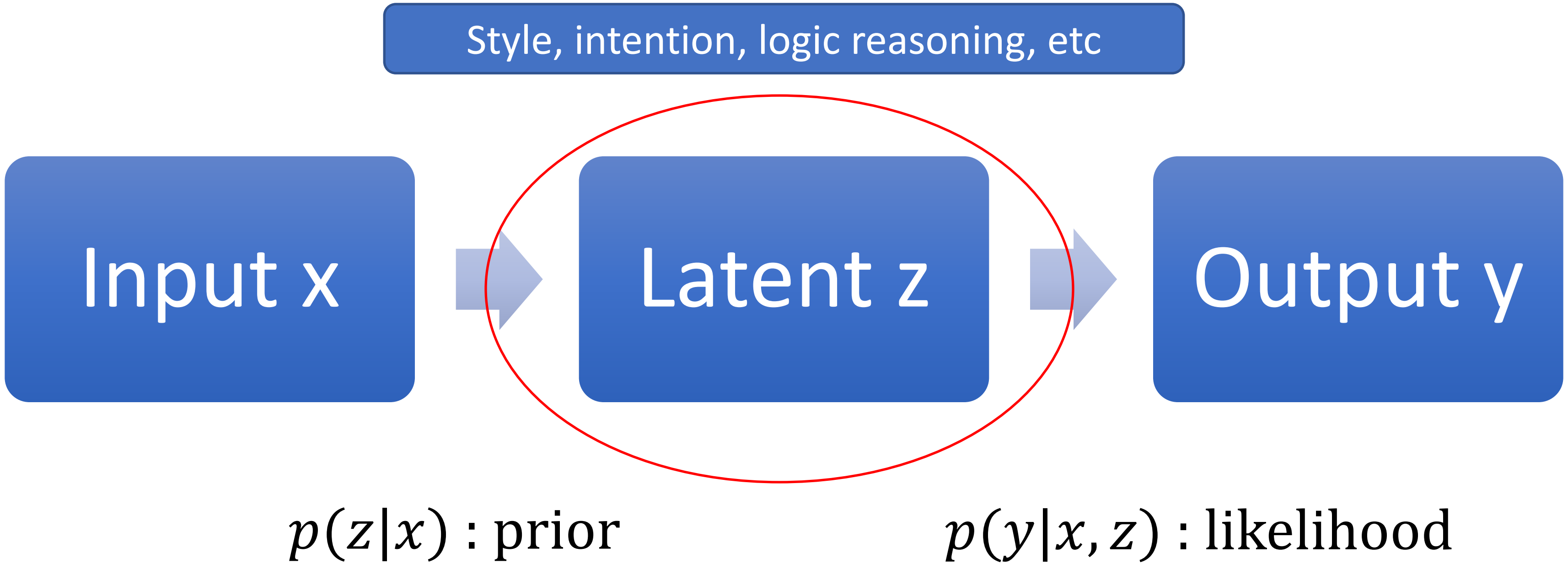
The ears are wide.
The nose is big.
It has yellow furs.
It is opening its mouth.
…

- Latent
- Stochastic

→ It is a dog.

Latent-variable Model for Text Generation

- Latent-variable model adds an intermediate variable z between the input x and output y.
- The model learns both the prior $p(z|x)$ and likelihood $p(y|x,z)$.



Optimization Challenge

- **Non-latent-variable model**: Easy to optimize, supervised learning via maximum likelihood for $p(y|x)$
- **Latent variable model**: Hard to optimize. As z is unknown, we don't have the label to optimize $p(z|x)$ and $p(y|x, z)$

	Labelling cost	Training stability	Performance	Applicability
Human label	Very high	Very good	Very good	Apply to some
Pseudo label	high	Very good	Not good	limited
Marginalization	low	good	good	limited
Variational inference	low	Very bad	good	Apply to all

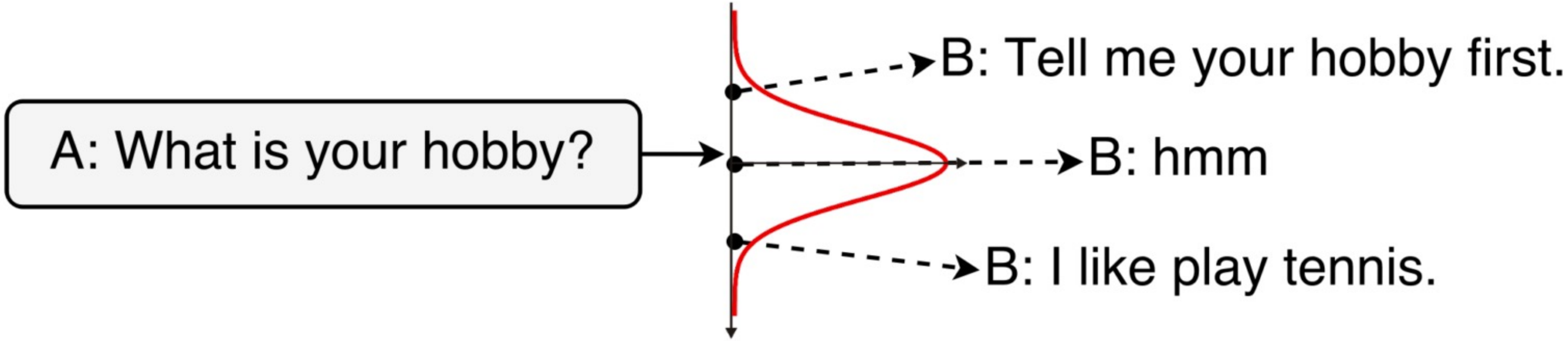
No free meal. Every method has pros and cons.

Shen (2021). Deep Latent-Variable Models for Neural Text Generation. PhD Thesis

Latent Variable in Dialogue Generation

- First sample a latent variable given the context, then generate a response based on it.
- P(z|x) follows a Gaussian distribution with learned mean/variance.

x **z~p(z|x)** **y~p(y|x,z)**



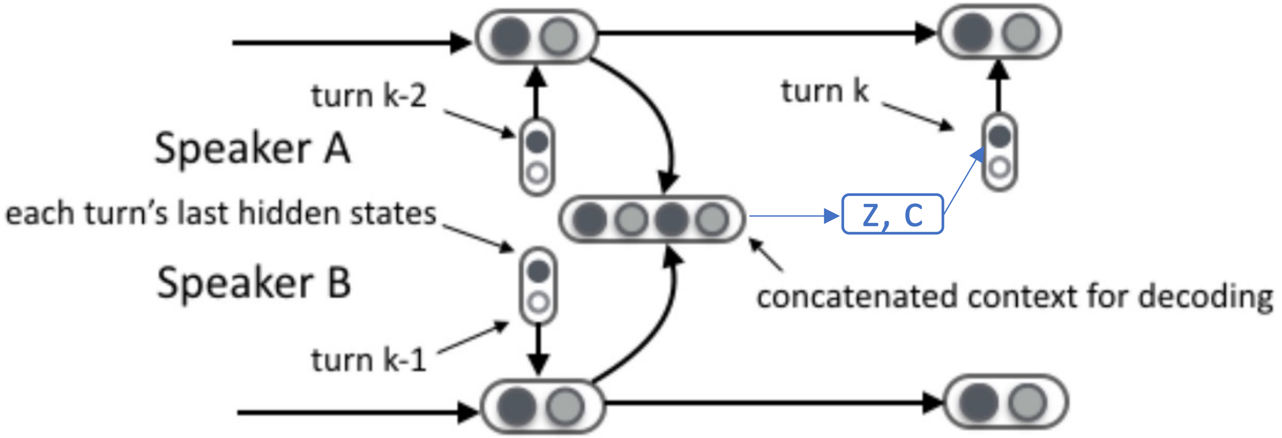
Shen et al. (2017). A Conditional Variational Framework for Dialog Generation. In ACL 2017

Optimization: Variational Inference

Maximize a lower bound of the marginal likelihood p(y|x)

$$\mathbb{E}_{z\sim \underbrace{q(z|x,y)}_{\text{Posterior}}} \underbrace{p(y|x,z)}_{\text{likelihood}} - \mathrm{KL}(\underbrace{q(z|x,y)}_{\text{Posterior}} \| \underbrace{p(z|x)}_{\text{prior}})$$

- **The KL term tends to collapse to 0. Assign a smaller weight to it.**
- **Propose a bilevel hierarchical encoder to capture user-specific attributes**
- **Combine a discrete latent variable c to indicate sentiment**

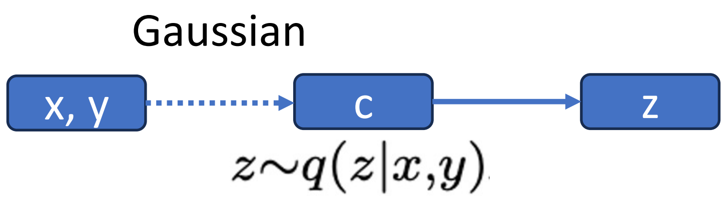


Improved Optimization

- We show that the objective is competition of two items, which is biased to utilizing or ignoring latent variables.

$$\mathbb{E}_{z\sim q(z|x,y)} p(y|x,z) - \mathbb{KL}(q(z|x,y)|p(z|x))$$
$$= \mathbb{E}_{z\sim p(z|x)} p(y|x,z) - \mathbb{KL}(q(z|x,y)|p(z|x,y))$$

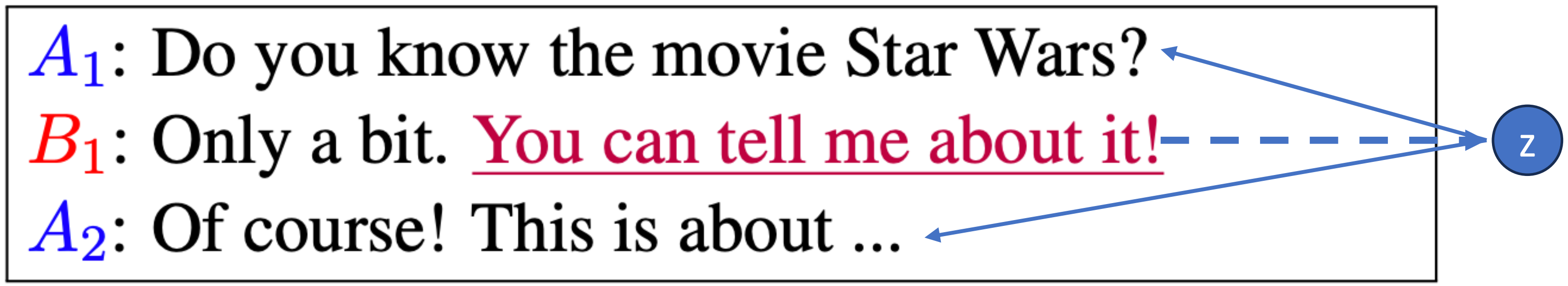
- We propose an iterative training pipeline to reduce KL-collapsing
- We prove that this new training pipeline leads to a tighter lower bound of the marginal likelihood



Shen et al. (2018). Improving variational encoder-decoders in dialogue generation. In AAAI 2018

Connecting the Preceding and Following

The latent variable should share high mutual information with both the preceding dialogue context and the following responses



The actual mutual information is not computable, we propose a way to approximate a lower bound of mutual information.

Shen et al. (2018). NEXUS Network: Connecting the Preceding and the Following in Dialogue Generation. In EMNLP 2018

Performance with/without Latent Variable

what are you going to do this sunday ?

W.o: Oh,that's great!
W.: Nothing really , what's up ?

Let's go for a drink !

W.o: Thanks!
W.: Sounds good . Where are we going ?

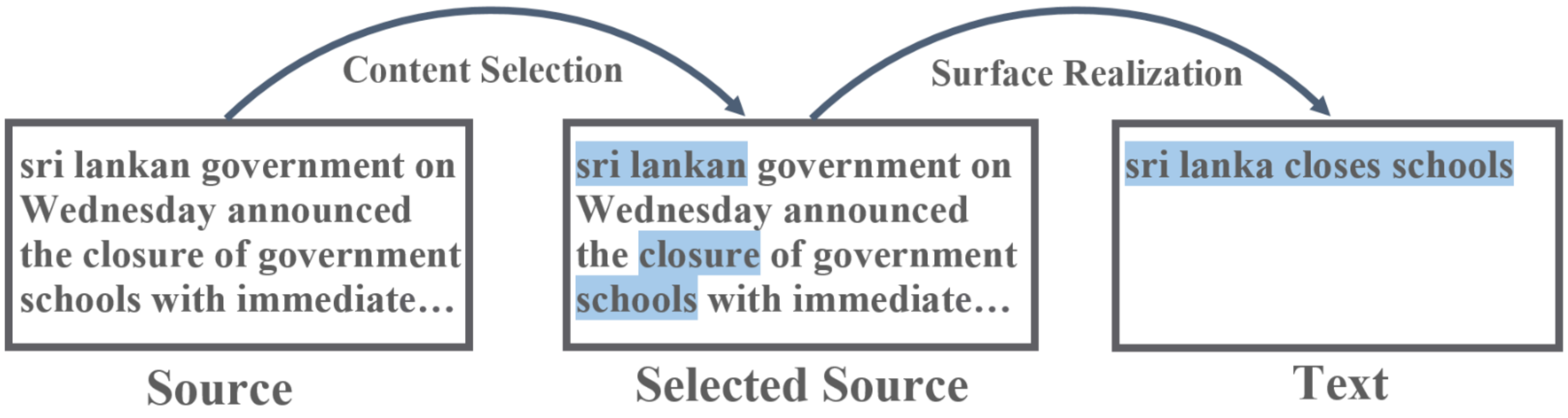
You won't want to miss our webinar tomorrow !

W.o: I think that's a good question.
W.: Thanks for your invitation! I'm free !

Latent variables help language models to think more, producing more meaningful responses.

Latent Variable as Content Selection

Latent variable z indicates if each word should be selected.



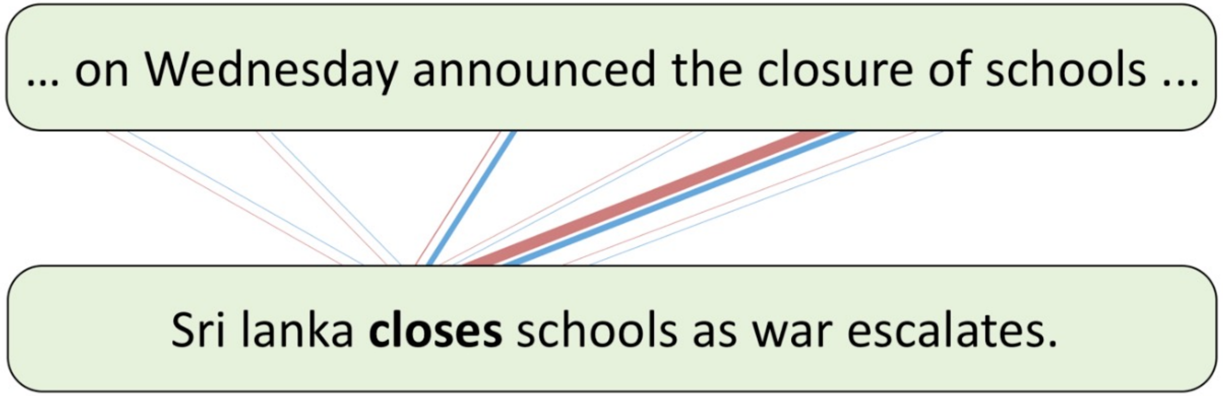
$$
\begin{aligned}
&\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta)\\
&= \log \mathbb{E}_{\beta\sim q_\phi} \frac{p_\theta(Y,\beta|X)}{q_\phi(\beta)}\\
&\geq \mathbb{E}_{\beta\sim q_\phi} \log \frac{p_\theta(Y,\beta|X)}{q_\phi(\beta)}\\
&= \mathbb{E}_{\beta\sim q_\phi} \log p_\theta(Y|X,\beta) - KL(q_\phi||\mathbf{B}(\gamma))
\end{aligned}
$$

Optimization: Importance sampling + reinforcement learning

Shen et al. (2019). Select and Attend: Towards Controllable Content Selection in Text Generation. In EMNLP 2019

Latent Variable as Word Alignment

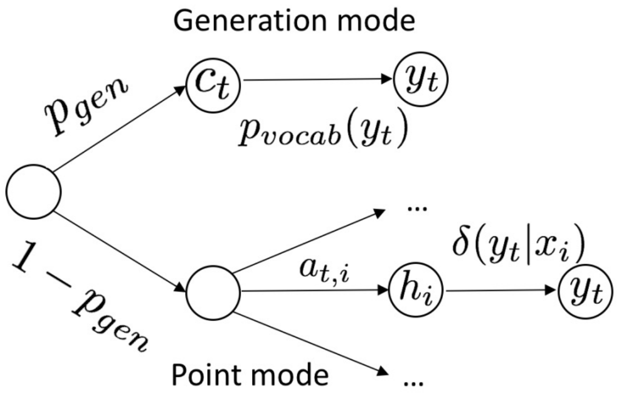
Source: The sri lankan government on Wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country.



Aligned to "announced":
Sri lanka announces closure of government schools.
Sri lanka declares closure of government schools.
Aligned to "closure":
Sri Lanka closes government schools.
Sri lanka shuts down government schools

Blue: prior
Red: posterior

Optimization: Top-k approximation



Shen et al. (2019). Improving Latent Alignment in Text Summarization by Generalizing the Pointer Generator. In EMNLP 2019

Latent Variable as Segment Alignment

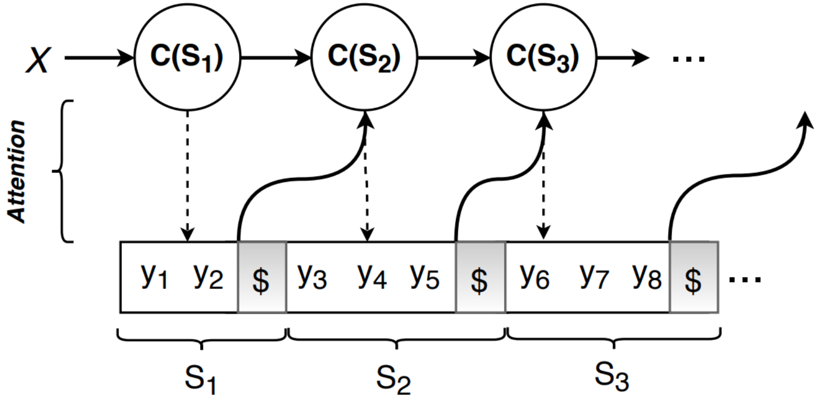
Latent variable z indicates segmentation and alignment.

Source data:

Name[Clowns], PriceRange[more than £30], EatType[pub], FamilyFriendly[no]

Generation:

①Name → ②(Clowns)
③FamilyFriendly → ④(is a child-free)
⑤PriceRange → ⑥(, expensive)
⑦EatType → ⑧(pub.)



Optimization: semi-markov model + forward/backward algorithm

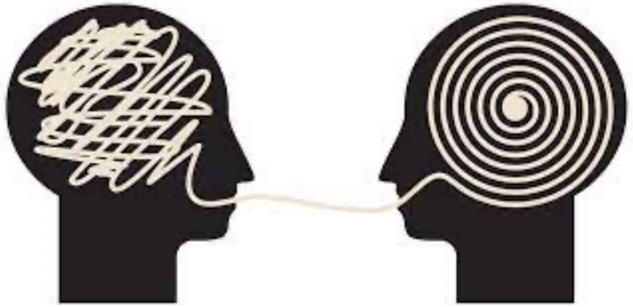
Shen et al. (2020). Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence. In ACL 2020

Summary: Latent Variable As Thinking Process

- Human thoughts are a complex combination of continuous and discrete latent variable.
- Latent variables are used to mimic the thinking process, whose distributions can be defined based on downstream tasks.

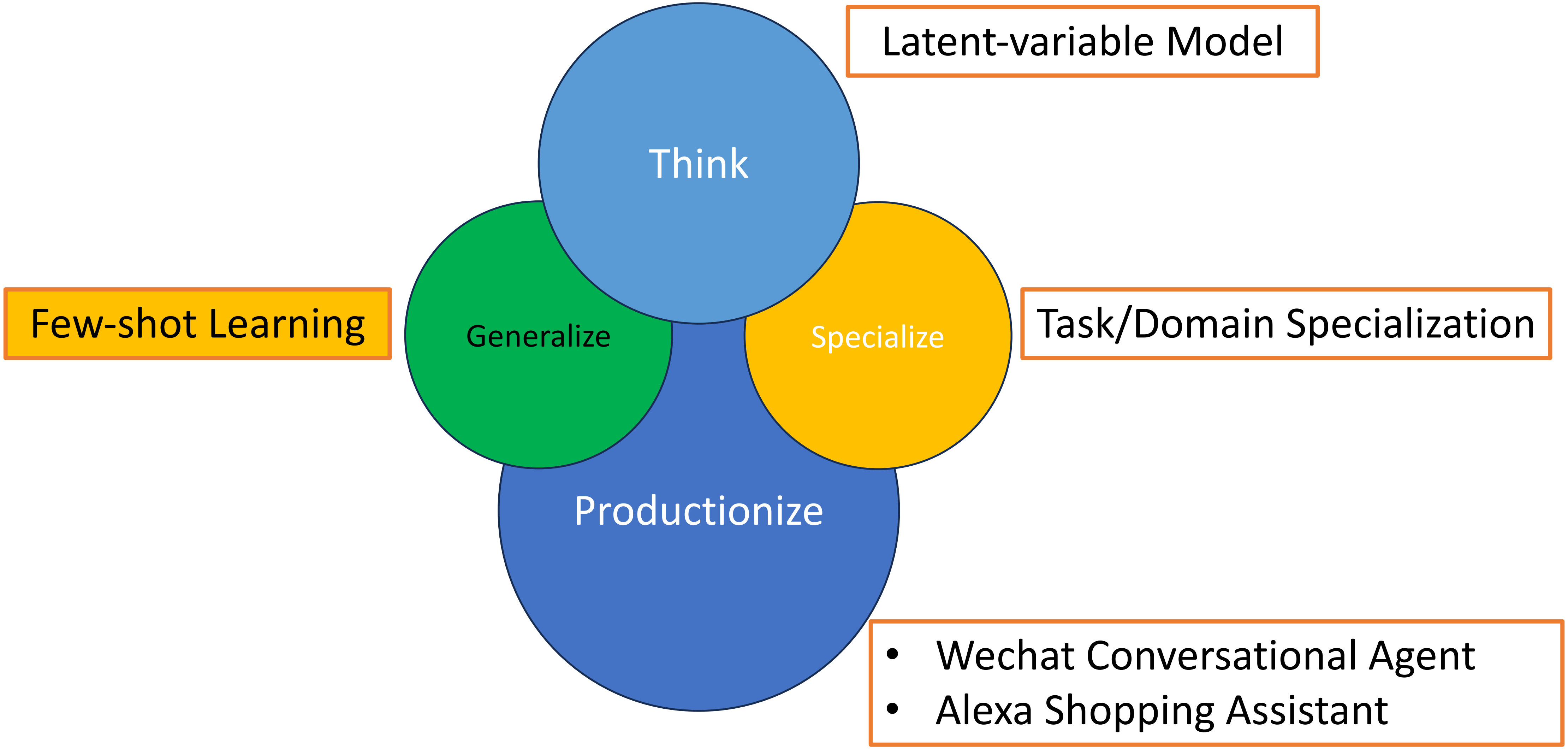
Challenges:

- Definition of latent variables
- Optimization of latent variables
- General representation of thoughts?



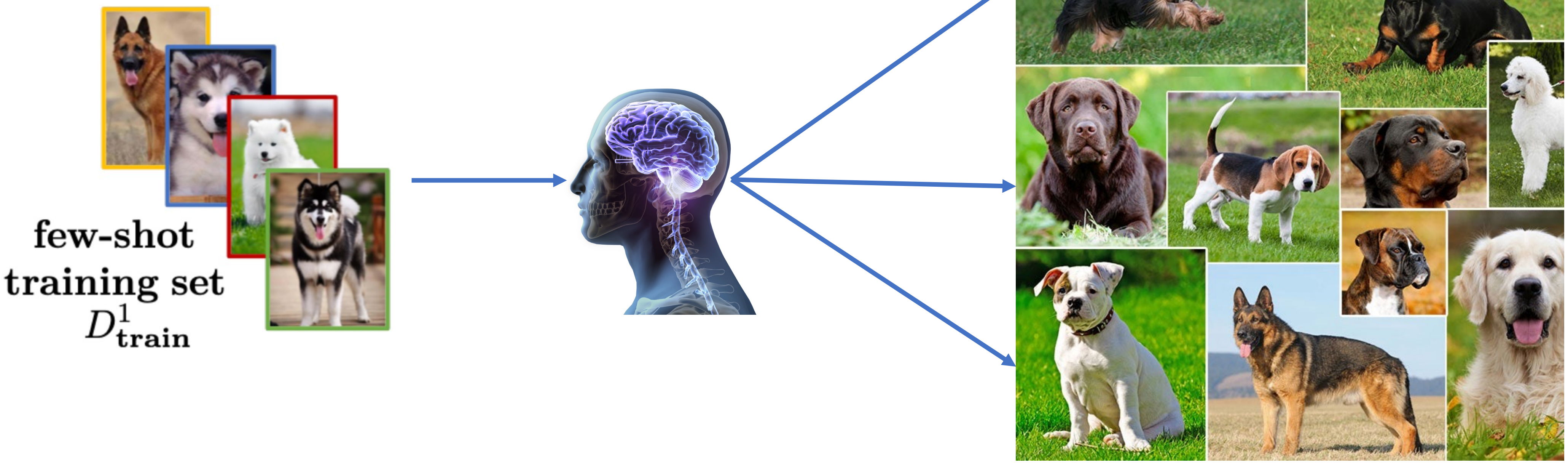
Josh won against Lio.	`(condition (won-against '(josh) '(lio)))`
Josh proceeded to claim victory against Alex.	`(condition (won-against '(josh) '(alex)))`
Even working as a team, Lio and Alex still could not beat Josh.	`(condition (not (won-against '(lio alex) '(josh)))`

Prior Research: Enable Language Model to



Teach Model to Generalize

- Human intelligence is able to to generalize from very few examples
- Few-shot learning teaches models to generalize from few training examples.



Data Augmentation

- Image augmentation can be easily done via truncating, rotating, resizing, changing brightness, adding white noise, etc
- Text Augmentation is much harder than image augmentation

Dog

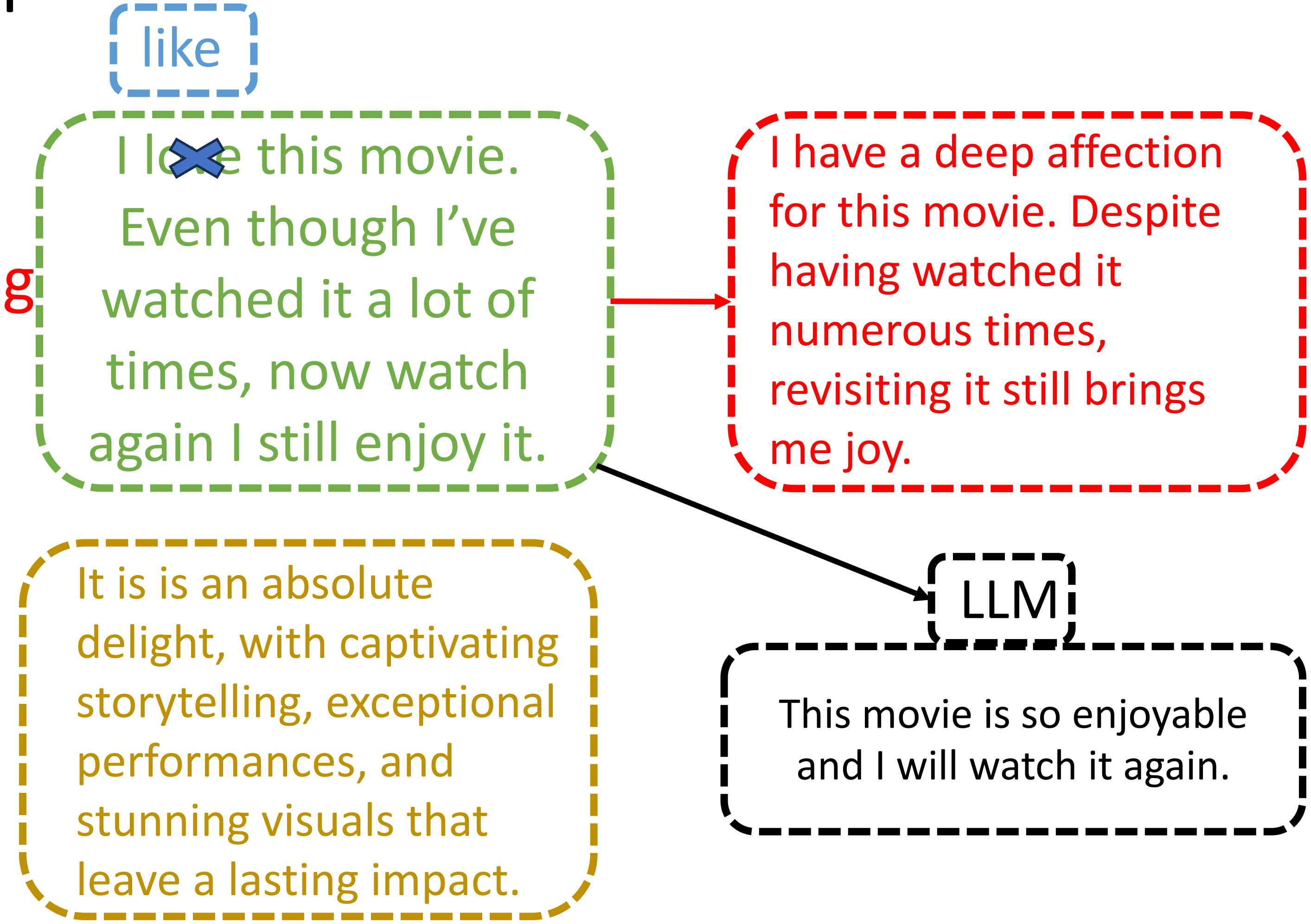
I love this movie. Even though I've watched it a lot of times, now watch again I still enjoy it.

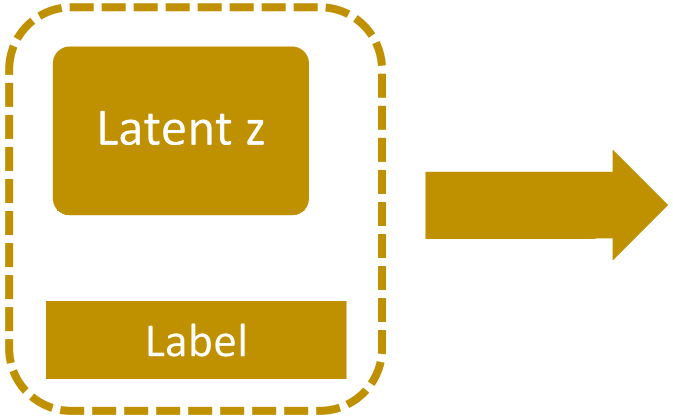
Positive

?

Text Augmentation

- Word-level replacement
- Sentence-level paraphrasing
- Latent-variable process
- Augmentation from LLM





Xu et al. (2020). Data Augmentation for Multiclass Utterance Classification–A Systematic Study. In COLING 2020
Qiu et at. (2020). EasyAug: An Automatic Textual Data Augmentation Platform for Classification Tasks. In WWW 2020
Chang et al. (2021). Neural Data-to-Text Generation with LM-based Text Augmentation. In EACL 2021
Su et al. (2022). Rocbert: Robust chinese bert with multimodal contrastive pretraining. In ACL 2022

Weak Supervision

Weak supervision leverages noisy labels from “weak” sources such as crowd-sourcing, heuristics, knowledge bases, augmentations, etc

I love this movie. Even though I’ve watched it a lot of times, now watch again I still enjoy it.

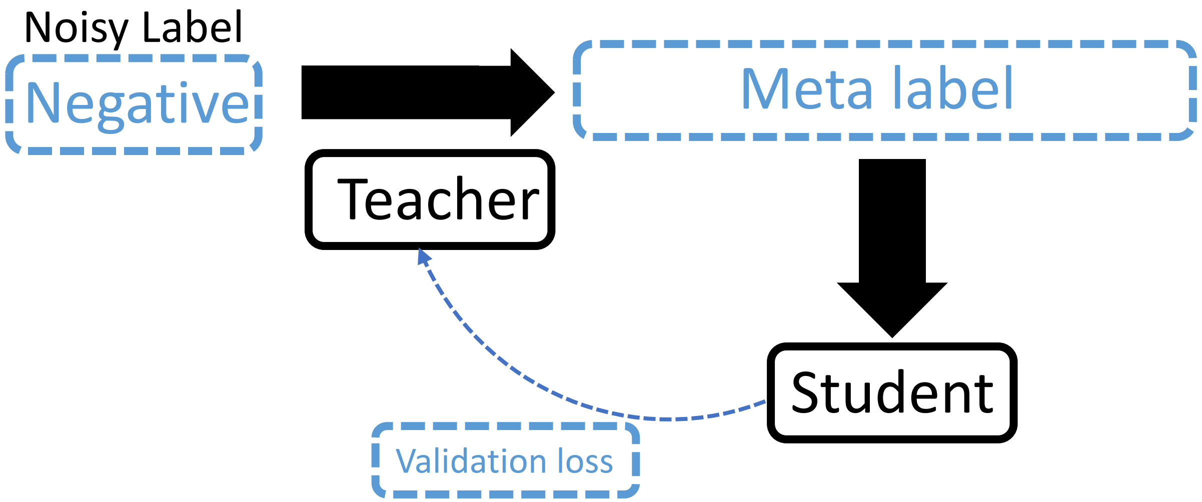
Noisy Label

Negative

Weak Supervision

Weak supervision leverages noisy labels from “weak” sources such as crowd-sourcing, heuristics, knowledge bases, augmentations, etc

I love this movie. Even though I’ve watched it a lot of times, now watch again I still enjoy it.



Zhu et al. (2023). Meta Self-Refinement for Robust Learning with Weak Supervision. In EACL 2023

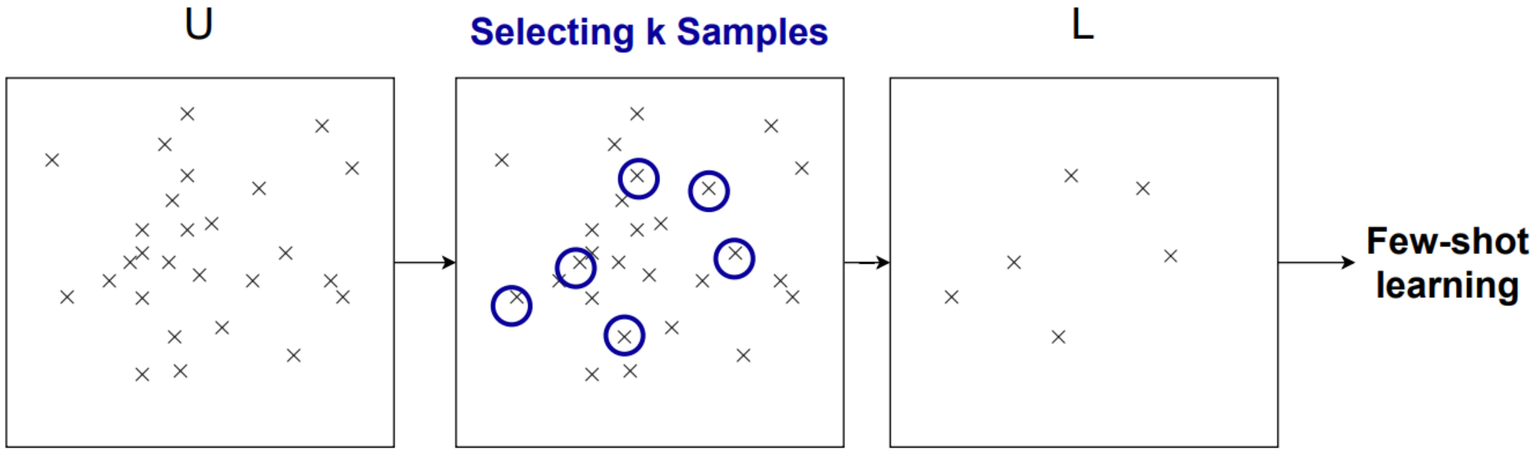
Weak Supervision

- Existing weak supervision techniques overestimated the performance in that they "under-used" clean data. Their advantage mostly vanishes when as few as 10 clean labels are available.
- Given the same budget, should we use it to get a few clean labels, or apply weak supervision on 1000× more noisy data?
- When the noisy labels come from human heuristics, we should rather spend it to get clean labels.

Zhu et at. (2023). Weaker Than You Think: A Critical Look at Weakly Supervised Learning. In ACL 2023 **Special theme paper award*

Data Selection

Apart from passively using existing dataset, we can also actively select which data to annotate and how to annotate

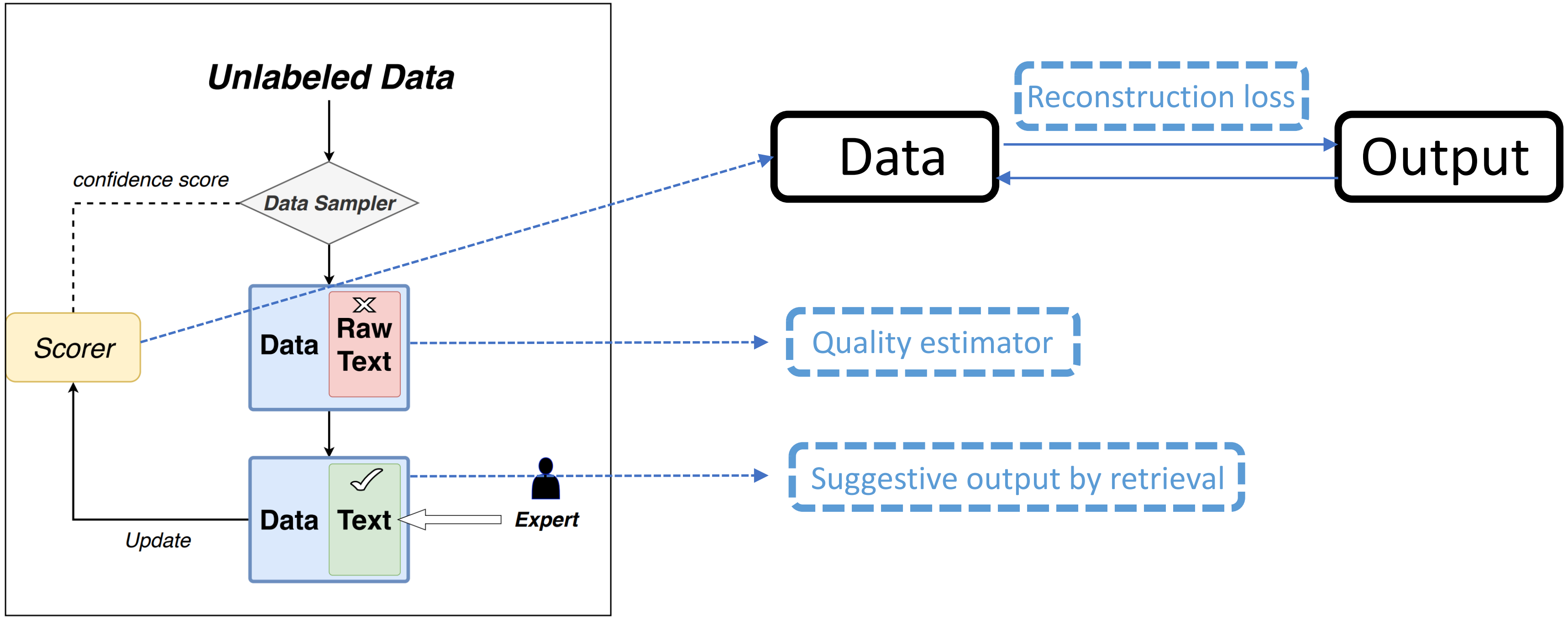


$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{i,j} ||x^i - \mu^j||_2^2$$

Shen et at. (2021). On Training Instance Selection for Few-Shot Neural Text Generation. In ACL 2021

Iterative Selection

An end-to-end pipeline which iteratively selects informative data points to annotate, suggest annotation and estimate quality.



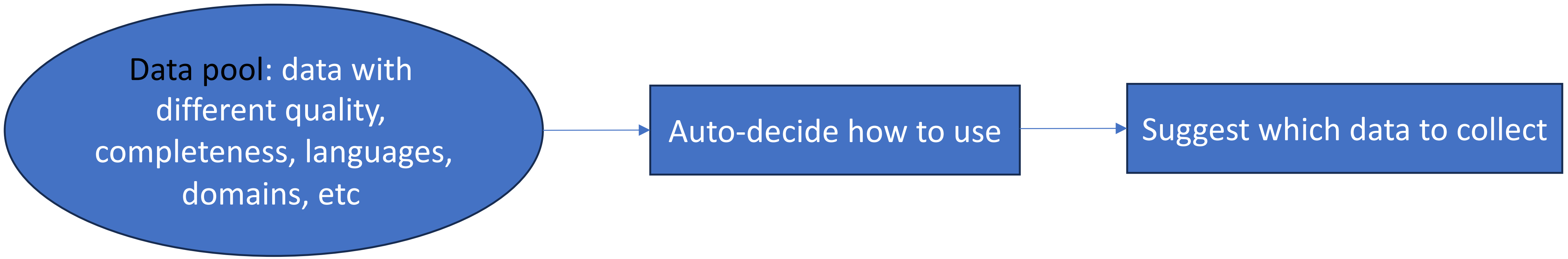
Chang et al. (2020). DART: A Lightweight Quality-Suggestive Data-to-Text Annotation Tool. In COLING 2020 **Best demo paper award*

Summary: Generalize From Few-shot Samples

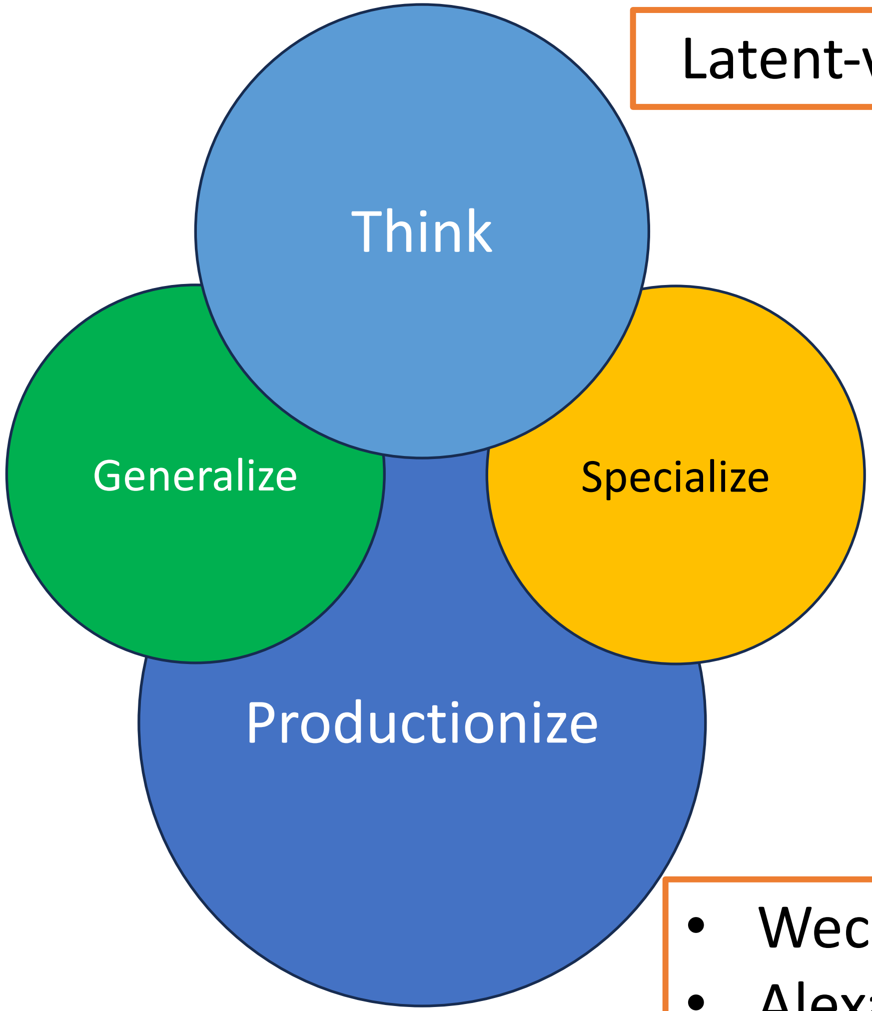
1. Select most informative samples to annotate (**data selection**)
2. Find most efficient way to annotate them (**annotate assistance**)
3. Augment the annotated data (**data augmentation**)
4. Train on augmented + clean annotations (**weak supervision**)

Challenges:

- Which kinds of augmentation/weak labels are useful?
- How to choose between annotate and augment?
- Is there a principled way to perform these four steps?



Prior Research: Enable Language Model to



Latent-variable Model

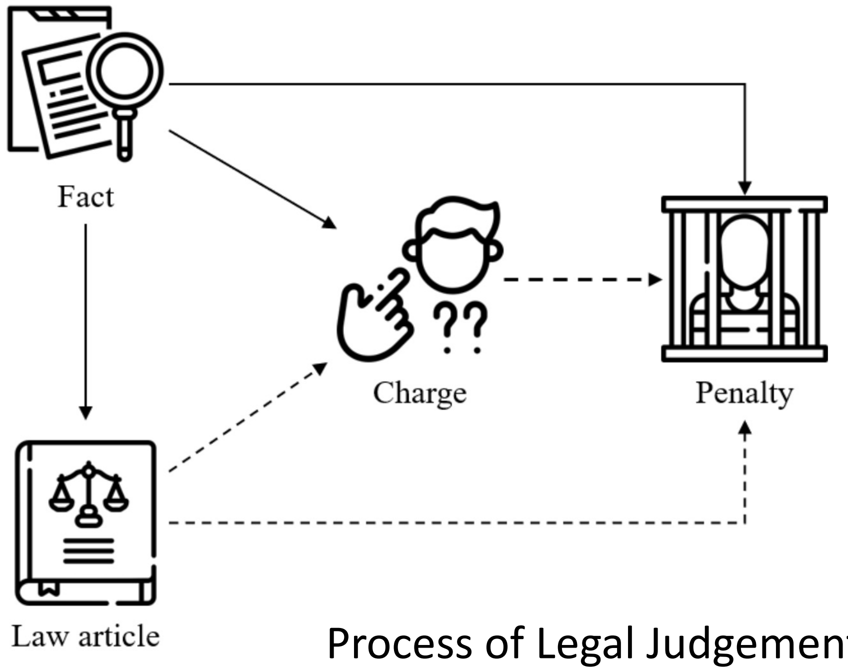
Few-shot Learning

Task/Domain Specialization

- Wechat Conversational Agent
- Alexa Shopping Assistant

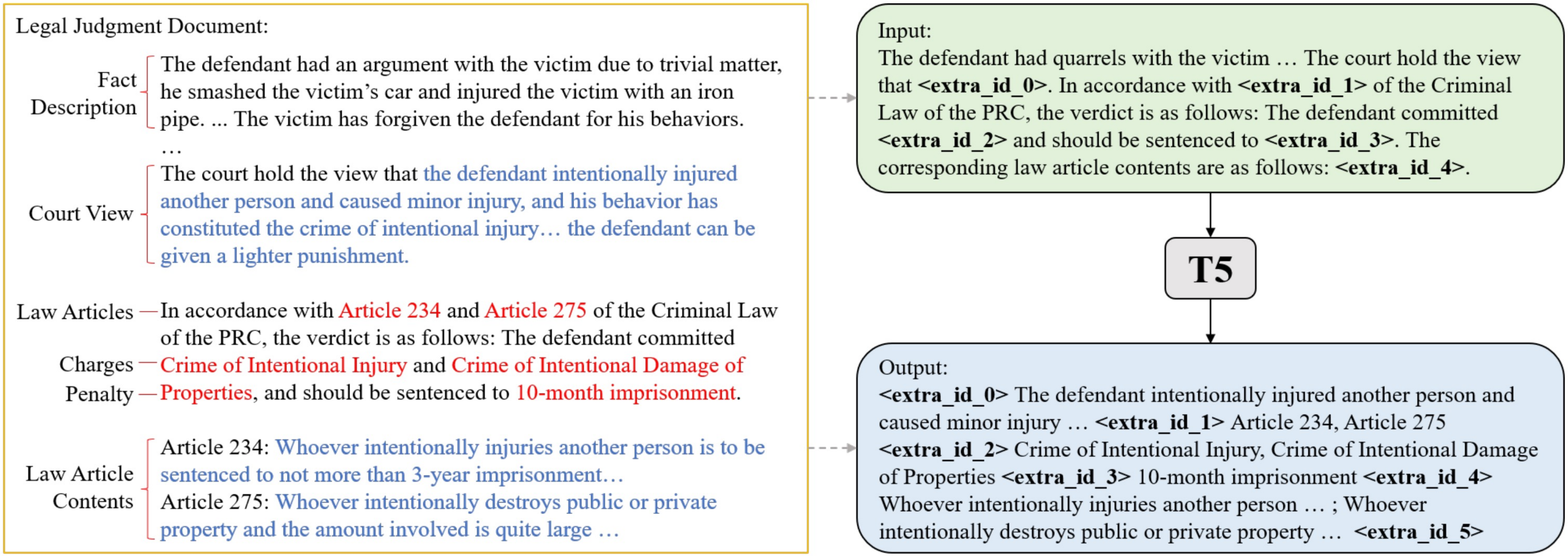
Teach Model to Specialize

- New tasks/domains require specialized knowledge
- Task/domain specialization aims to improve the performance by using these task/domain-specific knowledge



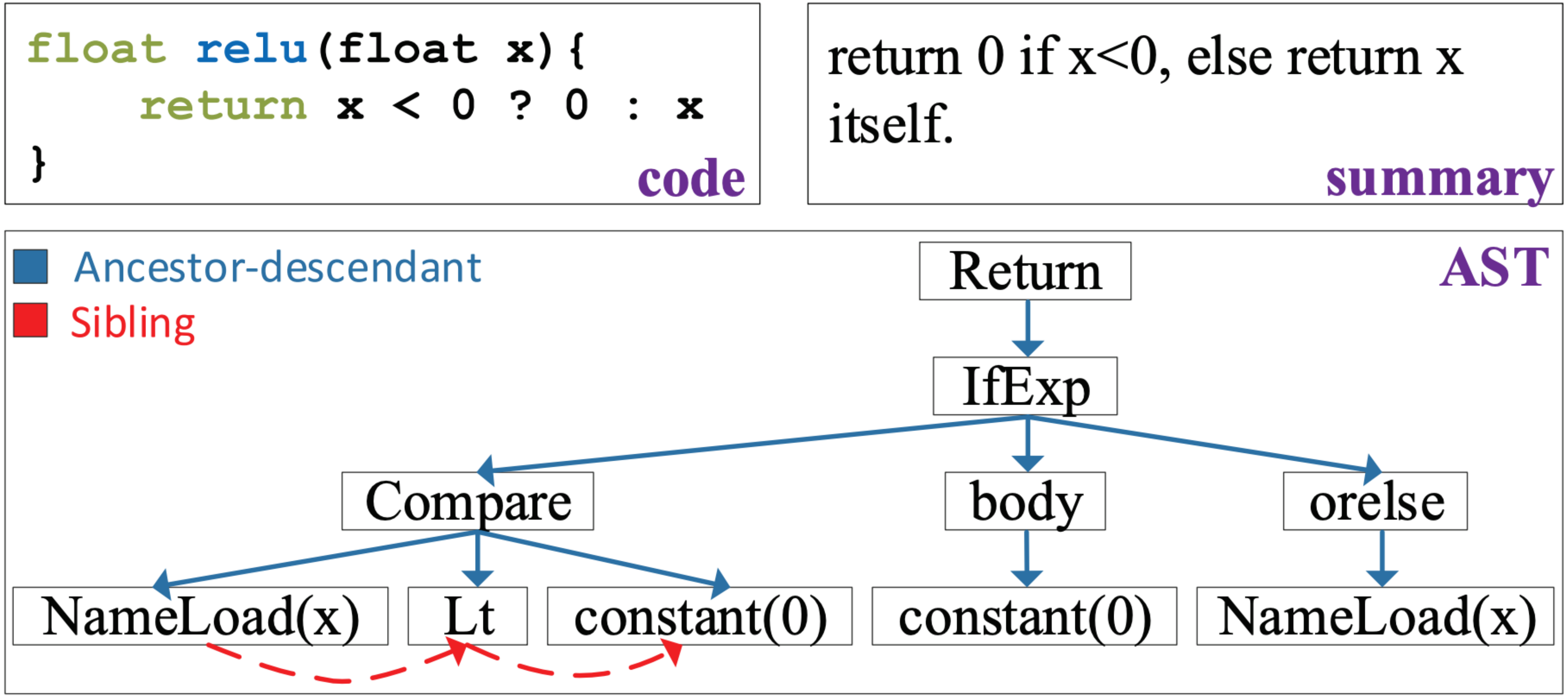
Process of Legal Judgement in China

Specialize to Legal Judgement Prediction



Ge et al. (2021). Learning Fine-Grained Fact-Article Correspondence in Legal Cases. In TASLP
Huang et al. (2021). Dependency Learning for Legal Judgment Prediction with a Unified Text-to-Text Transformer. In TASLP

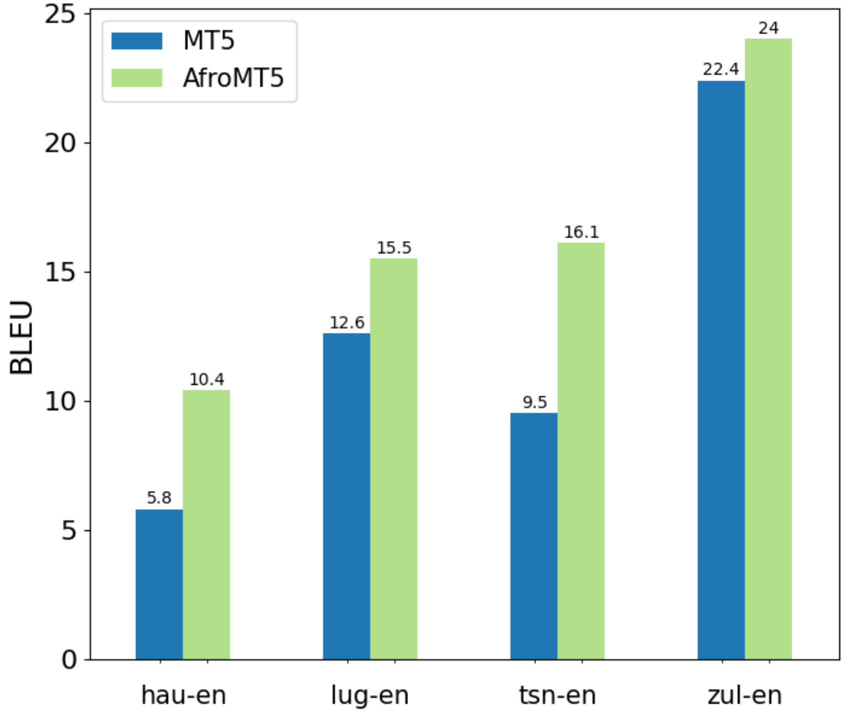
Specialize to Structured Code

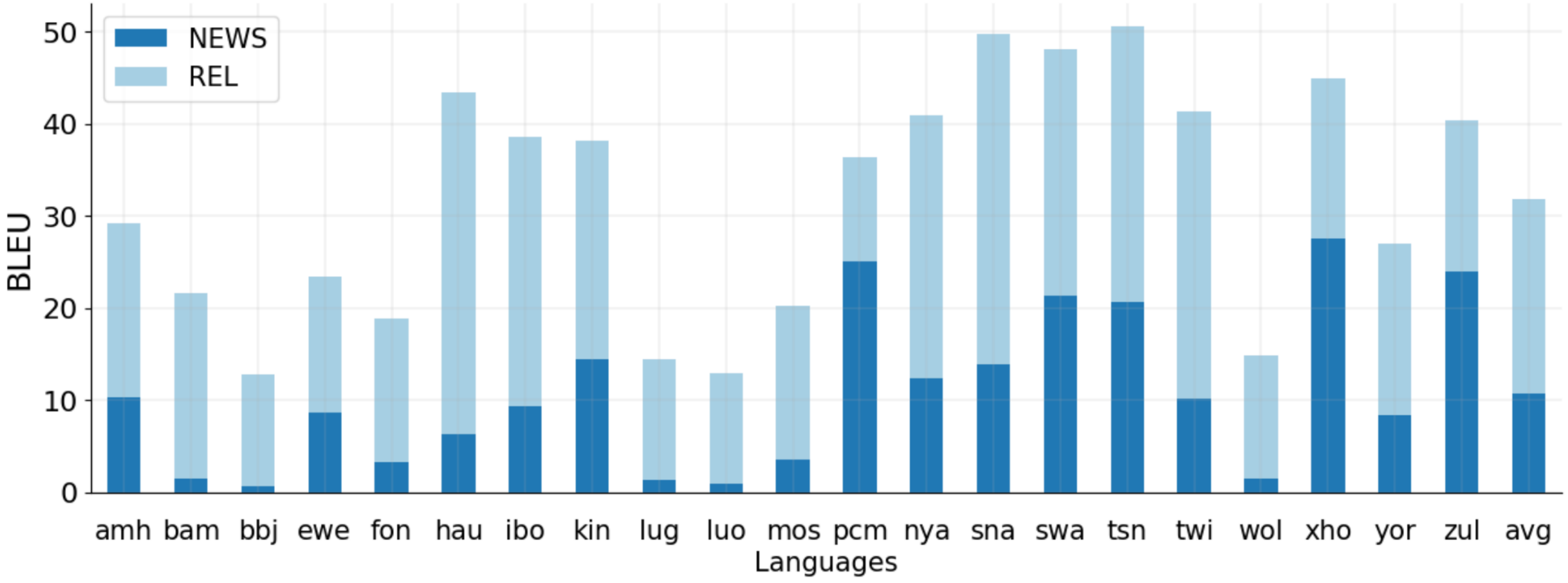


Tang et al. AST-trans: code summarization with efficient tree-structured attention. In ICSE 2022

Specialize to African News Translation

Pre-train on web text → Pre-train on African language text → Fine-tune on bible translation → Fine-tune on news translation





Adelani et al. A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation. In NAACL 2022

Summary: Specialize to New Tasks/Domains

- Specialize to new decision process and dependency relations
- Specialize to new input/output structure
- Specialize to new langauge and text domain

Challenges:

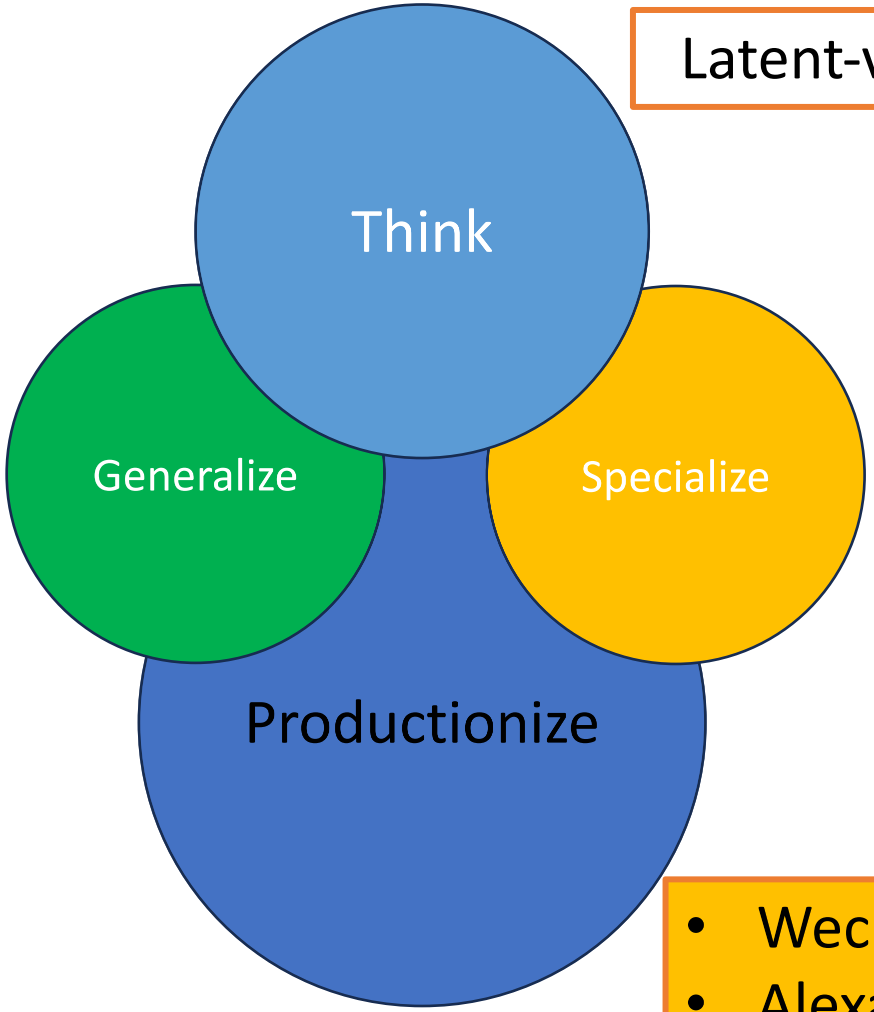
- Design proper model architecture and loss depending on the task
- Alignment of vocabulary and grammar
- Can we train agents the same way as humans? Learn from textbook and manuals?

Get into Law School → Study to get a JD Degree → Pass Bar Exam → Practice

Prior Research: Enable Language Model to

Latent-variable Model

Few-shot Learning



Task/Domain Specialization

- Wechat Conversational Agent
- Alexa Shopping Assistant

Wechat Dialogue Agent: Goal

- Chat with humans in the movie domain
- User engagement is the most important goal

A: 威尔史密斯演技真的很棒 (Will Smith's acting skill is really good).
B: 他的当幸福来敲门太经典了 (His The Pursuit of Happyness is a classic).
A: 一直都挂在电影排行榜靠前的位置 (That' always among top ranked movies).
B: 嗯嗯，这部电影真的很励志啊 (Yes, it's really motivational).
A: 威尔史密斯也演出了很惨的感觉了 (Will Smith plays like he is a real tragedy).
B: 演技特别好 (Yes, he acts pretty well).

Wechat Dialogue Agent: Think

Use utterance rewriter to uncover hidden information

	Context 1
Utterance 1 (Translation)	*Human*: 梅西有多高? *Human*: How tall is **Messi**?
Utterance 2	*ChatBot*: 官方说法他的身高是5英尺7英寸。 *ChatBot*: Officially **he** is 5ft 7 inches.
Utterance 3	*Human*: 他和C罗谁是最好的球员? *Human*: Who is the best, **he** or C.Ronaldo?
Utterance 3′	*Human*: 梅西和C罗谁是最好的球员? *Human*: *Who is the best, **Messi** or C.Ronaldo?*
	Context 2
Utterance 1	*Human*: 你最喜欢什么电影? *Human*: What movie do you **like most**?
Utterance 2	*ChatBot*: 泰坦尼克。 *ChatBot*: **Titanic**.
Utterance 3	*Human*: 为什么呢? *Human*: Why?
Utterance 3′	*Human*: 为什么最喜欢泰坦尼克? *Human*: Why do you **like Titanic most**?

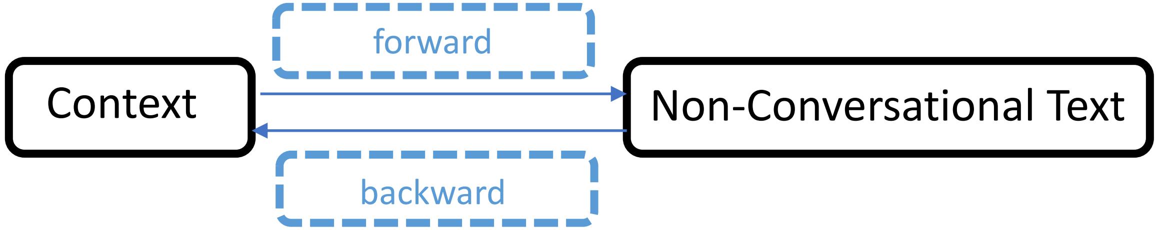
Model	Intention Precision	CPS
Original	80.77	6.3
With Rewrite	89.91	7.7

Shen et al. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. In ACL 2019

Wechat Dialogue Agent: Generalize

- Dialogue text are few and generic
- Non-conversational text are abundant and diverse
- Non-conversational text can be used to augment dialogue generation.

	Conversational Text
Context (Translation)	暗恋的人却不喜欢我 The one I have a crush on doesn't like me.
Response	摸摸头 Head pat.
	Non-Conversational Text
Forum Comments	暗恋这碗酒，谁喝都会醉啊 Crush is an alcoholic drink, whoever drinks it will get intoxicated.
Idiom	何必等待一个没有结果的等待 Why wait for a result without hope
Book Snippet	真诚的爱情之路永不会是平坦的 The course of true love never did run smooth (From *A Midsummer Night's Dream*)



Shen et al. Diversifying Dialogue Generation with Non-Conversational Text. In ACL 2020

Wechat Dialogue Agent: Specialize

- Pre-train in general-domain dialogue
- Design movie-specific dialog acts, aspects and processes

Dialogue Act	**Count(%)**	**Linked**	**Description**	**Example**
Request_fact	8.62	Fact	Request facts.	Who directed this movie?
Request_recommend	4.91	None	Ask recommendations.	Which other movies do you recommend?
Request_feeling	4.98	Comment	Request feelings.	How do you like its theme music?
Inform_fact	24.85	Fact	Inform facts.	Wong Kar-Wai directed this movie.
Inform_recommend	4.56	Movie	Give recommendations.	I can also recommend *Titanic*!
Inform_feeling	28.95	Comment	Convey feelings	Its music reminds me of my childhood!
Other	23.10	None	Greetings, echos, etc.	hahaha.

Aspects: name, director, actor, type, role, region, time, plot, line, award, gross, rating, website, music, others

Wechat Dialogue Agent: Specialize

Use a unified language model to encode and decode

Context	[context] *dialogue context*	**Encode**
Fact	[fact] *key-value pair(s)*	
Comment	[comment] *movie comment(s)*	
Recommend	[recommend] *movie name(s)*	
Track	[tracker] *[inherit] or a new movie name*	**Decode**
Intent	[intent] *DA sequence*	
Retrieve	[retrieve] *knowledge*	
Response	[response] *response*	

Wechat Dialogue Agent: Specialize

User Input	三傻大闹宝莱坞把我看哭了(I was moved into tears by 3 idiots)

Model input: [context] 三傻大闹宝莱坞把我看哭了

Model output: [track] [inherit] [intent] inform feeling others [retrieve]

Vector Match

Retrieve from user comment

Retrieved user comment: 感人的影片， 看了很多遍 (A touching movie, have watched many times)

Model input: [track] [inherit] [intent] inform feeling others [retrieve]感人的影片， 看了很多遍 [response]

Model output:] 我也是， 看了一百遍(Me too, have watched it one hundred times)

Su et al. MovieChats: Chat like Humans in a Closed Domain. In EMNLP 2020

Alexa Shopping Assistant: Goal

- Answer product-specific user questions
- Information accuracy is the most important goal.

Alexa, does this TV take wifi?

USER

- 电视支持WiFi吗?
- Wifiに対応していますか?
- هل يأخذ هذا التلفزيون شبكة Wifi
- est-ce que ce téléviseur prend le Wifi?
- Этот телевизор принимает Wi-Fi?
- 이 TV는 Wi-Fi를 사용합니까?

Yes, it is wifi enabled.

Alexa

- 对，支持WiFi的
- はい、Wi-Fi対応です
- نعم، تم تمكين wifi
- oui, le wifi est activé
- да, это Wi-Fi включен
- 예, Wi-Fi가 활성화되어 있습니다.

Alexa Shopping Assistant: Think

Alexa, does this TV take wifi?

Thinking process to get the answer:

- **Evidence ranker**
- Answer generator

Evidence 1: Just plug it in, connect to Wi-Fi, and enjoy.	0.95
Evidence 2: Voice remote with Alexa: Use your voice to watch live TV.	0.13
Evidence 3: Enhance your entertainment experience by easily controlling your smart home devices with Alexa.	0.03
Evidence 4: Smart but simple in every way.	0.01
…	

Alexa Shopping Assistant: Think

The PSQA systems contains two main components:

- Evidence ranker
- **Answer generator**

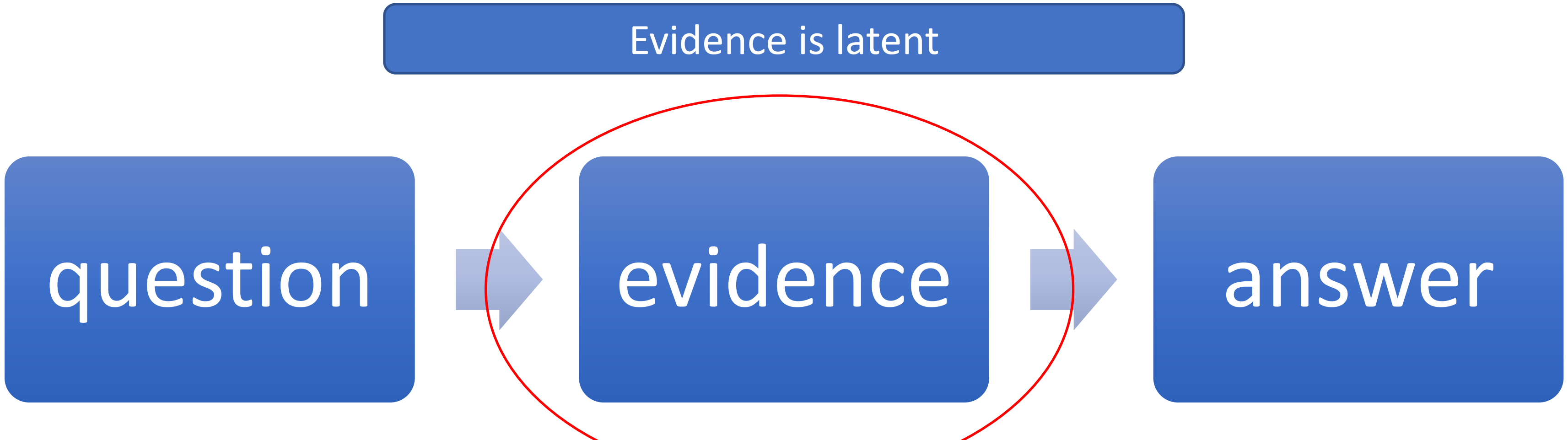
Alexa, does this TV take wifi?

Evidence: Just plug it in, connect to Wi-Fi, and enjoy.

Yes, it is wifi enabled.

Alexa Shopping Assistant: Think

- The evidence retrieval part is treated as latent variables.
- $p(evidence|question)$ is a categorical or **multinomial** distribution.
- Train on all historical Amazon question-answer pairs.



$p(evidence|question)$: prior $\quad G(answer|question, evidence)$: likelihood

Shen et al. Neural Ranking with Weak Supervision for Open-Domain Question Answering: A Survey In EACL 2023

Alexa Shopping Assistant: Generalize

Few-shot learning For Evidence Ranker:

- Use QA pairs from AmazonQA as pseudo pair
- Train a question generator to generate pseudo questions for evidence

Few-shot learning for Answer Generator:

- Train a backward model to generate pseudo evidence from QA pair
- Weak supervision from self-generated label

Shen et al. Product Answer Generation from Heterogeneous Sources: A New Benchmark and Best Practices. In ACL-ECNLP 2022

Alexa Shopping Assistant: Specialize

- Adapt to the shopping domain
- Adapt to multiple languages

Pre-train on QA → Pre-train on shopping-domain QA → Fine-tune on English Annotation → Fine-tune on Multilingual Annotation

Latent-variable model

Data Augmentation + Weak Supervision

Summary: Productionizing Process

1. Define the thinking process in the general domain of targeted task
2. Learn the thinking process from with as little annotation as possible
3. Specialize to the targeted domain

Future:

- Step 1 and 2 might not be needed. A foundation model with basic thinking ability would be available.
- Principled way to specialize to the targeted domain with little annotation cost

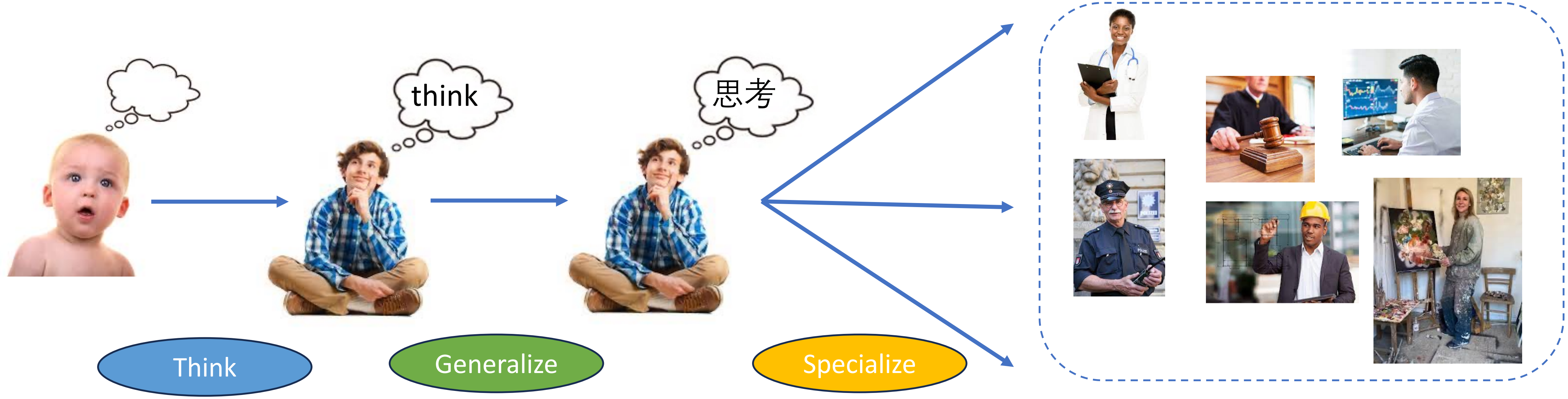
Future Direction

- Build agent with explainable and verifiable output
- From English agent to Chinese agent
- From Chinese agent to Chinese domain expert

Think

Generalize

Specialize




tusind tak
ngiyabonga
谢谢
dakujem vám
dziękuję
ありがとう
merci
suksema
danke
thank
you
baie dankie
धन्यवाद
molte grazie
gracias
takk
謝謝
obrigada
dank u
teşekkür edire
mahalo
ధన్యవాదాలు
obrigado
gràcies
teşekkür ederim
شكرا
tänan
tack så mycket
O ṣé
Спасибо