Advancements and Prospects in Dialogue Agents and LLMs

Jiahuan Pei

Centrum Wiskunde & Informatica (CWI)

ppsunrise99@gmail.com

Invited Talk at Bosch Center of Artificial Intelligence (BCAI)

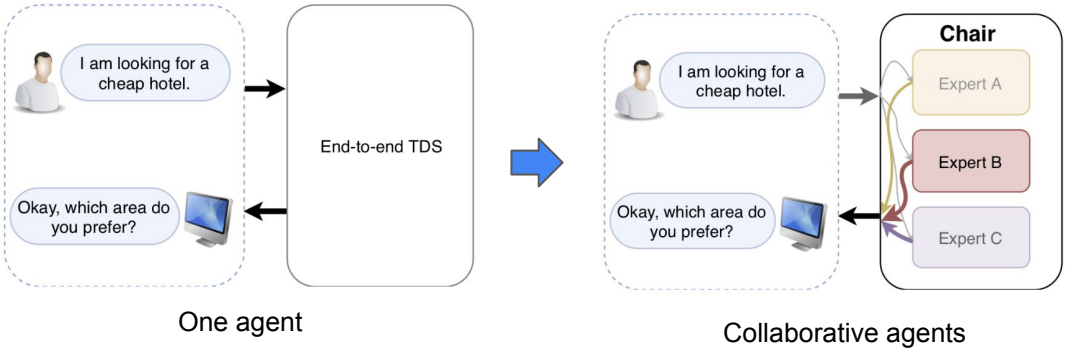
Self-Introduction

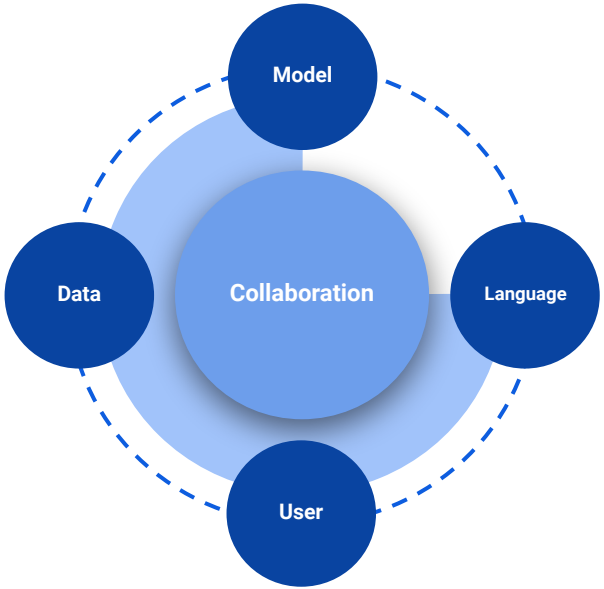
Jiahuan Pei

- Natural Language Processing (NLP)
  - **Dialogue systems**
  - Word embeddings
  - Functional phrase & clause identification
  - Microblog opinion summarization
- Information Retrieval (IR)
  - Matcher embedding
  - Conversational recommendation
  - Query understanding & disambiguation

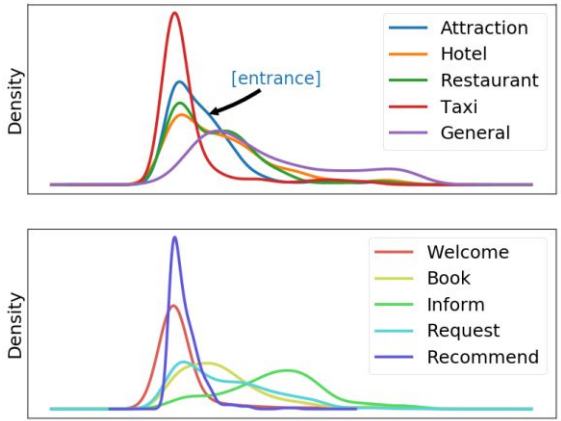
1. Recent Work

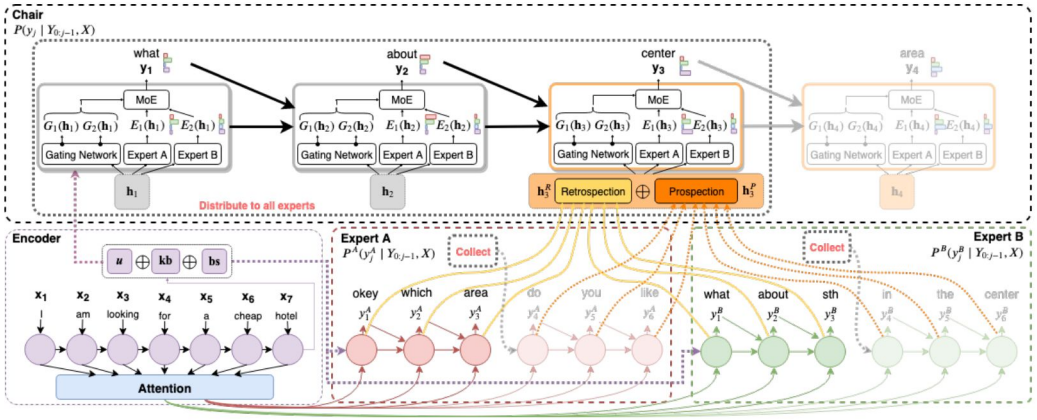
1.1 Cooperative Dialogue Agents





Model Collaboration for Response Generation





- A chair-expert model
- Retrospective and prospective collaboration mechanisms
- A global-local learning scheme

J. Pei, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020. (first author, CCF B conference).

Model Collaboration for Response Generation

	BLEU	Inform	Success	Score	PPL
S2SAttnLSTM	18.90%	71.33%	60.96%	85.05	**3.98**
S2SAttnGRU	18.21%	81.50%	68.80%	93.36	4.12
Structured Fusion [20]	16.34%	82.70%	72.10%	93.74	–
LaRLAttnGRU [36]	12.80%	82.78%	**79.20%**	93.79	5.22
MoGNet	**20.13%***	**85.30%***	73.30%	**99.43***	4.25

Automatic Evaluation

	S2SAttnGRU		LaRLAttnGRU		MoGNet	
	$\geqslant 1$	$\geqslant 2$	$\geqslant 1$	$\geqslant 2$	$\geqslant 1$	$\geqslant 2$
Informativeness	56.79%	31.03%	76.54%	44.83%	**80.25%**	**53.45%**
Consistency	45.21%	23.53%	71.23%	39.22%	**80.82%**	**50.98%**
Satisfactory	26.79%	25.00%	44.64%	21.88%	**60.71%**	**37.50%**

Human Evaluation

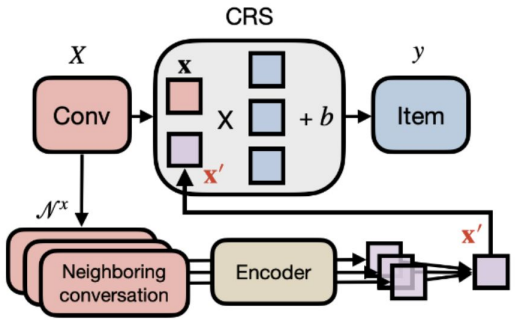
MoGNet outperforms all baselines by a large margin in terms of overall performance.

J. Pei, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020. (first author, CCF B conference).

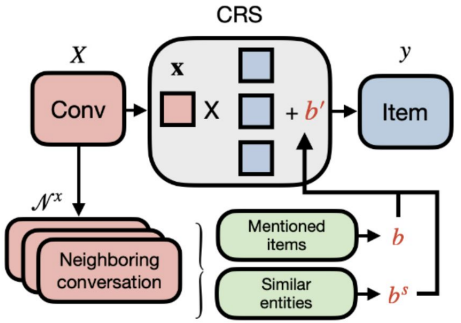
Data Collaboration for Conversational Recommendation

Context	**Conversation**: User: Hi, I am looking for Jim Carrey movies. Can you recommend any good ones? System: *Ace Ventura: Pet Detective* is a good one I remember from my childhood. User: I loved it! What's your second favorite? System: I also really enjoyed *Liar Liar* !
Ground truth	**Response**: *Liar Liar* was my next pick.
Entities from neighboring conversations	*Liar Liar*, Ace Ventura: Pet Detective, Dumb and Dumber, The Mask (1994), Bruce Almighty, Jumanji, *The Number 23*, Yes Man, The Mask (1988), Airplane!
C^2-CRS	**Generated response**: Oh yes lol.
NR-CRS	**Generated response**: *Liar Liar* and *The Number 23* would be pretty funny.

An example of how entities from neighboring conversations help with correct recommendation.



(b) CRS with enhanced user representation.



(c) CRS with enhanced user preference.

Liu, Y., **Pei, J.**, de Rijke, M. Augmentation with Neighboring Information for Conversational Recommendation (Ongoing work for TOIS 2023)

Data Collaboration for Conversational Recommendation

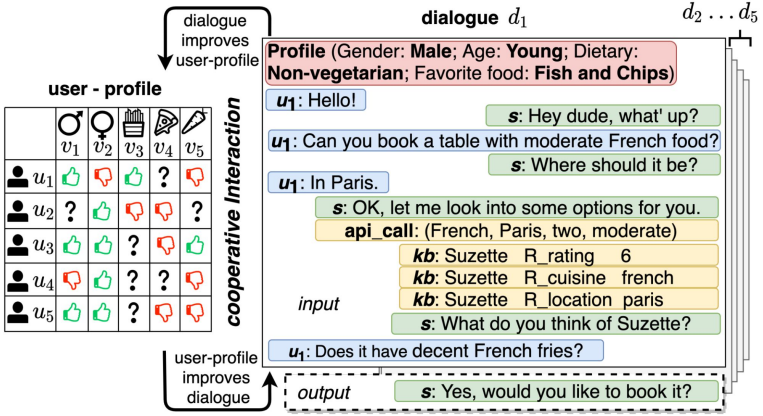
Method	Recall@K		MRR@K		NDCG@K	
	$K = 10$	$K = 50$	$K = 10$	$K = 50$	$K = 10$	$K = 50$
Popularity*	0.054	0.183	0.022	0.028	0.030	0.058
TextCNN*	0.063	0.162	0.022	0.026	0.031	0.053
ReDial*	0.156	0.303	0.064	0.072	0.086	0.119
KBRD*	0.168	0.333	0.064	0.072	0.088	0.125
KGSF*	0.183	0.369	0.072	0.081	0.098	0.139
KECRS*	0.159	0.308	0.064	0.073	0.087	0.120
RevCore*	0.187	0.377	0.073	0.082	0.100	0.140
SSCR*	0.204	0.385	0.080	0.088	0.109	0.149
C^2-CRS	0.233	0.407	0.101	0.109	0.132	0.171
NR-CRS	0.261▲	0.440▲	0.118▲	0.126▲	0.152▲	0.191▲

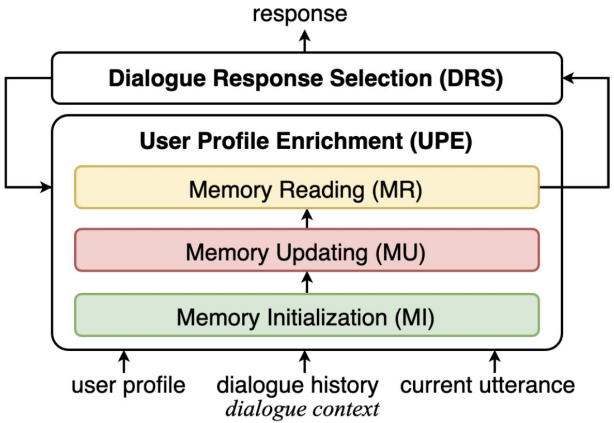
Recommendation performance on REDIAL dataset.

NR-CRS significantly outperforms all baselines in terms of metrics.

Liu, Y., **Pei, J.**, de Rijke, M. Augmentation with Neighboring Information for Conversational Recommendation (Ongoing work for TOIS 2023)

User Collaboration for Personalized TDSs





- A closed loop cooperative paradigm
  - Dialogue helps to enrich the user-item interactions.
  - Enriched user-item interactions help to select a better response.
- A learning algorithm with multiple hops

J. Pei, P. Ren, C. Monz, M. de Rijke. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. TheWebConf 2021. (first author, CCF A conference).

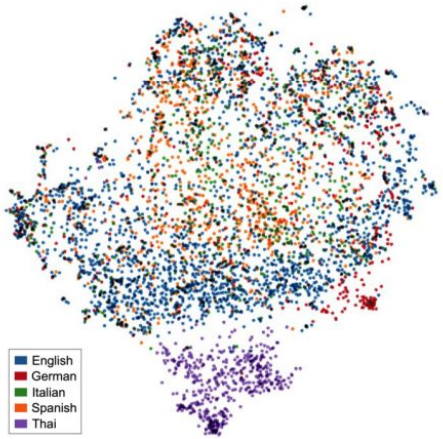
User Collaboration for Personalized TDSs

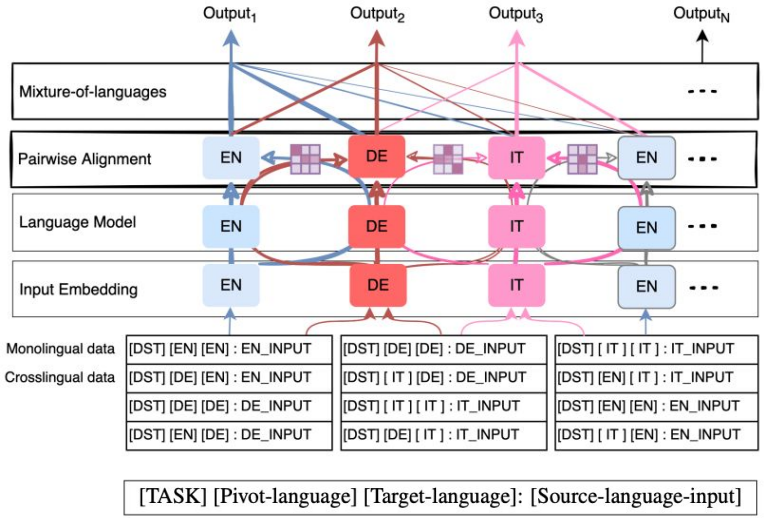
Discard Ratio	0%	10%	30%	50%	70%	90%	100%
NPMemNN	87.91	86.11	86.56	85.79	83.93	84.08	**84.83**
CoMemNN	**91.13***	**89.90***	**88.69***	**87.80***	**86.35***	**84.83***	82.85
Small Set/Diff.	3.22	3.79	2.13	2.01	2.42	0.75	−1.98
NPMemNN	97.49	97.01	96.05	95.52	95.40	90.96	90.50
CoMemNN	**98.13***	**97.94***	**97.68***	**97.53***	**96.98***	**96.63***	**92.73***
Large Set/Diff.	0.64	0.93	1.63	2.01	1.58	5.67	2.23

CoMemNN can infer missing values when the profile discard ratios range from 0% to 90%.

J. Pei, P. Ren, C. Monz, M. de Rijke. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. TheWebConf 2021. (first author, CCF A conference).

Language Collaboration for Multilingual TDSs





- A unified generation framework with mixture-of-language routing for Multilingual TDSs.
- Benefits from
  - Multilingual data argumentation;
  - Language characteristic modelling;
  - Mixture-of-language routing.

J. Pei, G. Yan, P. Ren, and M. de Rijke. Mixture-of-languages routing for multilingual task-oriented dialogue systems. (Under Review by TOIS 2023).

Language Collaboration for Multilingual TDSs

Model	DST: Joint Goal / Request Accuracy (%)					
	English (EN)		German (DE)		Italian (IT)	
mT5	89.53/97.02		79.06/95.92		87.58/95.44	
	EN,DE→EN	EN,IT→EN	DE,EN→DE	DE,IT→DE	IT,EN→IT	IT,DE→IT
mT5+bMOLR	91.42/97.32	91.11/**97.57**	**81.62**/96.65	**81.62**/96.23	**88.25/96.53**	86.98/96.41
	1.DE,IT→DE 2.EN,DE→EN	1.IT,DE→IT 2.EN,IT→EN	1.EN,IT→EN 2.DE,EN→DE	1.IT,EN→IT 2.DE,IT→DE	1.EN,DE→EN 2.IT,EN→IT	1.DE,EN→DE 2.IT,DE→IT
mT5+mMOLR	**91.84**/97.02	91.42/97.14	81.56/**97.02**	81.38/96.23	87.77/96.41	86.00/96.35

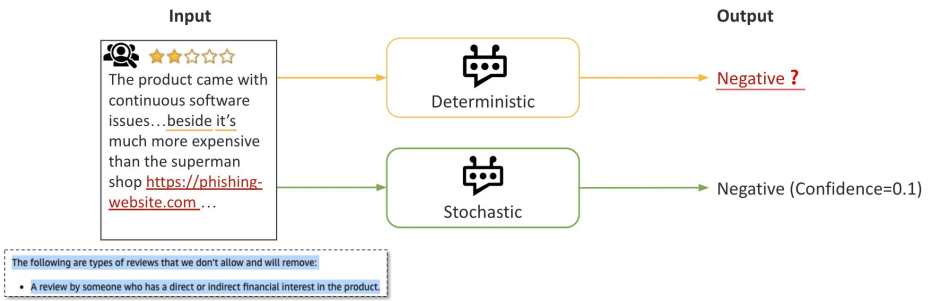
	DST: Similarity				NLU: Similarity		
Language pair	Genetic	Word	Sentence	Language pair	Genetic	Word	Sentence
(EN, DE)	0.1667	0.6725	0.8813	(EN, ES)	0.0833	0.7448	0.8777
(EN, IT)	0.1250	0.6711	0.9036	(EN, TH)	0.0000	0.4787	0.5706
(DE, IT)	0.0909	0.6486	0.9066	(ES, TH)	0.0000	0.4056	0.5512

$$\phi(EN, DE) > \phi(EN, IT) > \phi(DE, IT) > \phi(EN, ES) > \phi(EN, TH) > \phi(ES, TH)$$

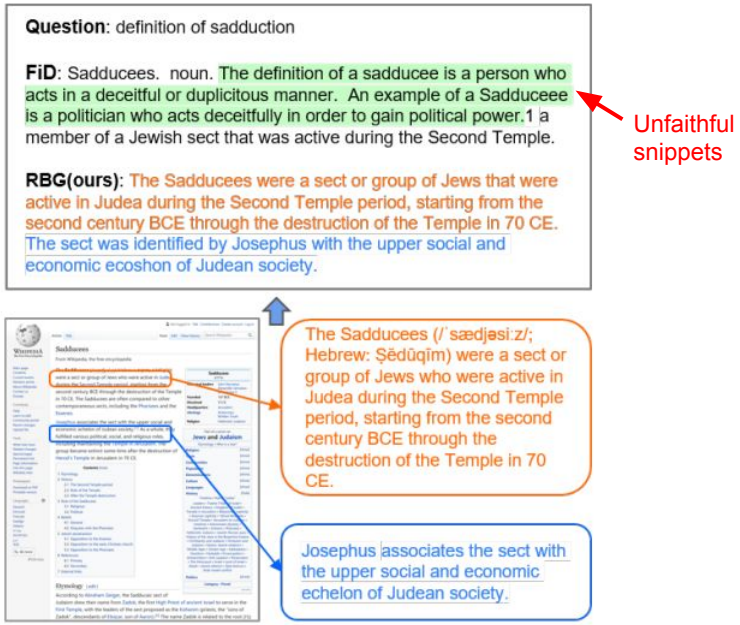
Gains are language-specific: better when the similarity between source and pivot language is larger.

J. Pei, G. Yan, P. Ren, and M. de Rijke. Mixture-of-languages routing for multilingual task-oriented dialogue systems. (Under Review by TOIS 2023).

1.2 Trustworthy Systems



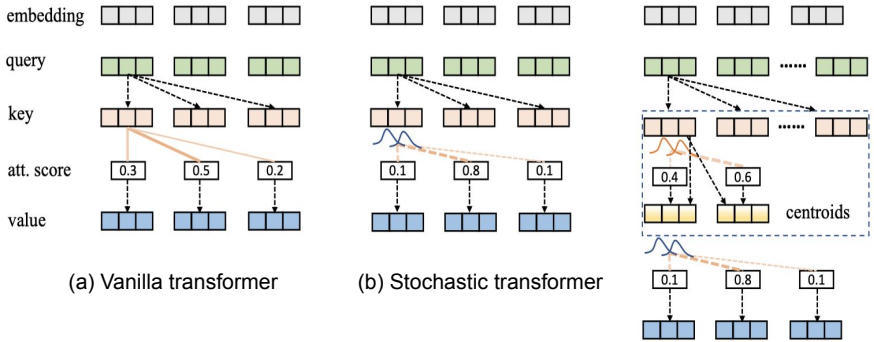
Uncertainty



Su, Dan, et al. "Read before Generate! Faithful Long Form Question Answering with Machine Reading." Findings of ACL 2022.

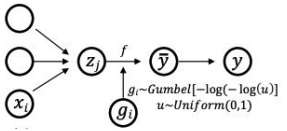
Faithfulness

Stochastic Transformers for Classification under Uncertainty



(a) Vanilla transformer (b) Stochastic transformer

(c) Hierarchical stochastic transformer



Gumbel-Softmax trick.

$$Q = \mathbf{W}_q\mathbf{x},\ K = \mathbf{W}_k\mathbf{x},\ V = \mathbf{W}_v\mathbf{x},$$
$$A = \mathrm{softmax}(\alpha^{-1}QK^\top),\ H = AV,$$

Sample attention weights

$$\hat{a}_i \sim \mathcal{G}\left(\frac{q_i k_i^\top}{\tau}\right)$$

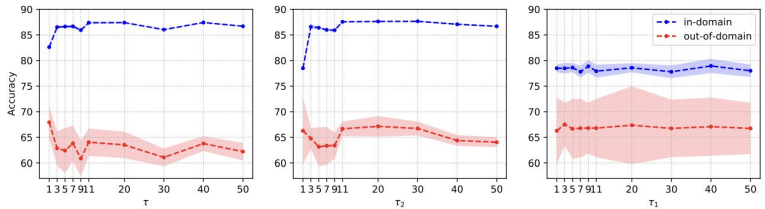
Key stochastically to attend to learnable centroids

$$\hat{a}_c \sim \mathcal{G}(\tau_1^{-1} k_i C), \hat{a}_c \in \mathbb{R}^{l\times c},$$
$$\hat{k}_i = \hat{a}_c C^\top, \hat{k}_i \in \mathbb{R}^{l\times d_h},$$
$$\hat{a}_v \sim \mathcal{G}(\tau_2^{-1} q_i \hat{k}_i^\top), \hat{a}_v \in \mathbb{R}^{l\times l},$$

Pei, J., Wang, C., Szarvas, G. Transformer uncertainty estimation with hierarchical stochastic attention. AAAI 2022.

Stochastic Transformers for Classification under Uncertainty

	ID (%)	OOD (%)	▽ID (%)	▽OOD (%)
TRANS ($\eta = 0.1$)	87.00	65.00	/	/
TRANS ($\eta = 0.5$)	87.51	63.40	0.51 ↑	1.60 ↓
MC-DROPOUT ($\eta = 0.5$)	86.06 ± 0.087	63.38 ± 1.738	0.94 ↑	1.62 ↓
MC-DROPOUT ($\eta = 0.1$)	87.01 ± 0.075	63.38 ± 0.761	0.10 ↑	1.62 ↓
ENSEMBLE	86.89 ± 0.230	64.20 ± 1.585	0.11 ↓	0.80 ↓
STO-TRANS ($\tau = 1$)	82.62 ± 0.092	67.92 ± 0.634	4.38 ↓	2.92 ↑
STO-TRANS ($\tau = 40$)	87.42 ± 0.022	63.78 ± 0.289	0.42 ↑	1.22 ↓
H-STO-TRANS ($\tau_1 = 1, \tau_2 = 20$)	87.63 ± 0.017	67.14 ± 0.400	0.63 ↑	2.14 ↑
H-STO-TRANS ($\tau_1 = 1, \tau_2 = 30$)	87.66 ± 0.022	66.72 ± 0.271	0.66 ↑	1.72 ↑



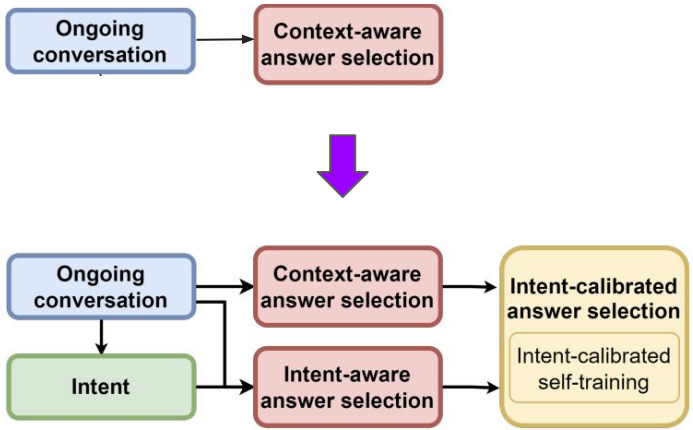
- Enable transformers with uncertainty estimation while retain the original predictive performance.
- Stochastic transformer has difficulties in the trade-off between in-domain and out-of-domain performance.

Pei, J., Wang, C., Szarvas, G. Transformer uncertainty estimation with hierarchical stochastic attention. AAAI 2022.

Intent-calibrated Self-training for Answer Selection

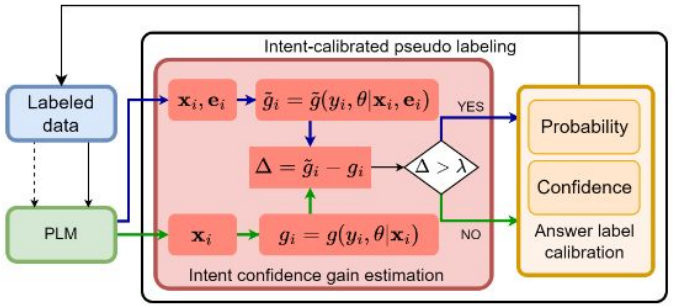
Context Utterances			Intents
User: How does a photon picture make the pattern?			OQ
Agent: Photons in mainstream physics, are quantum mechanical entities which in great numbers build up the classical electromagnetic radiation...			PA
User: Do you know why the photon which is hitting forward is causing an electron to move up-down?			IR
Candidate Answers	**Model**	**ICG**	**Probablity**
A1: The theories of quantum mechanics for electron photon interactions can be found in https://www.website.com.	TSST	/	0.00
	ICAST	0.14	0.99
A2: The energy of a photon is equal to the level spacing of a two-level system. It is a result of energy conservation...	TSST	/	0.96
	ICAST	-0.13	0.71

An example: predicted intents help with selecting a correct answer.

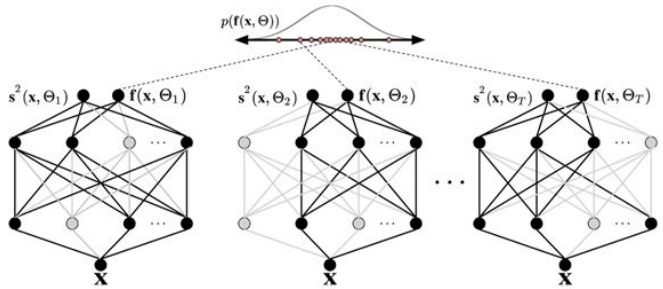


Deng, W., **Pei, J.**, Ren, P. Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues. Transactions of ACL 2023

Intent-calibrated Self-training for Answer Selection



More confident with the predicted intents?
If YES, it's a high-qualitied data for training!



Monte Carlo (MC) dropout

$$\tilde{g}(y_i, \beta | \mathbf{x}_i, \mathbf{e}_i)$$
$$= \mathbf{H}[y_i | \mathbf{x}_i, \mathbf{e}_i] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i | \mathbf{x}_i, \mathbf{e}_i; \beta]]$$
$$\approx \mathbf{H}[\mathbf{E}_{P(\beta)}[y_i | \mathbf{x}_i, \mathbf{e}_i; \beta]] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i | \mathbf{x}_i, \mathbf{e}_i; \beta]]$$
$$\approx -\frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t \cdot \log \frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t + \frac{1}{T}\sum_{t=1}^{T} p_i^e \tilde{p}_t \cdot \log p_i^e \tilde{p}_t,$$

$>$

$$g(y_i, \beta | \mathbf{x}_i)$$
$$= \mathbf{H}[y_i | \mathbf{x}_i] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i | \mathbf{x}_i; \beta]]$$
$$\approx \mathbf{H}[\mathbf{E}_{P(\beta)}[y_i | \mathbf{x}_i; \beta]] - \mathbf{E}_{P(\beta)}[\mathbf{H}[y_i | \mathbf{x}_i; \beta]]$$
$$\approx -\frac{1}{T}\sum_{t=1}^{T} p_t \cdot \log \frac{1}{T}\sum_{t=1}^{T} p_t + \frac{1}{T}\sum_{t=1}^{T} p_t \cdot \log p_t,$$

Confidence scores with and without intents

Deng, W., **Pei, J.**, Ren, P. Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues. Transactions of ACL 2023

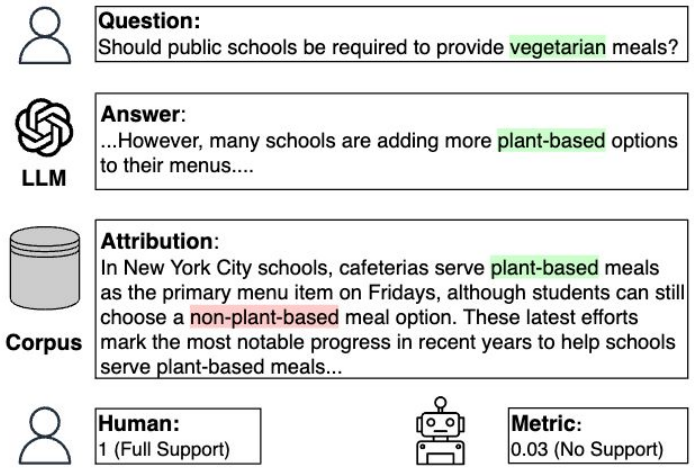
Intent-calibrated Self-training for Answer Selection

		MSDIALOG							MANTIS						
Setting	Model	P	R	F1	R@1	R@2	R@5	MAP	P	R	F1	R@1	R@2	R@5	MAP
1% labeled	IART†	22.18	46.75	30.08	25.65	46.28	77.58	47.74	48.29	52.22	50.18	50.40	68.34	86.22	66.12
	SAM†	44.17	44.36	44.26	46.89	59.06	77.02	60.72	57.75	58.62	58.18	65.10	76.32	88.54	75.60
	JM‡	44.80	44.59	44.70	44.54	60.76	84.30	61.26	62.95	62.62	62.78	62.64	77.32	92.30	75.25
	BIG‡	44.07	44.78	44.42	50.93	66.30	87.50	66.15	57.91	57.42	57.66	70.22	83.04	95.12	80.78
	GRAY‡	41.68	42.15	41.91	51.26	66.40	85.62	66.10	61.30	60.72	61.01	64.67	77.34	88.32	75.57
	GRN‡	43.41	43.37	43.39	43.28	61.60	86.46	61.19	61.75	61.10	61.42	61.06	76.64	93.66	74.56
	BERT_FP†	44.32	42.95	43.62	56.76	**72.08**	**91.25**	**70.90**	66.26	62.86	64.51	75.62	**86.14**	**95.22**	**84.11**
	BERT‡	48.56	45.34	46.90	54.79	68.32	85.80	68.04	67.28	65.62	66.44	74.82	83.00	92.16	82.41
	ICAST (Teacher)	**49.82**	**46.33**	**48.01**	**56.86**	67.81	85.38	69.03	**68.48**	**66.12**	**67.28**	**77.28**	86.12	94.98	82.98
1% labeled	TSST‡	53.72	52.58	53.14	61.04	73.91	89.70	73.04	73.73	72.60	73.16	82.94	**91.08**	**97.88**	**89.18**
+all unlabeled	ICAST	**57.05**	**54.32**	**55.65**	**62.21**	**76.31**	**91.07**	**73.77**	**74.89**	**72.72**	**73.79**	**83.68**	90.68	96.42	88.31

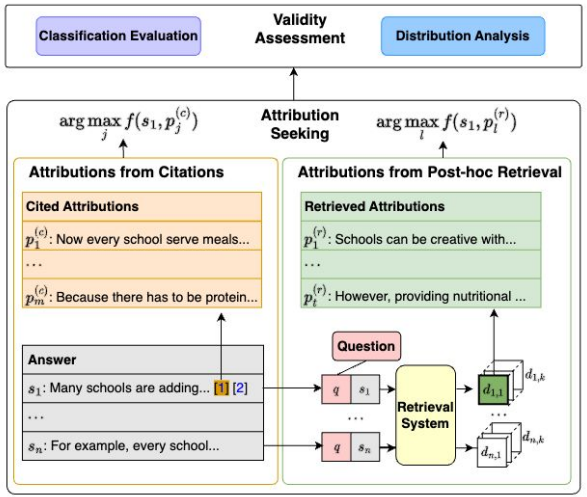
The predicted intents can provide more information that are useful for selecting correct answers

Deng, W., **Pei, J.**, Ren, P. Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues. Transactions of ACL 2023

Validity of Faithfulness Metrics in Attribution Seeking for LLMs



An example of an invalid faithfulness metric when the attribution contains partial support evidence.

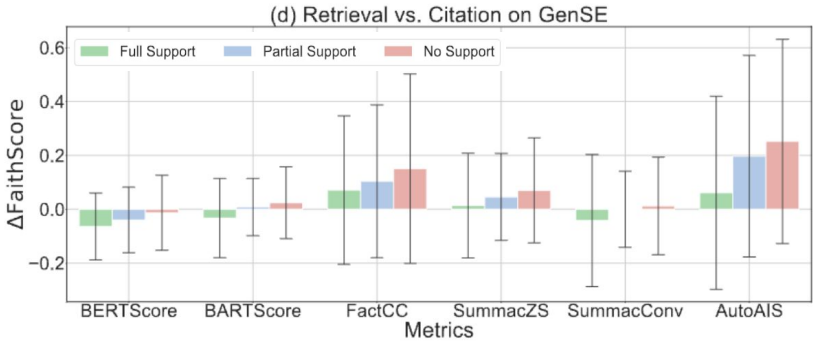


25.5% of citations are NOT fully support the statements.

Zhang, W., **Pei, J.**, Kanoulas, E. Exploring the Validity of Faithfulness Metrics in Attribution Seeking for LLMs (Under Review by CoNLL 2023)

Validity of Faithfulness Metrics in Attribution Seeking for LLMs

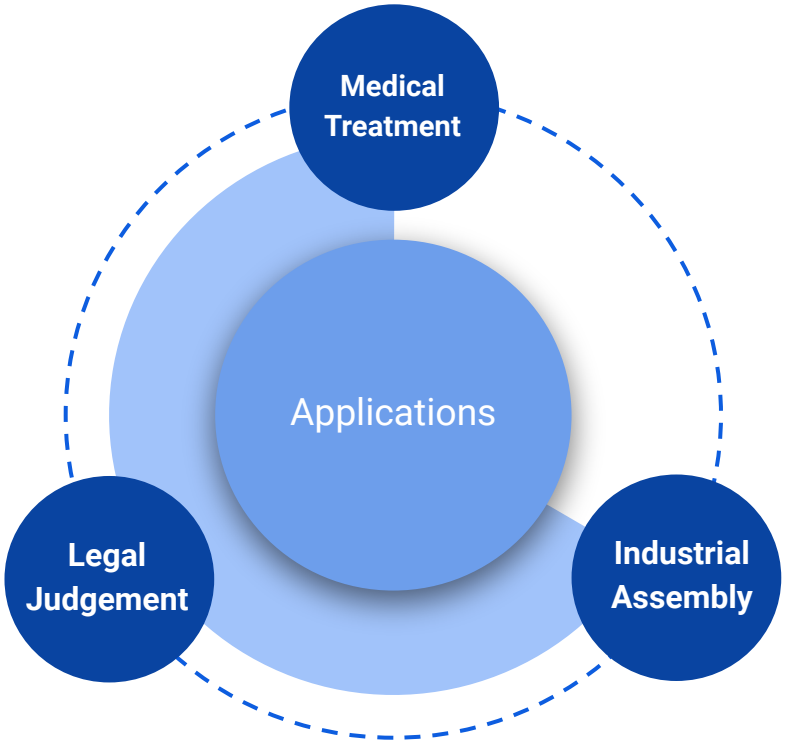
Metric	GENSE						ATOMIC			
	Evidence			Citation			Sentence			Fact
	F/N	F/P	P/N	F/N	F/P	P/N	F/N	F/P	P/N	F/N
BERTScore	**92.78**	79.68	**78.33***	74.59	69.20	**60.50**	85.15	73.63	**68.47**	67.97
BARTScore	87.71	76.33	72.09	72.29	68.88	56.82	85.79	75.35	67.82	68.45
FactCC	72.09	62.56	60.74	63.16	58.79	55.08	68.83	68.41	50.99	52.56
SUMMAC_{ZS}	89.67	79.19	73.16	69.22	67.83	52.92	83.24	78.96	58.83	79.32
SUMMAC_{Conv}	71.77	78.87	35.02	70.67	69.93	52.32	81.70	79.09	56.28	83.05
AutoAIS	92.01	**83.46***	72.20	**76.83**	**72.63**	58.34	**88.36***	**83.59***	60.33	**84.90**
Ensemble	93.93	84.64	77.08	76.84	74.11	58.05	90.33	83.88	65.72	84.03



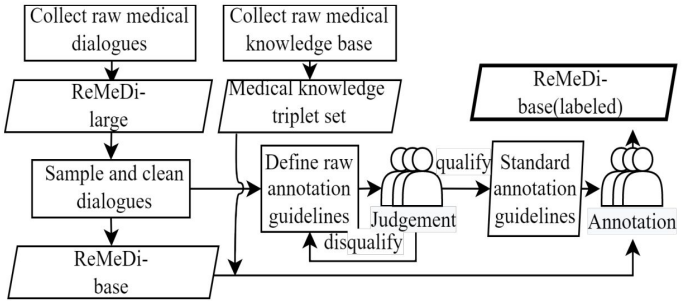
- Partial support (F/P and P/N) drops ~10% performance compared with F/N;
- The larger the score, the better the metrics are at finding attributions that possibly contain more supporting evidence.

Zhang, W., **Pei, J.**, Kanoulas, E. Exploring the Validity of Faithfulness Metrics in Attribution Seeking for LLMs (Under Review by CoNLL 2023)

1.3 LLMs Tailored for Specific Applicable Domains



Multi-Domain Multi-service Medical Dialogues



- A dataset contains 96,965 conversations between doctors and patients, including 1,557 conversations with fine-gained labels;
- Benchmarks using pretrained models (i.e., BERT-WWM, BERT-MED, GPT2, and MT5);
- A proposed self-supervised contrastive learning(SCL) model.

G. Yan, **J. Pei**, P. Ren, Z. Ren, X. Xin, H. Liang, M. de Rijke. ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues. SIGIR 2022.

Syllogistic Reasoning for Legal Judgment Analysis

Input: Fact

Defendant A used a credit card from Bank B in May 2008 for malicious overdraft consumption of 18060 yuan without repayment.

Task 1: Article Retrieval

Generated Query:
A324 Credit Card Fraud Crime Using Fiction of Identity or Overdraft of Credit Cards for Fraud.
...
Retrived Document: Article 196 [Crime of Credit Card Fraud] Whoever Maliciously overdrawns, with a relatively large amount, shall be sentenced to not more than five years or criminal detention.
...

Task 2: Criminal Element Generation

Subject: Defendant A who is a fully capable person.
Subjective Element: Intent

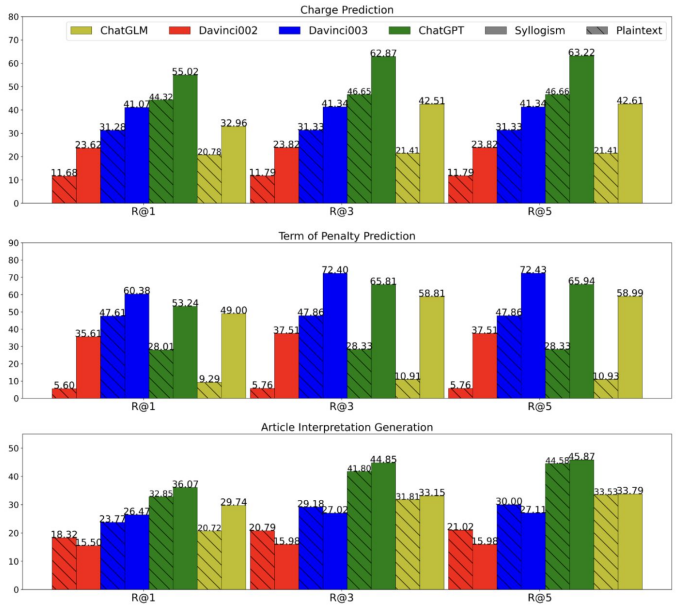
Object: Bank B's property security
Objective Elements:
1. Time: May 2008.
2. Location: Bank B.
3. Action: Overdraft consumption through credit card.
4. Results:
The overdraft amount reaches 18060 yuan.

Task 3: Article Interpretation Generation

Article 196: Defendant A maliciously overdrawn his credit card with a huge amount, which meets the Articel 196 of credit card fraud.

Task 4: Legal Judgment Prediction

Charge: Crime of Credit Card Fraud
Criminal Penalty: Imprisonment of up to five years



- A dataset contains 11,239 legal cases, including 1,337 facts, 124 articles, 80 charges and 5 terms.
- LLM Benchmarks: ChatGLM (6.2B+RLHF), Davinci002 (175B), Davinci003 (Davinci002+RLHF), ChatGPT(Davinci002+RLHF+Prompt&Context)

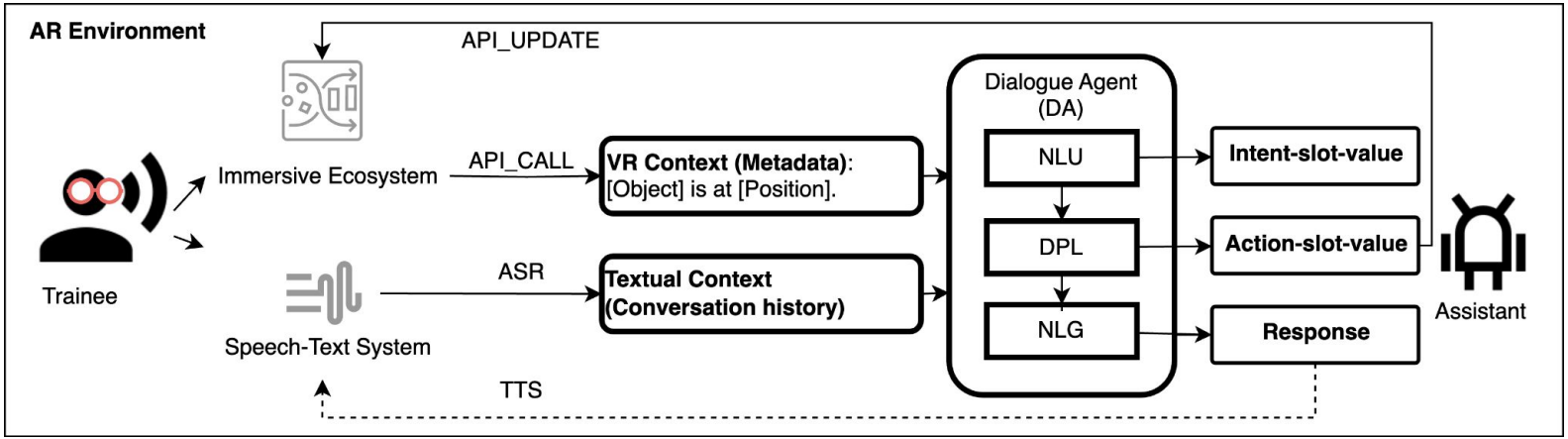
Deng, W., **Pei, J.**, Ren, P. Syllogistic Reasoning for Legal Judgment Analysis (Under Review by EMNLP 2023)

Augmented Reality Guided Dialogue Agent for Industrial Assembly

Turn	Trainee-Assistant Conversation	Dialogue Modeling
1	Trainee: Hi, I'm ready to start assembling the car.	Trainee: StartAssembly
2	Agent: Great! Let me check the progress for you.	Agent: ProvideProgressStatus
3	Trainee: Can you please tell me where I can find the engine?	Trainee: GetPartLocation {PartName: Engine}
4	Agent: Sure, the engine is located in the storage area on the right side of the workshop.	Agent: ProvidePartLocation {PartName: Engine}
5	Trainee: Thank you. Now, I need instructions for step 5 of the assembly.	Trainee: GetAssemblyInstructions {StepNumber: 5}

Pei, J., Ren, P, Cesar, P. AR-Dialog: Augmented Reality Guided Dialogue Agent for Industrial Assembly (Ongoing work for ACL 2024, ARR)

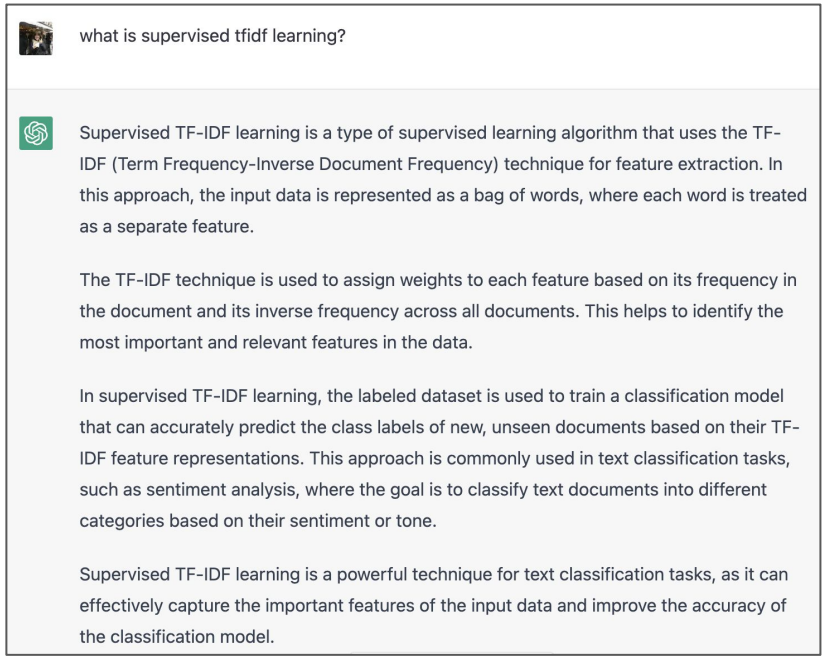
Augmented Reality Guided Dialogue Agent for Industrial Assembly

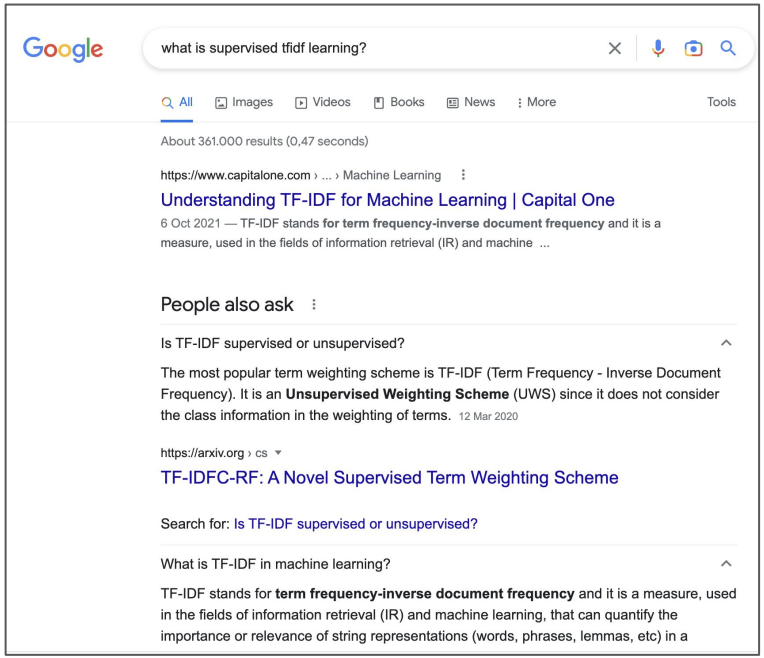


Pei, J., Ren, P, Cesar, P. AR-Dialog: Augmented Reality Guided Dialogue Agent for Industrial Assembly (Ongoing work for ACL 2024, ARR)

2. Next Steps

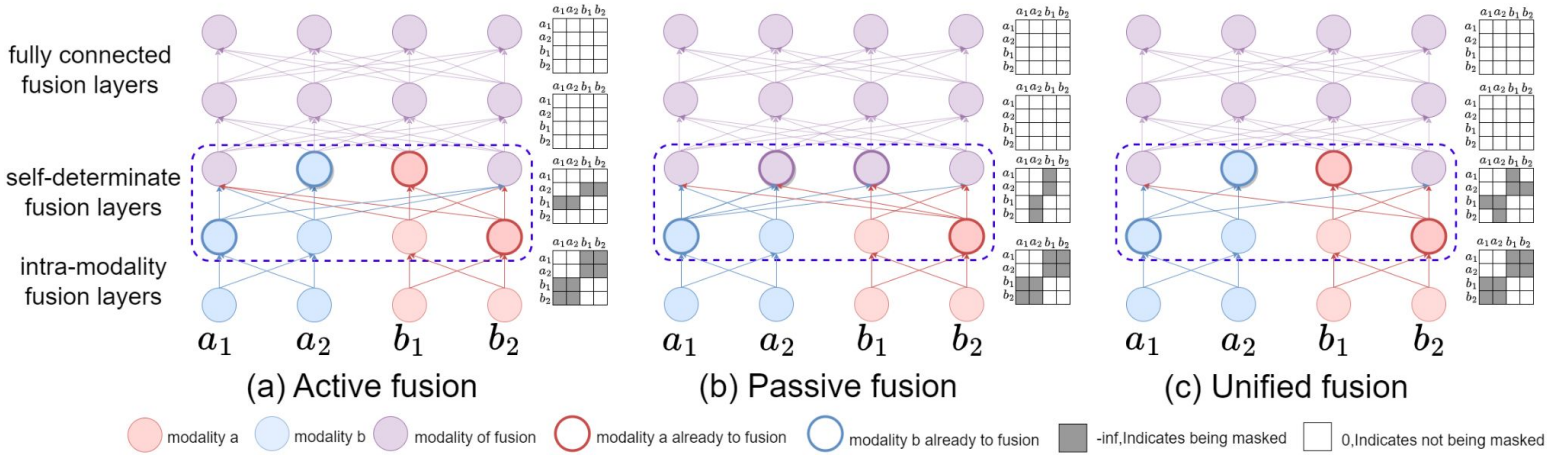
2.1 Faithfulness / Uncertainty Estimation





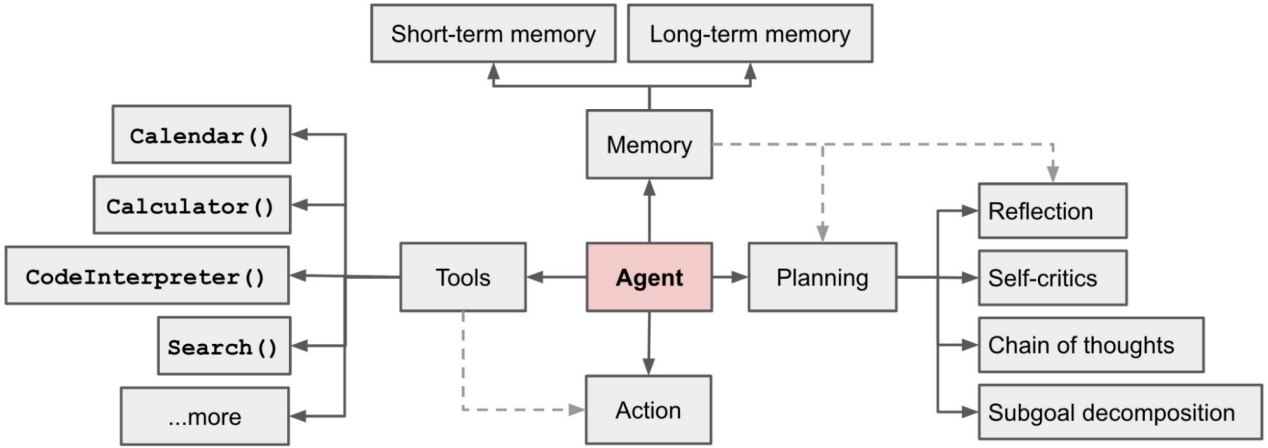
The screenshots are from Chatgpt (left) and Google (right), respectively.

2.2 Multimodal LLMs



Self-determinate Multi-modality Fusion for Transformers (Ongoing)

2.3 LLM Powered Autonomous Agents



The figure is from https://lilianweng.github.io/posts/2023-06-23-agent/

Thank you for your attention!
Q & A