Conversational AI: Unveiling the Magic Behind Pre-trained Large Language Models in Digital Assistants

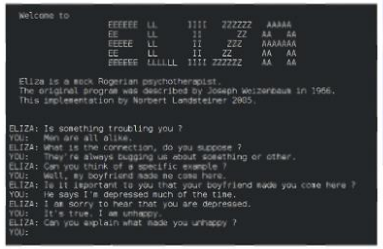
Meng Chen

https://chenmengdx.github.io

2023/11/30

History of Conversational AI

- Conversational AI is a subfield of artificial intelligence that focuses on enabling machines to engage in human-like conversation. It uses technologies like machine learning, natural language processing, and speech recognition to understand and respond to human language in a natural and intuitive way



```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Turing Test
[Turing, 1950]

MIT Eliza
[Weizenbaum, 1966]

ATIS (Airline Travel Information Systems)
[Pallett, et al. 1990]

TREC QA
[Voorhees, et al. 2000]
POMDP
[Williams and Young, 2007]

Apple Siri
Feb. 2010
IBM Watson won on *Jeopardy!*
[Ferrucci, et al. 2011]
DSTC1
[Jason, et al. 2013]

Amazon Echo
Nov. 2014
Microsoft Cortana
Apr. 2014
Seq2Seq Dialogue
[Vinyals, et al. 2015]
SQuAD
[Pranav, et al. 2016]
Microsoft XiaoIce
[Shum et al. 2018]

MultiWOZ
[Budzianowski, et al. 2018]
CoQA
[Reddy et al. 2018]
QuAC
[Choi et al. 2018]
MMD
[Amrita, et al. 2018]

Google Meena
[Adiwardana, et al. 2020]
Meta Blender
[Roller, et al. 2020]
Meta SIMMC
[Seungwhan, et al. 2020]
Meta Image-Chat
[Kurt, et al. 2020]
Salesforce SimpleTOD
[Hosseini-Asl, et al. 2020]

Open AI ChatGPT
[Ouyang et al, 2022]
Claude 2
[Anthropic, 2022]
LLAMA
[Touvron, et al, 2023]
Google Bard
Mar. 2023
Baidu ERNIE Bot
Mar. 2023

Before 1990 2000 2010 2014 2018 2020 **2022**

Challenges of Conversational AI

- **NLU (Natural Language Understanding)**

Ambiguities, Handling variability, Context management, Slangs, typos, and abbreviations, domain knowledge

- **Understanding User**

User profile, personality

- **Difficulty in Communicating**

Multi-lingual, spoken language

- **Costly Investment**

High-quality real-scenario training data, computing resources

- **Public Skepticism**

public suspicion, misconceptions

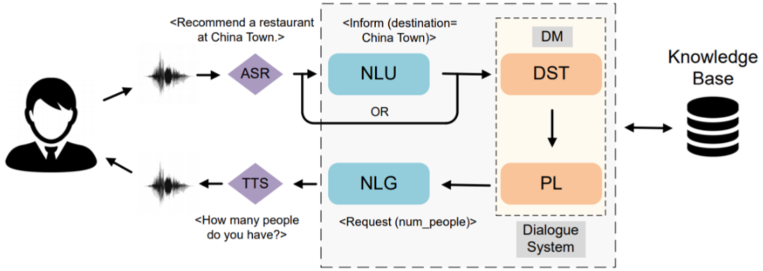
- **Unexpected Questions**

unanticipated customer interactions

Category of Dialogue System

Task-oriented Dialogue System

- Designed to help users achieve pre-defined tasks, such as booking tickets, ordering food, or scheduling calls
- Communicate with humans via text, speech, gestures, graphics, and other means.
- Often use a pipeline approach that employs a variety of modules, breaking the task into smaller sub-tasks.

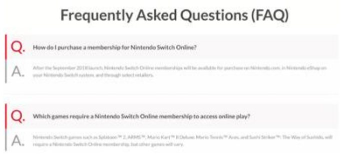

<Recommend a restaurant at China Town.>
ASR
<Inform (destination= China Town)>
NLU
OR
DM
DST
PL
Knowledge Base
TTS
<How many people do you have?>
NLG
<Request (num_people)>
Dialogue System

Question & Answering

- usually answer info-seeking questions posed by humans in natural language

- use different sources of information, such as qa pairs, tables/graphs, documents, or the web, to find the relevant answers

IR-Bot


Frequently Asked Questions (FAQ)

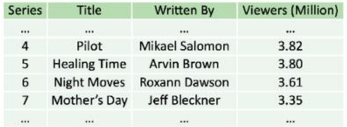
Q. How do I purchase a membership for Nintendo Switch Online?

A.

Q. Which games require a Nintendo Switch Online membership to access online play?

A.

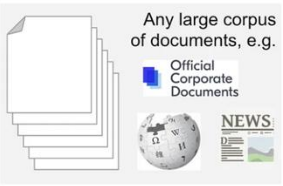
Table QA



Series	Title	Written By	Viewers (Million)
...	...	...	...
4	Pilot	Mikael Salomon	3.82
5	Healing Time	Arvin Brown	3.80
6	Night Moves	Roxann Dawson	3.61
7	Mother's Day	Jeff Bleckner	3.35
...	...	...	...

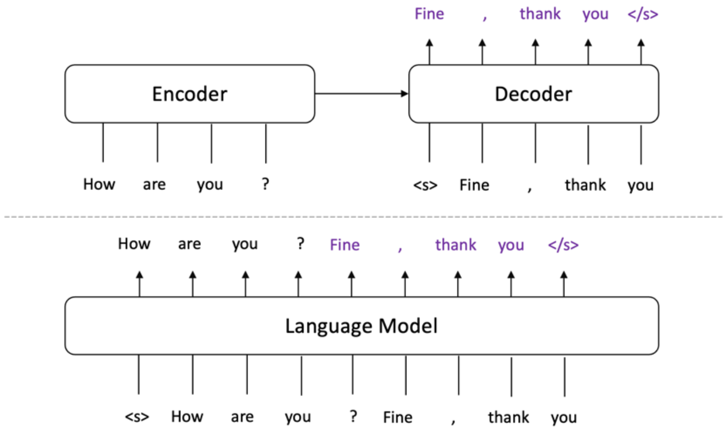
KBQA

DocQA


Any large corpus of documents, e.g.
Official Corporate Documents
NEWS

Open Domain Chit-chat

- a type of conversation that does not have a specific goal or topic, but rather aims to make the dialogue engaging, interesting, and natural
- chatbots that can generate responses to any user input, without being restricted by a predefined domain or task



Pre-trained Large Language Models

Large language models (LLMs): use transformer models and are trained using massive datasets, which enables them to recognize, translate, predict, or generate text or other content

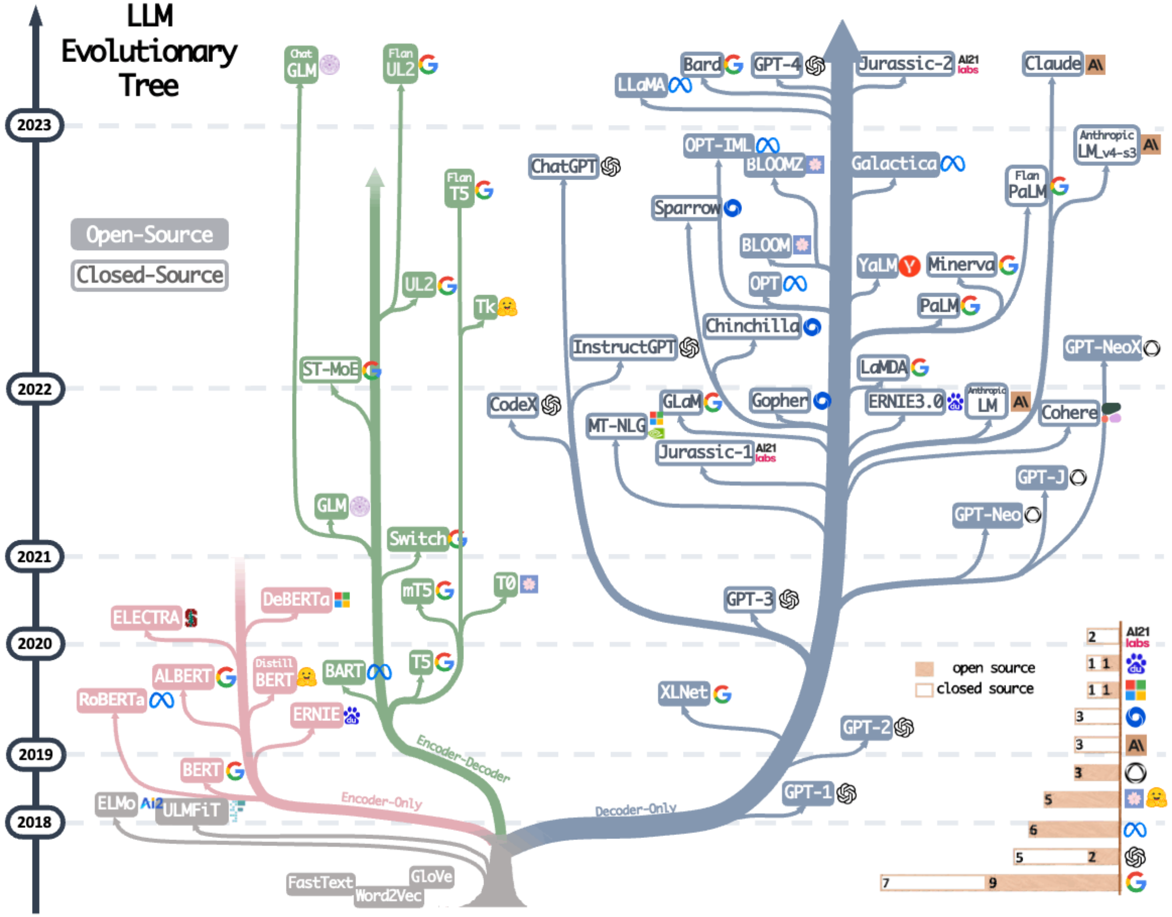
Pre-training: first training a model on one task or dataset, then using the parameters or model from this training to train another model on a different task or dataset

Fine-tuning: the procedure of re-training a pre-trained language model using smaller, task-specific data

Prompt-tuning: adds a small number of tunable embeddings to an otherwise frozen model

In-Context learning: LLMs are prompted with instructions or demonstrations to solve a new task without any additional training

Emergent ability: An ability is emergent if it is not present in smaller models but is present in larger models. [Wei et al, 2022]



Source: https://github.com/Hannibal046/Awesome-LLM

Dialogue Pre-training

Dialog-Post: Multi-Level Self-Supervised Objectives and Hierarchical Model for Dialogue Post-Training
Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen, Xiaodong He
The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)

POSPAN: Position-Constrained Span Masking for Language Model Pre-training
Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen, Xiaodong He
The 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023)

Transformer is All You Need!!!

- The original transformer architecture consisting of an encoder- and decoder part. The internal core parts include the scaled dot product attention mechanism, multi-head attention blocks, and positional input encoding
- With the success of GPT serials, decoder-style architecture and pretraining via next-word prediction become mainstream structure.

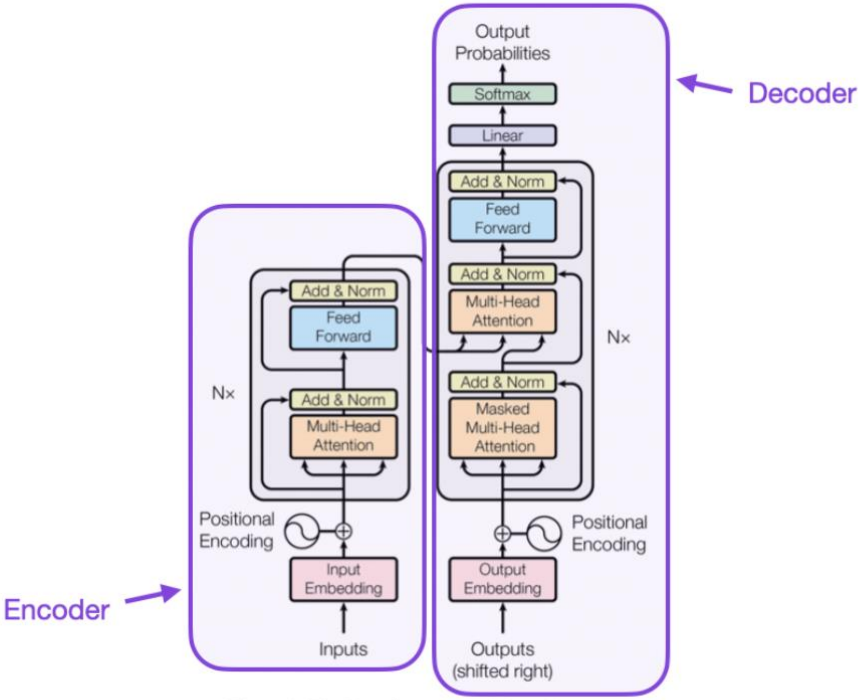
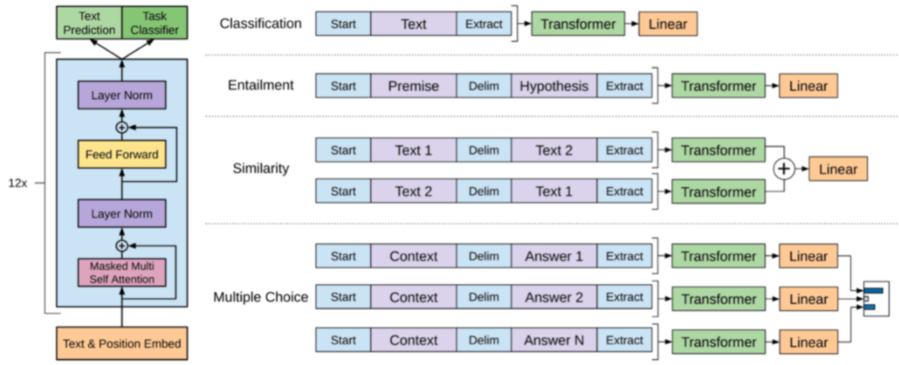


Figure 1: The Transformer - model architecture.

[Vaswani et al, 2017]

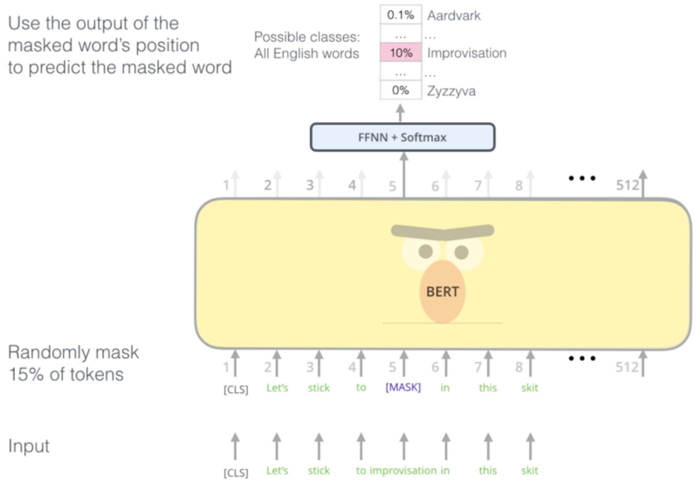
Decoder-style GPT model (originally for predictive modeling)



[Radford et al, 2018]

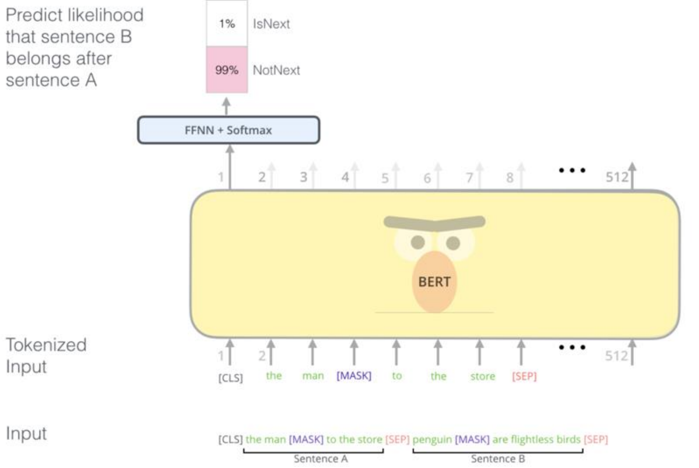
Self-supervised Learning for Pre-training

- Self-supervised learning: supervise using labels generated from the data without any manual or weak label sources


Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
...
10% Improvisation
...
0% Zyzzyva
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Randomly mask 15% of tokens
[CLS] Let's stick to [MASK] in this skit
Input
[CLS] Let's stick to improvisation in this skit

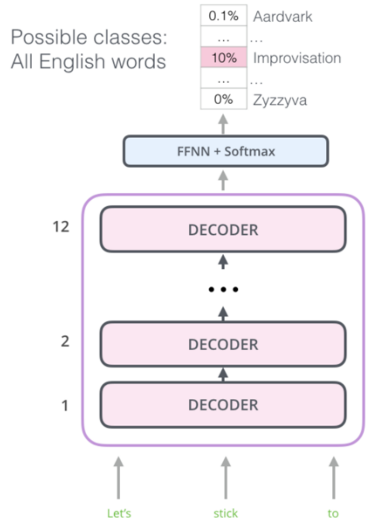
BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Masked Language Modeling (MLM)


Predict likelihood that sentence B belongs after sentence A
1% IsNext
99% NotNext
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Tokenized Input
[CLS] the man [MASK] to the store [SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

Next Sentence Prediction (NSP)



The OpenAI Transformer is now ready to be trained to predict the next word on a dataset made up of 7,000 books.

Next Word Prediction (NWP)

Existing Drawbacks & Motivations

Characteristics of dialogues

- Hierarchical semantic structure (Serban et al., 2016; Xing et al., 2018; Zhang et al., 2019), i.e., dialogue → utterance → token
- Multi-facet attributes (See et al., 2019; Shen et al., 2021a), such as speaker-shift, content-relatedness, fact-awareness, and coherence

Motivations

- How can we improve our modeling of the hierarchical semantic relations in dialogues?
- Is it possible to design auxiliary pre-text tasks that capture the multi-faceted attributes of dialogues?
- With the classic token/span masking method, are we overlooking anything?

HSSA: Hierarchical Segment-wise Self-Attention Network

- HSSA model contains several layers, and each layer is a block consisting of inner-segment self-attention, intersegment self-attention, segment updater, and feedforward sub-layers
- HSSA can reduce the memory cost from $O(n^2)$ to $O(nB+(\frac{n}{B})^2+n)$

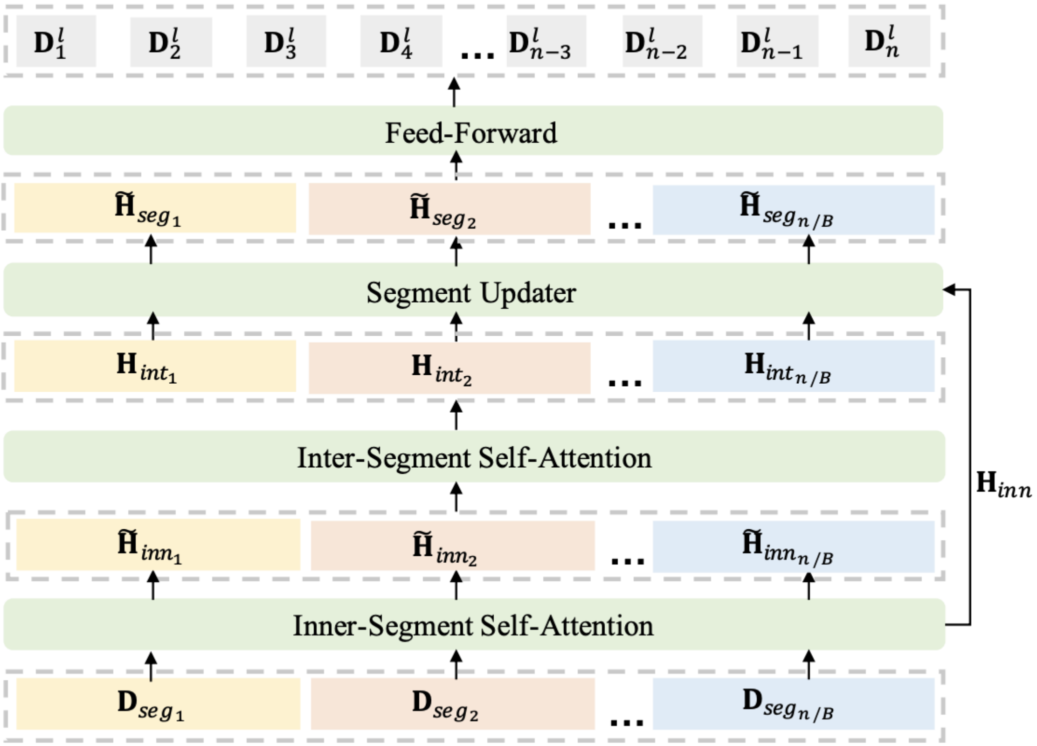


Figure 2: Overview of a HSSA layer.

$$\mathbf{H}_{inn_i} = \mathrm{SA}(\mathbf{D}_{seg_i}) \in \mathbb{R}^{B\times d}$$

$$\mathrm{Agg}(\mathbf{H}_{inn_i}) = \frac{1}{\sum e^{\mathbf{M}_j}} \sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * e^{\mathbf{M}_j},$$

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{\mathrm{Agg}(\mathbf{H}_{inn_i})\mathbf{H}_{inn_{i,j}}^T}{\sqrt{d}}\right), j \in [1, B],$$

$$\tilde{\mathbf{H}}_{inn_i} = \mathbf{W}_p\left(\sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * \alpha_{ij}\right)^T + \mathbf{b}_p,$$

$$\tilde{\mathbf{H}}_{inn} = [\tilde{\mathbf{H}}_{inn_1}, \tilde{\mathbf{H}}_{inn_2}, \ldots, \tilde{\mathbf{H}}_{inn_{n/B}}],$$

$$\mathbf{H}_{int} = \mathrm{SA}(\tilde{\mathbf{H}}_{inn})$$

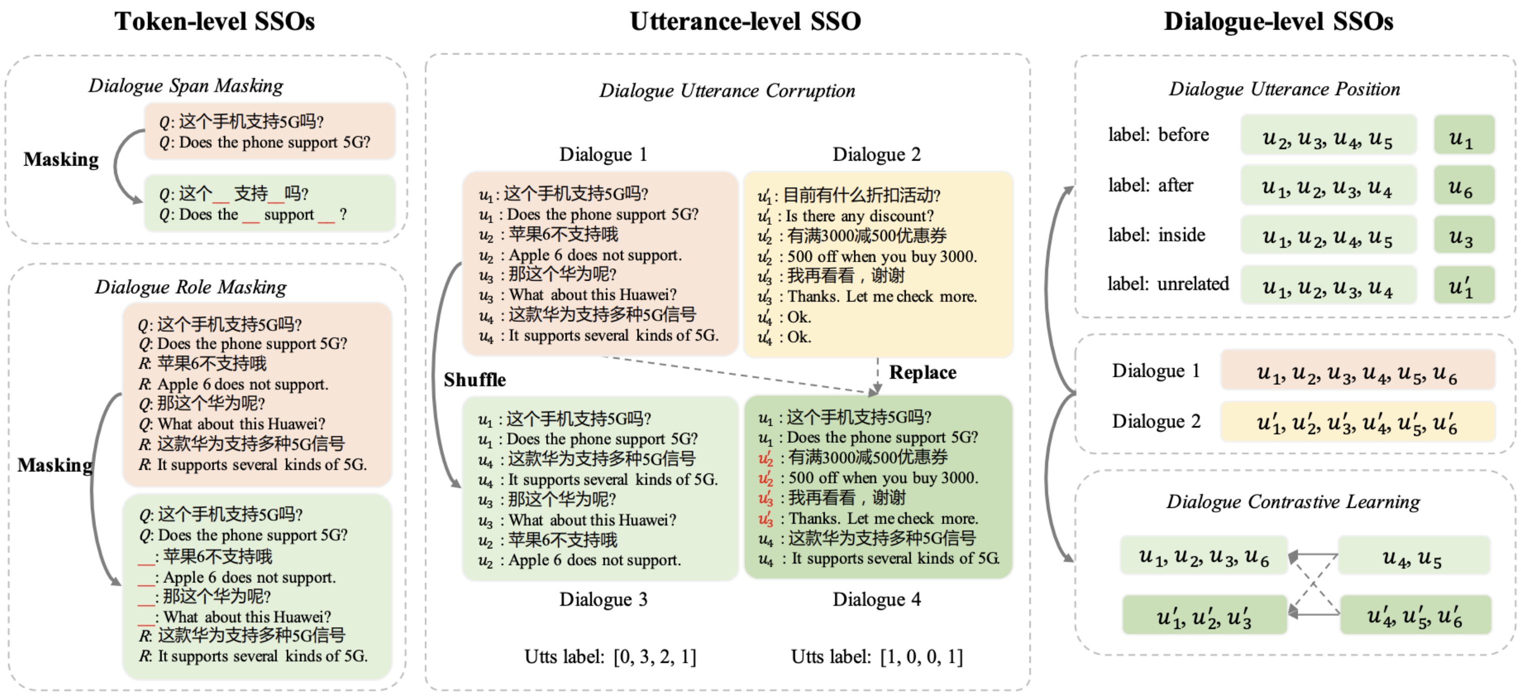
$$\tilde{\mathbf{H}}_{seg_{i,j}} = \beta_{i,j} * \mathbf{H}_{int_i} + \mathbf{H}_{inn_{i,j}},$$

$$\beta_{i,j} = \mathrm{softmax}\left(\frac{\mathbf{H}_{inn_{i,j}}\mathbf{H}_{int_i}^T}{\sqrt{d}}\right), j \in [1, B].$$

SSOs: Multi-level self-supervised objectives

$$\mathcal{L} = \mathcal{L}_{DSM} + \mathcal{L}_{DRM} + \mathcal{L}_{DUC} + \mathcal{L}_{DUP} + \mathcal{L}_{DCL}$$

- We design five multilevel SSOs to post-train the dialogue encoder, which consist of two token-level SSOs, one utterance-level SSO, and two dialogue-level SSOs
- Apply the popular continuous multi-task learning (CMTL) framework for model training, which can pre-train models with multitask objectives efficiently and prevent knowledge forgetting of previous tasks when training with the current task objective(s)



Q: 这个手机支持5G吗?
Q: Does the phone support 5G?

Q: 这个__支持__吗?
Q: Does the __ support __ ?

Q: 这个手机支持5G吗?
Q: Does the phone support 5G?
R: 苹果6不支持哦
R: Apple 6 does not support.
Q: 那这个华为呢?
Q: What about this Huawei?
R: 这款华为支持多种5G信号
R: It supports several kinds of 5G.

Q: 这个手机支持5G吗?
Q: Does the phone support 5G?
__: 苹果6不支持哦
__: Apple 6 does not support.
__: 那这个华为呢?
__: What about this Huawei?
R: 这款华为支持多种5G信号
R: It supports several kinds of 5G.

u_1 : 这个手机支持5G吗?
u_1 : Does the phone support 5G?
u_2 : 苹果6不支持哦
u_2 : Apple 6 does not support.
u_3 : 那这个华为呢?
u_3 : What about this Huawei?
u_4 : 这款华为支持多种5G信号
u_4 : It supports several kinds of 5G.

u'_1 : 目前有什么折扣活动?
u'_1 : Is there any discount?
u'_2 : 有满3000减500优惠券
u'_2 : 500 off when you buy 3000.
u'_3 : 我再看看，谢谢
u'_3 : Thanks. Let me check more.
u'_4 : Ok.
u'_4 : Ok.

u_1 : 这个手机支持5G吗?
u_1 : Does the phone support 5G?
u_4 : 这款华为支持多种5G信号
u_4 : It supports several kinds of 5G.
u_3 : 那这个华为呢?
u_3 : What about this Huawei?
u_2 : 苹果6不支持哦
u_2 : Apple 6 does not support.

u_1 : 这个手机支持5G吗?
u_1 : Does the phone support 5G?
u'_2 : 有满3000减500优惠券
u'_2 : 500 off when you buy 3000.
u'_3 : 我再看看，谢谢
u'_3 : Thanks. Let me check more.
u_4 : 这款华为支持多种5G信号
u_4 : It supports several kinds of 5G.

Figure 1: Illustration of multi-level SSOs in DIALOG-POST. Q and R represent speaker roles. u_i represents utterance. The utterance/dialogue in green color represents the corrupted utterance/dialogue.

POSPAN: Position-Constrained Span Masking

- Existing span masking only considers span length with some discrete distributions, while the dependencies among spans are ignored
- We present POSPAN, a general framework to allow diverse position-constrained span masking strategies via the combination of span length distribution and position constraint distribution

- **Case 1**: There are barely any dependency or semantic relationship between S_i and S_j, i.e., we can predict S_i and S_j independently without knowing each other.
- **Case 2**: $S_i \rightarrow S_j$, i.e., S_i is the premise of S_j. When S_i appears, S_j will appear most of the time.
- **Case 3**: $S_j \rightarrow S_i$, i.e., S_j is the premise of S_i.

$$P(S_i, S_j|R_{ij}) = \frac{P(R_{ij}|S_i, S_j) * P(S_i, S_j)}{P(R_{ij})},$$
$$\log P(S_i, S_j|R_{ij}) \propto \underbrace{\log P(R_{ij}|S_i, S_j)}_{①} + \underbrace{\log P(S_i, S_j)}_{②}, \tag{1}$$

$$\sum_{i,j} \log P(S_i, S_j) = \frac{(M-1)\log P(S_1, S_2, ..., S_M)}{2}$$
$$\propto \sum_{i=1}^{M} \log P(S_i).$$
$$\mathcal{L}_S :\rightarrow max(\mathbb{E}[\log P(S_i|len_i)]),\ where \tag{2}$$
$$\mathbb{E}[\log P(S_i|len_i)] = \mathbb{E}_{len_i \sim F_M}(\sum_{l=0}^{len_i-1} \log P(x_{i+l})).$$

$$P(R_{ij}|d) = P(R_{ij}|x_{pos_i-1}, ..., x_{pos_j+len_j})$$
$$= \frac{P(x_{pos_i-1}, ..., x_{pos_j+len_j}|R_{ij}) * P(R_{ij})}{P(x_{pos_i-1}, ..., x_{pos_j+len_j})}$$
$$= P(R_{ij}) * \frac{P(S_i, S_j|R_{ij}) * \prod_{k=pos_j-d}^{pos_j-1} P(x_k|R_{ij})}{P(S_i, S_j) * \prod_{k=pos_j-d}^{pos_j-1} P(x_k)} \tag{3}$$
$$= P(R_{ij}|S_i, S_j) * \frac{\prod_{k=pos_j-d}^{pos_j-1} P(x_k|R_{ij})}{\prod_{k=pos_j-d}^{pos_j-1} P(x_k)}$$
$$where\ d \sim F_D.$$

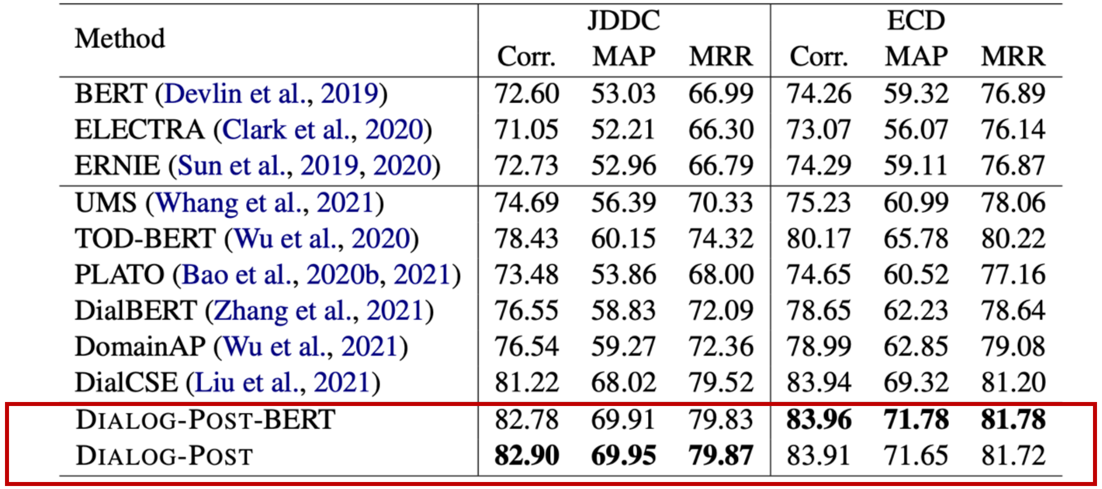
$$P(R_{ij}|S_i, S_j) \propto P(R_{ij}|d). \tag{4}$$

Finally, the pre-training with masked language modeling can be decomposed into two losses:

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_S,$$
$$\mathcal{L}_R :\rightarrow max(\mathbb{E}[\log P(R_{ij}|F_D)]), \tag{5}$$

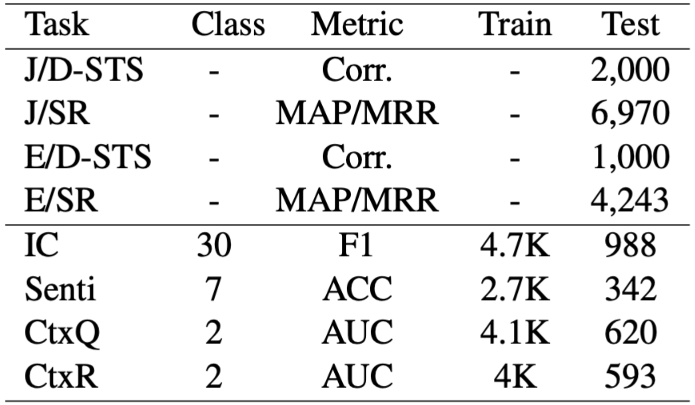
Experiments

- **Datasets**
  - Pre-training: JDDC (Chen et al., 2020) and ECD (Zhang et al., 2018)
  - POSPAN: 9 public NLU tasks
- **Evaluation**
  - Dialogue Representation Evaluation: SR & STS
  - Dialogue Understanding Evaluation: IC, Senti, CtxQ, CtxR



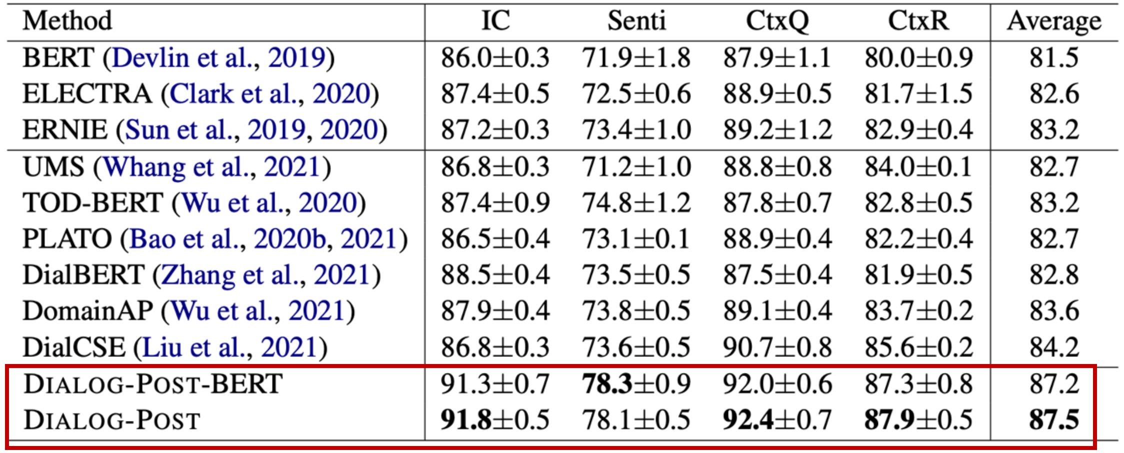
Method	JDDC			ECD		
	Corr.	MAP	MRR	Corr.	MAP	MRR
BERT (Devlin et al., 2019)	72.60	53.03	66.99	74.26	59.32	76.89
ELECTRA (Clark et al., 2020)	71.05	52.21	66.30	73.07	56.07	76.14
ERNIE (Sun et al., 2019, 2020)	72.73	52.96	66.79	74.29	59.11	76.87
UMS (Whang et al., 2021)	74.69	56.39	70.33	75.23	60.99	78.06
TOD-BERT (Wu et al., 2020)	78.43	60.15	74.32	80.17	65.78	80.22
PLATO (Bao et al., 2020b, 2021)	73.48	53.86	68.00	74.65	60.52	77.16
DialBERT (Zhang et al., 2021)	76.55	58.83	72.09	78.65	62.23	78.64
DomainAP (Wu et al., 2021)	76.54	59.27	72.36	78.99	62.85	79.08
DialCSE (Liu et al., 2021)	81.22	68.02	79.52	83.94	69.32	81.20
DIALOG-POST-BERT	82.78	69.91	79.83	**83.96**	**71.78**	**81.78**
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	81.72

Table 2: Evaluation results on semantic retrieval (SR) and dialogue-based semantic textual similarity (D-STS) tasks.



Task	Class	Metric	Train	Test
J/D-STS	-	Corr.	-	2,000
J/SR	-	MAP/MRR	-	6,970
E/D-STS	-	Corr.	-	1,000
E/SR	-	MAP/MRR	-	4,243
IC	30	F1	4.7K	988
Senti	7	ACC	2.7K	342
CtxQ	2	AUC	4.1K	620
CtxR	2	AUC	4K	593

Table 3: Details of evaluation tasks. "J" and "E" represent JDDC and ECD.



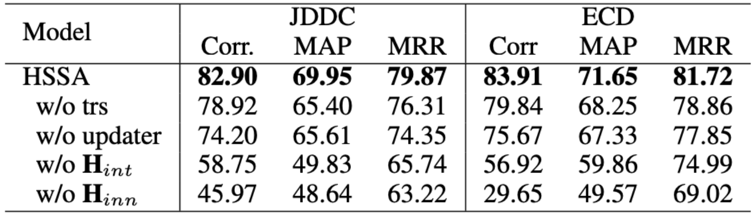
Method	IC	Senti	CtxQ	CtxR	Average
BERT (Devlin et al., 2019)	86.0±0.3	71.9±1.8	87.9±1.1	80.0±0.9	81.5
ELECTRA (Clark et al., 2020)	87.4±0.5	72.5±0.6	88.9±0.5	81.7±1.5	82.6
ERNIE (Sun et al., 2019, 2020)	87.2±0.3	73.4±1.0	89.2±1.2	82.9±0.4	83.2
UMS (Whang et al., 2021)	86.8±0.3	71.2±1.0	88.8±0.8	84.0±0.1	82.7
TOD-BERT (Wu et al., 2020)	87.4±0.9	74.8±1.2	87.8±0.7	82.8±0.5	83.2
PLATO (Bao et al., 2020b, 2021)	86.5±0.4	73.1±0.1	88.9±0.4	82.2±0.4	82.7
DialBERT (Zhang et al., 2021)	88.5±0.4	73.5±0.5	87.5±0.4	81.9±0.5	82.8
DomainAP (Wu et al., 2021)	87.9±0.4	73.8±0.5	89.1±0.4	83.7±0.2	83.6
DialCSE (Liu et al., 2021)	86.8±0.3	73.6±0.5	90.7±0.8	85.6±0.2	84.2
DIALOG-POST-BERT	91.3±0.7	**78.3**±0.9	92.0±0.6	87.3±0.8	87.2
DIALOG-POST	**91.8**±0.5	78.1±0.5	**92.4**±0.7	**87.9**±0.5	**87.5**

Table 4: Evaluation results on dialogue understanding tasks (all with significance value $p < 0.05$).

Ablation Study

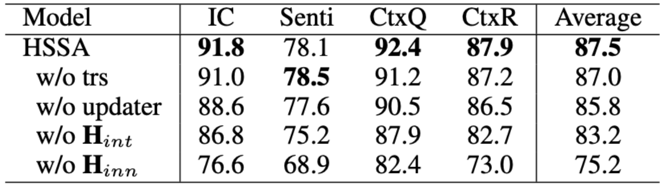
Ablation of HSSA

- We stack 10 layers of HSSA blocks and 2 layers of Transformer blocks, the last 2 Transformer layers are devised to capture the full dialogue semantics based on the global self-attention (SA) mechanism. Here, we first replace the last 2 Transformer layers with 2 HSSA layers (denoted as “w/o trs”)
- The performance of Senti becomes slightly better with all HSSA blocks. Since the input of Senti task is an utterance without context, it is possible that the 12-layer HSSA focusing on the local attention has some advantages



Model	JDDC Corr.	MAP	MRR	ECD Corr	MAP	MRR
HSSA	**82.90**	**69.95**	**79.87**	**83.91**	**71.65**	**81.72**
w/o trs	78.92	65.40	76.31	79.84	68.25	78.86
w/o updater	74.20	65.61	74.35	75.67	67.33	77.85
w/o $\mathbf{H}_{int}$	58.75	49.83	65.74	56.92	59.86	74.99
w/o $\mathbf{H}_{inn}$	45.97	48.64	63.22	29.65	49.57	69.02

Table 9: Experimental results of HSSA Ablation Study on all dialogue representation tasks.

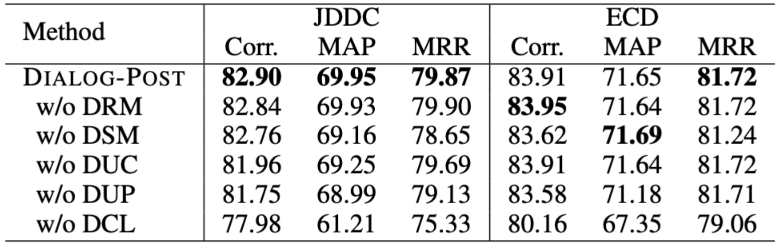


Model	IC	Senti	CtxQ	CtxR	Average
HSSA	**91.8**	78.1	**92.4**	**87.9**	**87.5**
w/o trs	91.0	**78.5**	91.2	87.2	87.0
w/o updater	88.6	77.6	90.5	86.5	85.8
w/o $\mathbf{H}_{int}$	86.8	75.2	87.9	82.7	83.2
w/o $\mathbf{H}_{inn}$	76.6	68.9	82.4	73.0	75.2

Table 10: Experimental results of HSSA Ablation Study on all dialogue understanding tasks.

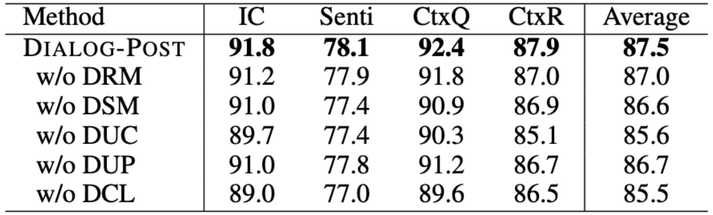
Ablation of SSOs

- We remove one training objective each time while keeping the remaining four, each training objective contributes to the overall performance to some extent, indicating the multi-level SSOs are complementary
- DCL brings the most benefits, which implies the effectiveness of DCL on capturing the content-relatedness of context-context pairs



Method	JDDC Corr.	MAP	MRR	ECD Corr.	MAP	MRR
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	**81.72**
w/o DRM	82.84	69.93	79.90	**83.95**	71.64	81.72
w/o DSM	82.76	69.16	78.65	83.62	**71.69**	81.24
w/o DUC	81.96	69.25	79.69	83.91	71.64	81.72
w/o DUP	81.75	68.99	79.13	83.58	71.18	81.71
w/o DCL	77.98	61.21	75.33	80.16	67.35	79.06

Table 11: Experimental results of SSOs Ablation Study on all dialogue representation tasks.

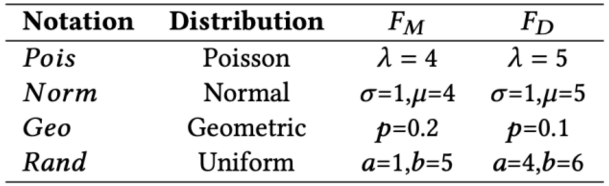


Method	IC	Senti	CtxQ	CtxR	Average
DIALOG-POST	**91.8**	**78.1**	**92.4**	**87.9**	**87.5**
w/o DRM	91.2	77.9	91.8	87.0	87.0
w/o DSM	91.0	77.4	90.9	86.9	86.6
w/o DUC	89.7	77.4	90.3	85.1	85.6
w/o DUP	91.0	77.8	91.2	86.7	86.7
w/o DCL	89.0	77.0	89.6	86.5	85.5

Table 12: Experimental results of SSOs Ablation Study on all dialogue understanding tasks.

Experimental Results of POSPAN

- All post-training models bring further improvements compared to the strong baseline DeBERTaV3, which shows the effectiveness of post-training
- Compared with single-token masking, all span-level masking methods yield substantial improvements, indicating the advantage of span-level masking on capturing the critical semantics of language.
- POSPAN obtained the best performance across different tasks, which demonstrates the superiority and necessity of position constraint for span masking



Notation	Distribution	F_M	F_D
Pois	Poisson	$\lambda = 4$	$\lambda = 5$
Norm	Normal	σ=1,μ=4	σ=1,μ=5
Geo	Geometric	p=0.2	p=0.1
Rand	Uniform	a=1,b=5	a=4,b=6

Table 1: Hyper-parameters of different distributions. We tune hyper-parameters of the distributions via grid search and find the best settings.

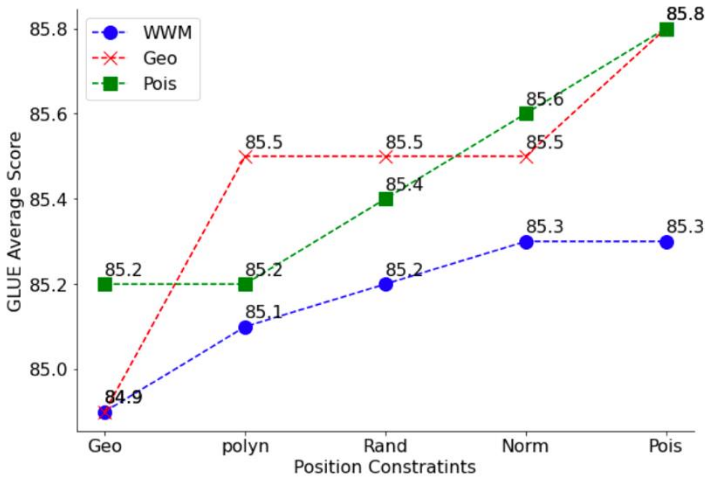
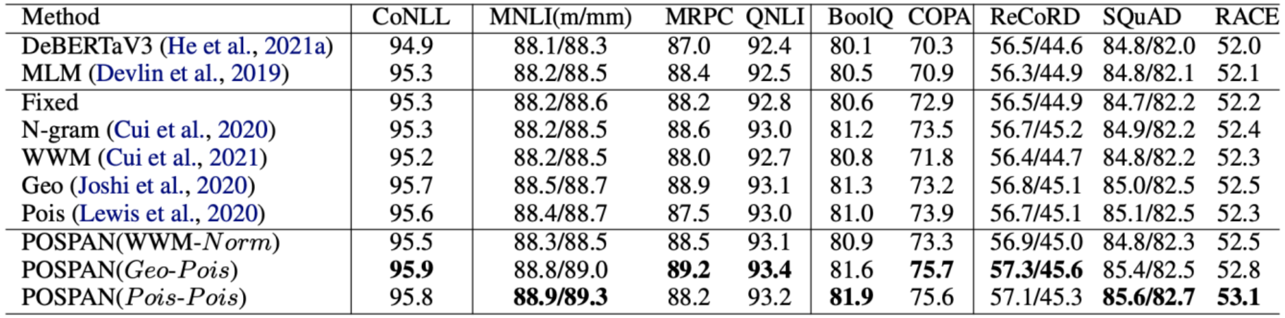
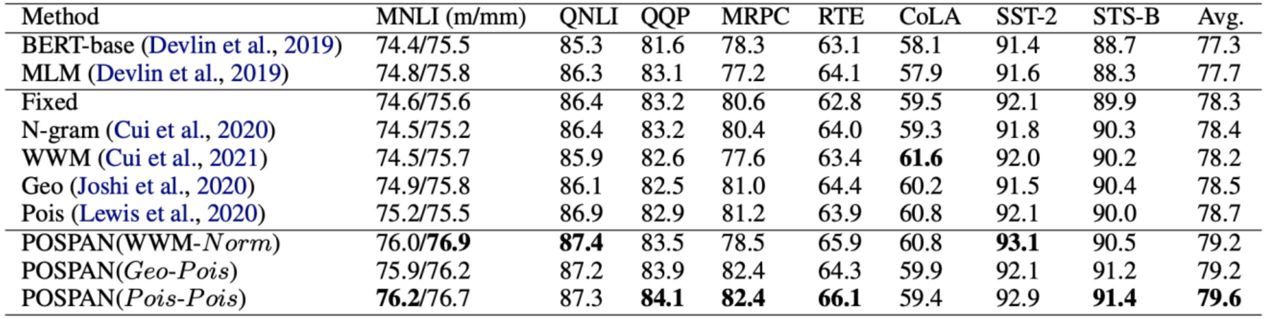


Figure 1: The model performance of POSPAN with different position constraints (x-axis).



Method	CoNLL	MNLI(m/mm)	MRPC	QNLI	BoolQ	COPA	ReCoRD	SQuAD	RACE
DeBERTaV3 (He et al., 2021a)	94.9	88.1/88.3	87.0	92.4	80.1	70.3	56.5/44.6	84.8/82.0	52.0
MLM (Devlin et al., 2019)	95.3	88.2/88.5	88.4	92.5	80.5	70.9	56.3/44.9	84.8/82.1	52.1
Fixed	95.3	88.2/88.6	88.2	92.8	80.6	72.9	56.5/44.9	84.7/82.2	52.2
N-gram (Cui et al., 2020)	95.3	88.2/88.5	88.6	93.0	81.2	73.5	56.7/45.2	84.9/82.2	52.4
WWM (Cui et al., 2021)	95.2	88.2/88.5	88.0	92.7	80.8	71.8	56.4/44.7	84.8/82.2	52.3
Geo (Joshi et al., 2020)	95.7	88.5/88.7	88.9	93.1	81.3	73.2	56.8/45.1	85.0/82.5	52.5
Pois (Lewis et al., 2020)	95.6	88.4/88.7	87.5	93.0	81.0	73.9	56.7/45.1	85.1/82.5	52.3
POSPAN(WWM-*Norm*)	95.5	88.3/88.5	88.5	93.1	80.9	73.3	56.9/45.0	84.8/82.3	52.5
POSPAN(*Geo-Pois*)	**95.9**	88.8/89.0	**89.2**	**93.4**	81.6	**75.7**	**57.3/45.6**	85.4/82.5	52.8
POSPAN(*Pois-Pois*)	95.8	**88.9/89.3**	88.2	93.2	**81.9**	75.6	57.1/45.3	**85.6/82.7**	**53.1**

Table 2: Experimental results of POSPAN. POSPAN(*Geo-Pois*) denotes $F_M \sim Geo$ and $F_D \sim Pois$. CoNLL and SQuAD represent ConNLL 2003 and SQuAD v2.0. MNLI (m/mm) represents the two versions of MNLI, MNLI-matched and MNLI-mismatched. The complete evaluation results are reported in Appendix A.4.

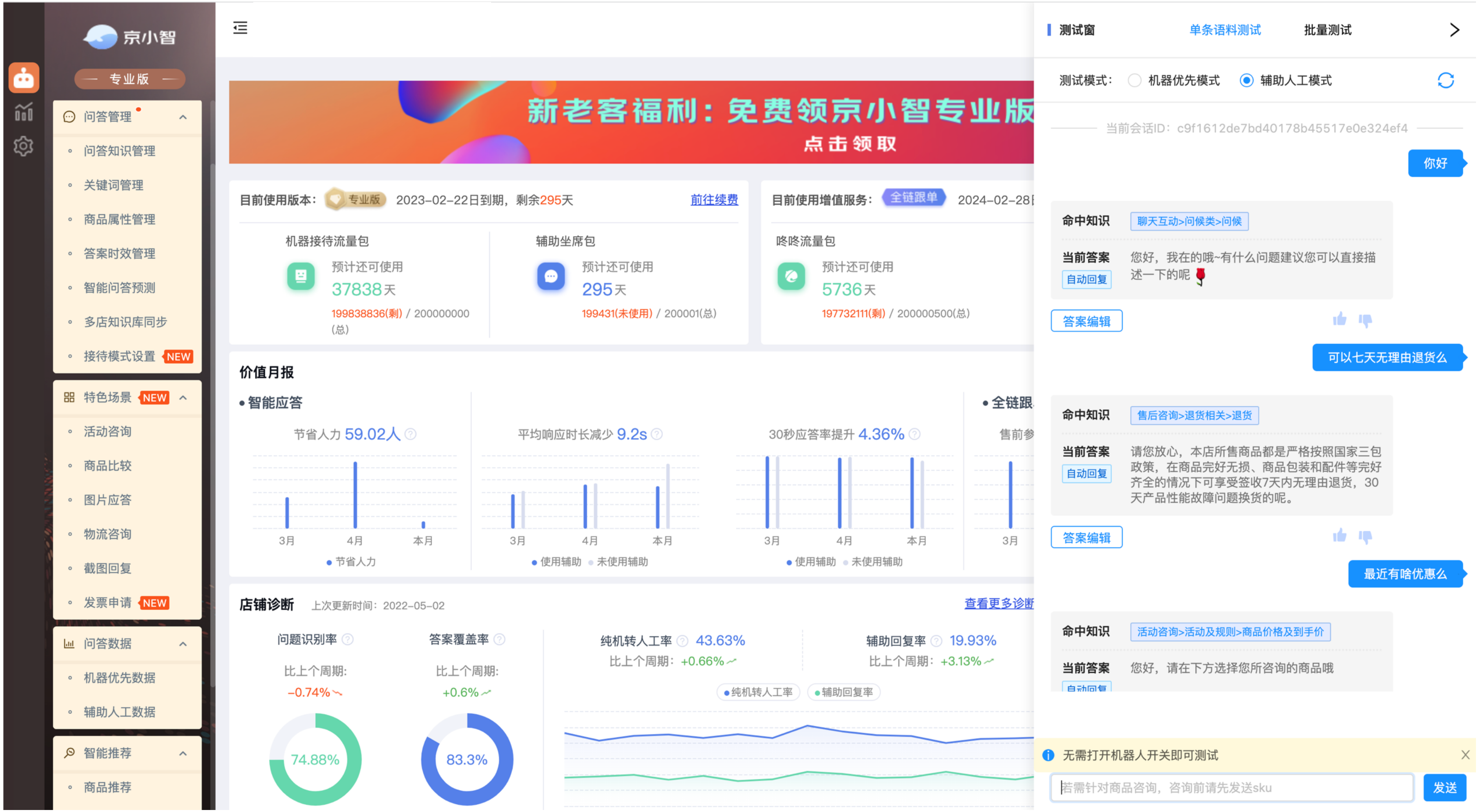


Method	MNLI (m/mm)	QNLI	QQP	MRPC	RTE	CoLA	SST-2	STS-B	Avg.
BERT-base (Devlin et al., 2019)	74.4/75.5	85.3	81.6	78.3	63.1	58.1	91.4	88.7	77.3
MLM (Devlin et al., 2019)	74.8/75.8	86.3	83.1	77.2	64.1	57.9	91.6	88.3	77.7
Fixed	74.6/75.6	86.4	83.2	80.6	62.8	59.5	92.1	89.9	78.3
N-gram (Cui et al., 2020)	74.5/75.2	86.4	83.2	80.4	64.0	59.3	91.8	90.3	78.4
WWM (Cui et al., 2021)	74.5/75.7	85.9	82.6	77.6	63.4	**61.6**	92.0	90.2	78.2
Geo (Joshi et al., 2020)	74.9/75.8	86.1	82.5	81.0	64.4	60.2	91.5	90.4	78.5
Pois (Lewis et al., 2020)	75.2/75.5	86.9	82.9	81.2	63.9	60.8	92.1	90.0	78.7
POSPAN(WWM-*Norm*)	76.0/**76.9**	**87.4**	83.5	78.5	65.9	60.8	**93.1**	90.5	79.2
POSPAN(*Geo-Pois*)	75.9/76.2	87.2	83.9	82.4	64.3	59.9	92.1	91.2	79.2
POSPAN(*Pois-Pois*)	**76.2**/76.7	87.3	**84.1**	**82.4**	**66.1**	59.4	92.9	**91.4**	**79.6**

Table 7: Experimental results of POSPAN in GLUE with BERT as base model. POSPAN(*Geo-Pois*) denotes $F_M \sim Geo$ and $F_D \sim Pois$. MNLI (m/mm) represents the two versions of MNLI, MNLI-matched and MNLI-mismatched.

Pre-sales E-commerce Chatbot

JD AlphaSales



目前使用版本：专业版 2023-02-22日到期，剩余295天

目前使用增值服务：全链跟单 2024-02-28

机器接待流量包
预计还可使用
37838天
199838836(剩) / 200000000(总)

辅助坐席包
预计还可使用
295天
199431(未使用) / 200001(总)

咚咚流量包
预计还可使用
5736天
197732111(剩) / 200000500(总)

价值月报

- 智能应答

节省人力 59.02人

平均响应时长减少 9.2s

30秒应答率提升 4.36%

店铺诊断 上次更新时间：2022-05-02

问题识别率
比上个周期：
-0.74%
74.88%

答案覆盖率
比上个周期：
+0.6%
83.3%

纯机转人工率 43.63%
比上个周期：+0.66%

辅助回复率 19.93%
比上个周期：+3.13%

测试模式：○ 机器优先模式 ● 辅助人工模式

当前会话ID：c9f1612de7bd40178b45517e0e324ef4

你好

命中知识 聊天互动>问候类>问候
当前答案 自动回复 您好，我在的哦~有什么问题建议您可以直接描述一下的呢🌹

可以七天无理由退货么

命中知识 售后咨询>退货相关>退货
当前答案 自动回复 请您放心，本店所售商品都是严格按照国家三包政策，在商品完好无损、商品包装和配件等完好齐全的情况下可享受签收7天内无理由退货，30天产品性能故障问题换货的呢。

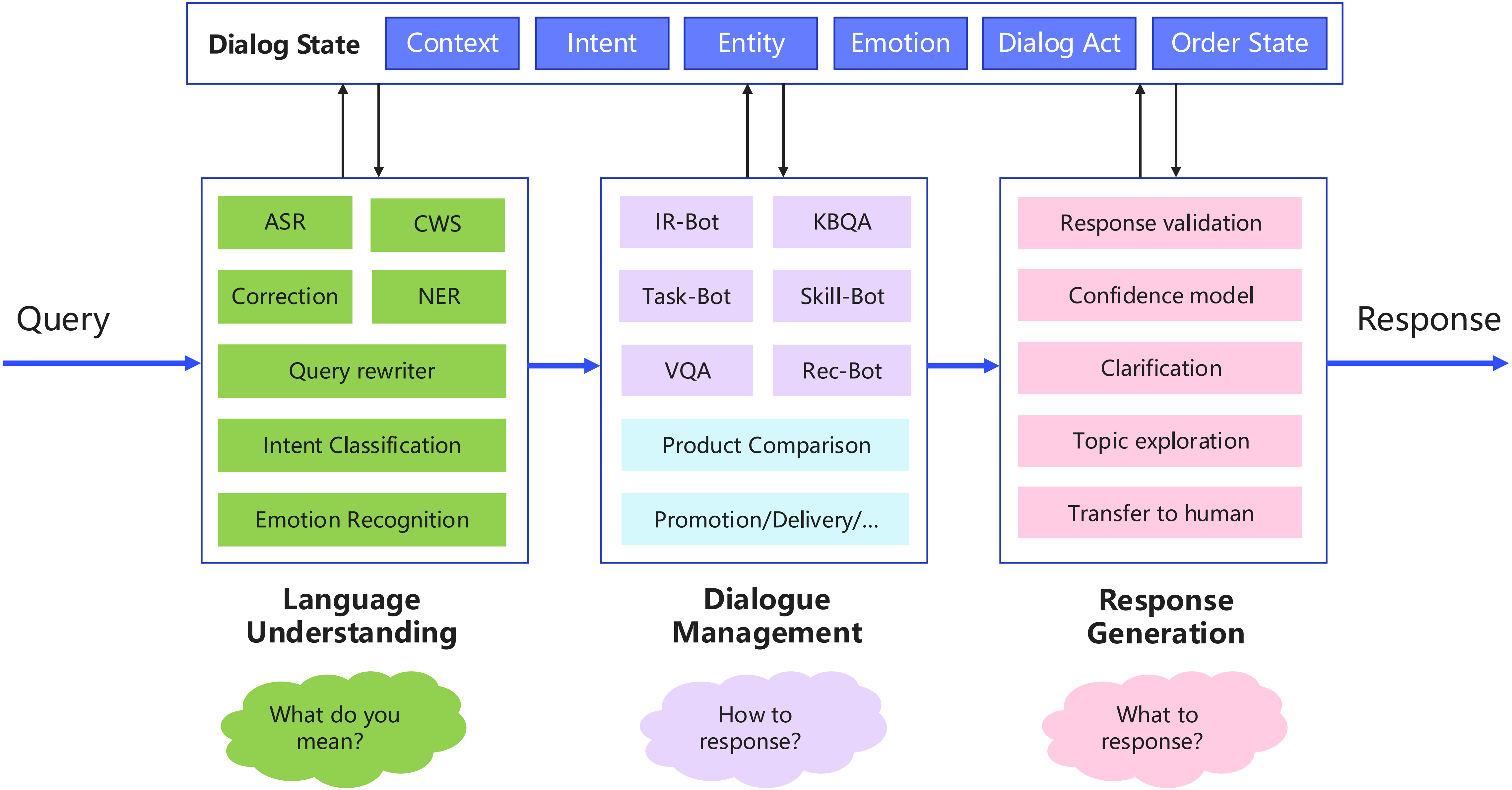
最近有啥优惠么

命中知识 活动咨询>活动及规则>商品价格及到手价
当前答案 您好，请在下方选择您所咨询的商品哦

无需打开机器人开关即可测试

若需针对商品咨询，咨询前请先发送sku

Architecture of Dialogue System



Dialog State

Context

Intent

Entity

Emotion

Dialog Act

Order State

Query

ASR

CWS

Correction

NER

Query rewriter

Intent Classification

Emotion Recognition

IR-Bot

KBQA

Task-Bot

Skill-Bot

VQA

Rec-Bot

Product Comparison

Promotion/Delivery/…

Response validation

Confidence model

Clarification

Topic exploration

Transfer to human

Response

Language Understanding

Dialogue Management

Response Generation

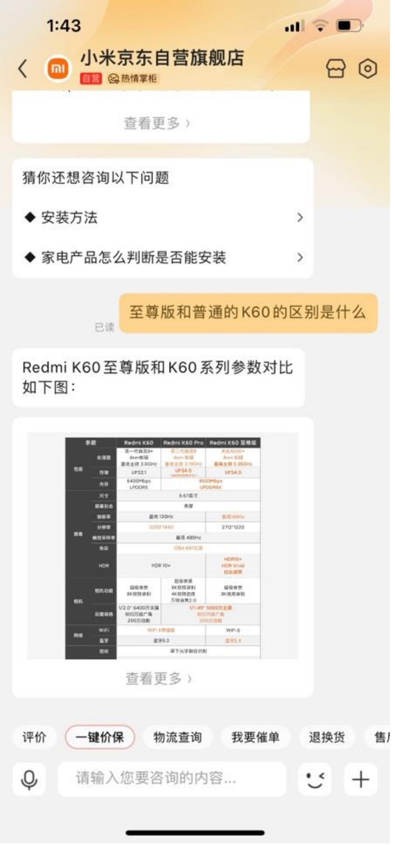
What do you mean?

How to response?

What to response?

Main Features

Product Comparison



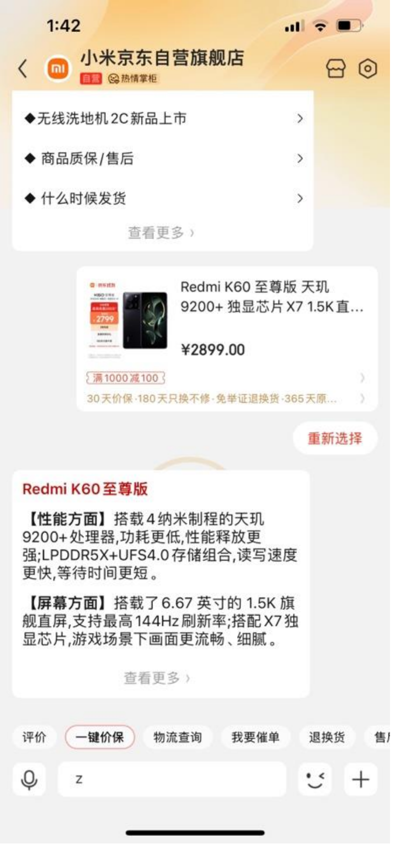
猜你还想咨询以下问题

- ◆ 安装方法
- ◆ 家电产品怎么判断是否能安装

至尊版和普通的K60的区别是什么

Redmi K60至尊版和K60系列参数对比如下图：

Selling Point Generation



- ◆无线洗地机2C新品上市
- ◆ 商品质保/售后
- ◆ 什么时候发货

Redmi K60 至尊版 天玑 9200+ 独显芯片 X7 1.5K 直...

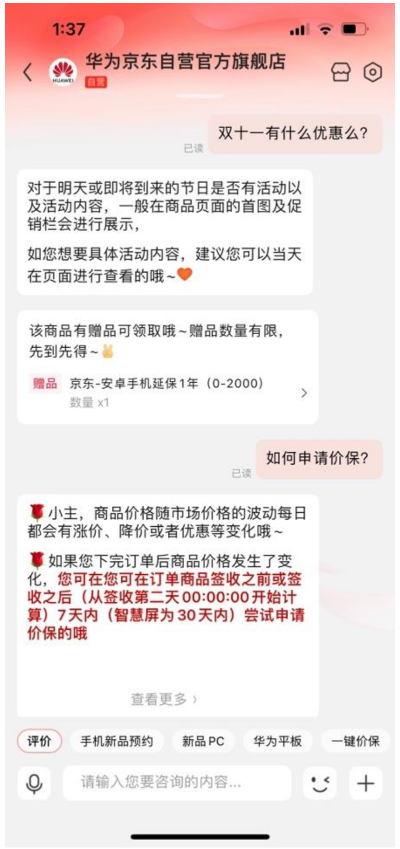
¥2899.00

Redmi K60至尊版

【性能方面】搭载4纳米制程的天玑9200+处理器,功耗更低,性能释放更强;LPDDR5X+UFS4.0存储组合,读写速度更快,等待时间更短。

【屏幕方面】搭载了6.67英寸的1.5K旗舰直屏,支持最高144Hz刷新率;搭配X7独显芯片,游戏场景下画面更流畅、细腻。

Promotion Checking



双十一有什么优惠么?

对于明天或即将到来的节日是否有活动以及活动内容，一般在商品页面的首图及促销栏会进行展示，

如您想要具体活动内容，建议您可以当天在页面进行查看的哦~🧡

该商品有赠品可领取哦~赠品数量有限，先到先得~✌

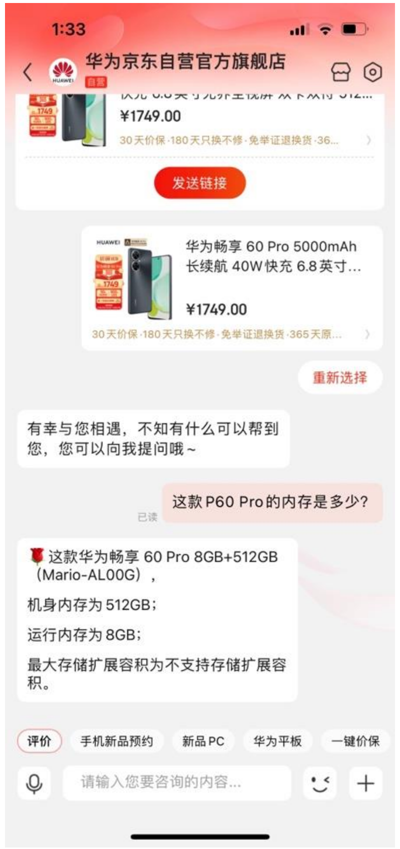
京东-安卓手机延保1年（0-2000）
数量 x1

如何申请价保?

🌹小主，商品价格随市场价格的波动每日都会有涨价、降价或者优惠等变化哦~

🌹如果您下完订单后商品价格发生了变化，**您可在您可在订单商品签收之前或签收之后（从签收第二天00:00:00开始计算）7天内（智慧屏为30天内）尝试申请价保的哦**

Attribute Inquiry



¥1749.00

华为畅享 60 Pro 5000mAh 长续航 40W快充 6.8英寸...

¥1749.00

有幸与您相遇，不知有什么可以帮到您，您可以向我提问哦~

这款P60 Pro的内存是多少?

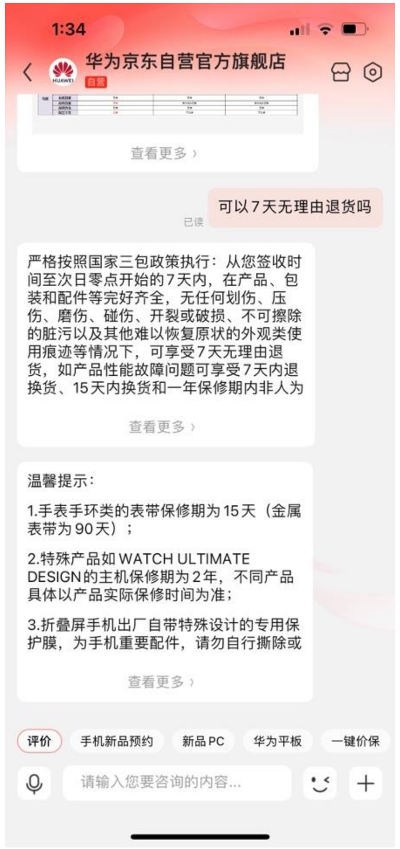
🌹这款华为畅享 60 Pro 8GB+512GB（Mario-AL00G），

机身内存为512GB;

运行内存为8GB;

最大存储扩展容积为不支持存储扩展容积。

Policy Inquiry



可以7天无理由退货吗

严格按照国家三包政策执行：从您签收时间至次日零点开始的7天内，在产品、包装和配件等完好齐全，无任何划伤、压伤、磨伤、碰伤、开裂或破损、不可擦除的脏污以及其他难以恢复原状的外观类使用痕迹等情况下，可享受7天无理由退货，如产品性能故障问题可享受7天内退换货、15天内换货和一年保修期内非人为

温馨提示：

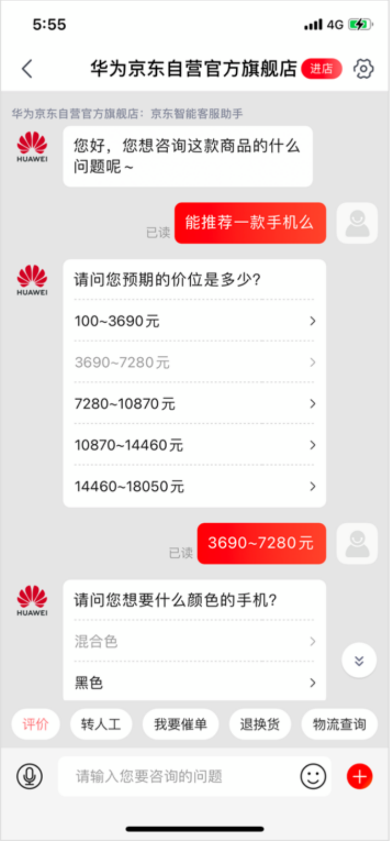
1.手表手环类的表带保修期为15天（金属表带为90天）；

2.特殊产品如WATCH ULTIMATE DESIGN的主机保修期为2年，不同产品具体以产品实际保修时间为准；

3.折叠屏手机出厂自带特殊设计的专用保护膜，为手机重要配件，请勿自行撕除或

Main Features (Cont.)

Conversational Recommendation



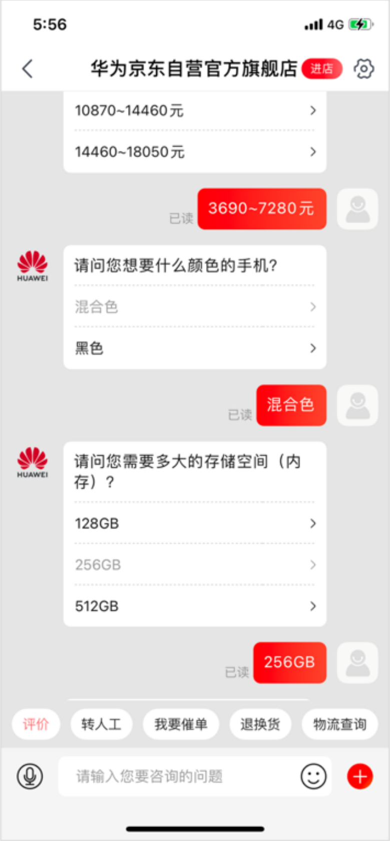
您好，您想咨询这款商品的什么问题呢～

能推荐一款手机么

请问您预期的价位是多少？
- 100~3690元
- 3690~7280元
- 7280~10870元
- 10870~14460元
- 14460~18050元

3690~7280元

请问您想要什么颜色的手机？
- 混合色
- 黑色



- 10870~14460元
- 14460~18050元

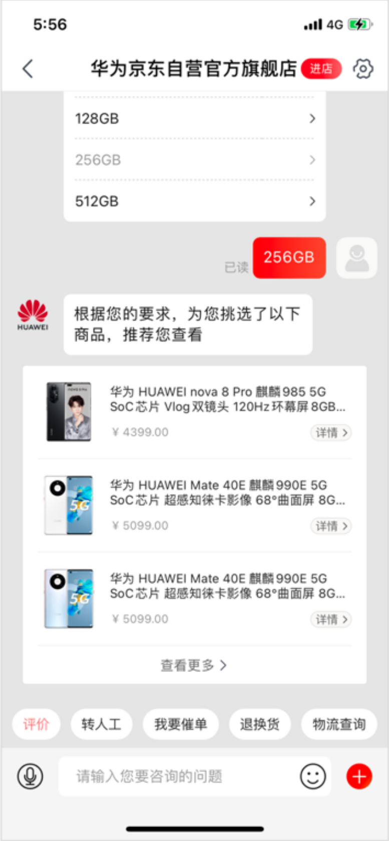
3690~7280元

请问您想要什么颜色的手机？
- 混合色
- 黑色

混合色

请问您需要多大的存储空间（内存）？
- 128GB
- 256GB
- 512GB

256GB



- 128GB
- 256GB
- 512GB

256GB

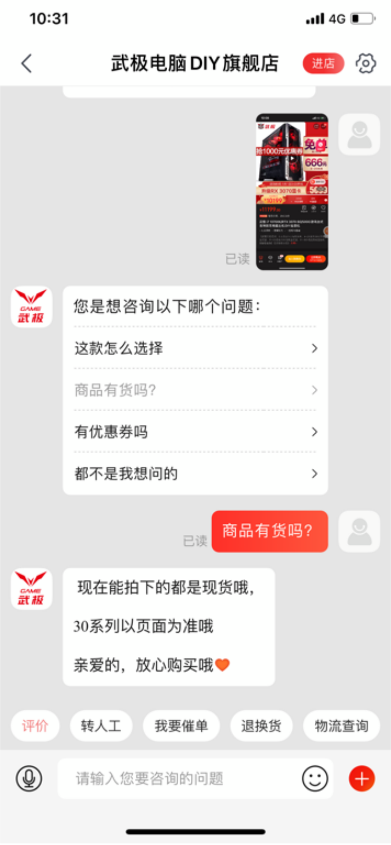
根据您的要求，为您挑选了以下商品，推荐您查看

华为 HUAWEI nova 8 Pro 麒麟985 5G SoC芯片 Vlog双镜头 120Hz环幕屏 8GB...
¥ 4399.00

华为 HUAWEI Mate 40E 麒麟990E 5G SoC芯片 超感知徕卡影像 68°曲面屏 8G...
¥ 5099.00

华为 HUAWEI Mate 40E 麒麟990E 5G SoC芯片 超感知徕卡影像 68°曲面屏 8G...
¥ 5099.00

Multimodal Conversation



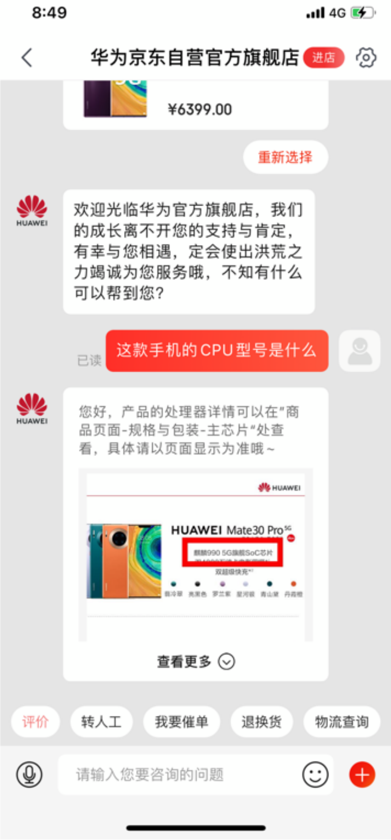
您是想咨询以下哪个问题：
- 这款怎么选择
- 商品有货吗？
- 有优惠券吗
- 都不是我想问的

商品有货吗？

现在能拍下的都是现货哦，

30系列以页面为准哦

亲爱的，放心购买哦❤



¥6399.00

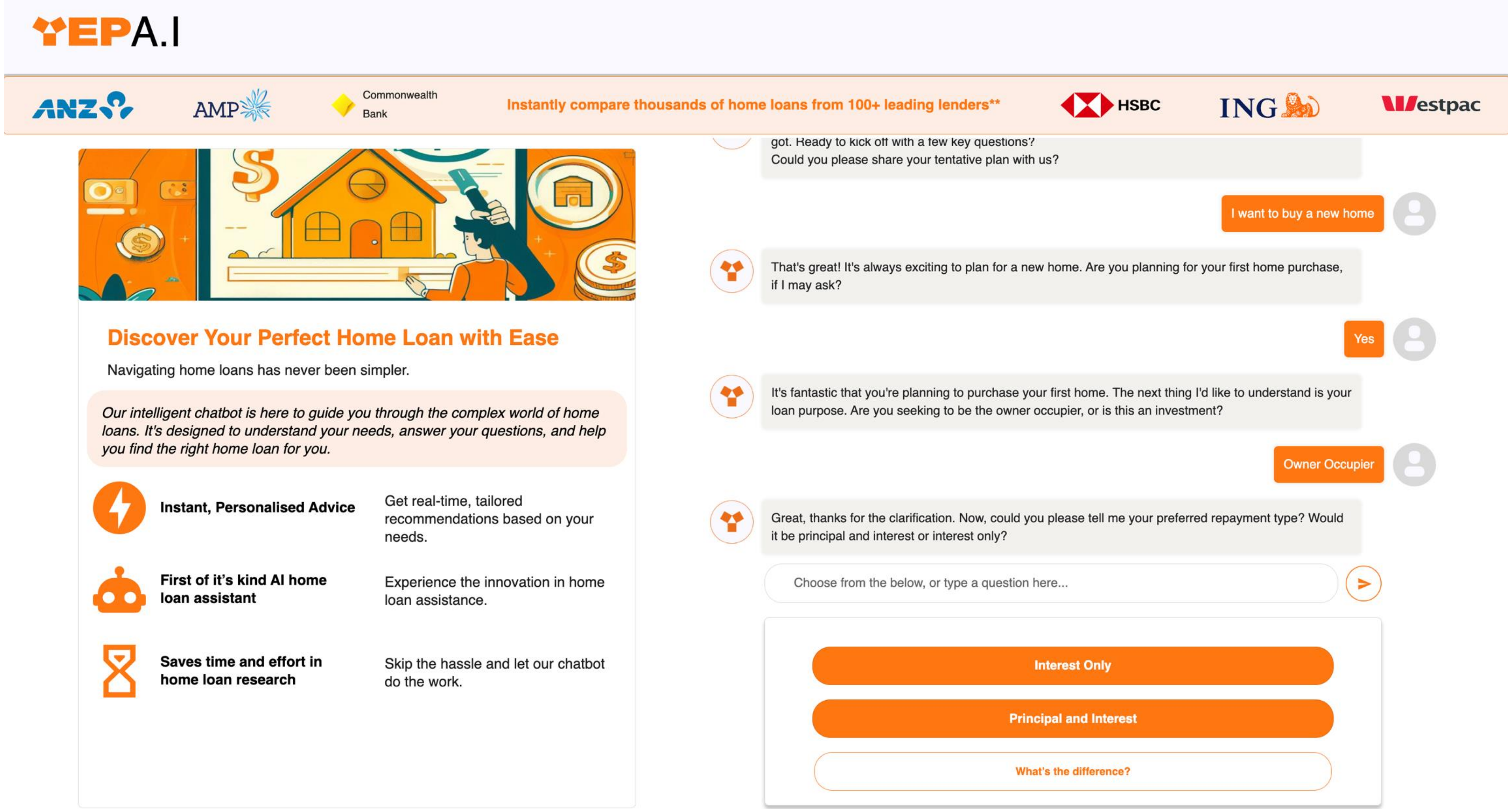
欢迎光临华为官方旗舰店，我们的成长离不开您的支持与肯定，有幸与您相遇，定会使出洪荒之力竭诚为您服务哦，不知有什么可以帮到您？

这款手机的CPU型号是什么

您好，产品的处理器详情可以在"商品页面-规格与包装-主芯片"处查看，具体请以页面显示为准哦～

Financial Home Loan Chatbot

LLM-based AI Home Loan



Instantly compare thousands of home loans from 100+ leading lenders**

Discover Your Perfect Home Loan with Ease

Navigating home loans has never been simpler.

Our intelligent chatbot is here to guide you through the complex world of home loans. It's designed to understand your needs, answer your questions, and help you find the right home loan for you.

Instant, Personalised Advice	Get real-time, tailored recommendations based on your needs.
First of it's kind AI home loan assistant	Experience the innovation in home loan assistance.
Saves time and effort in home loan research	Skip the hassle and let our chatbot do the work.

got. Ready to kick off with a few key questions?
Could you please share your tentative plan with us?

I want to buy a new home

That's great! It's always exciting to plan for a new home. Are you planning for your first home purchase, if I may ask?

Yes

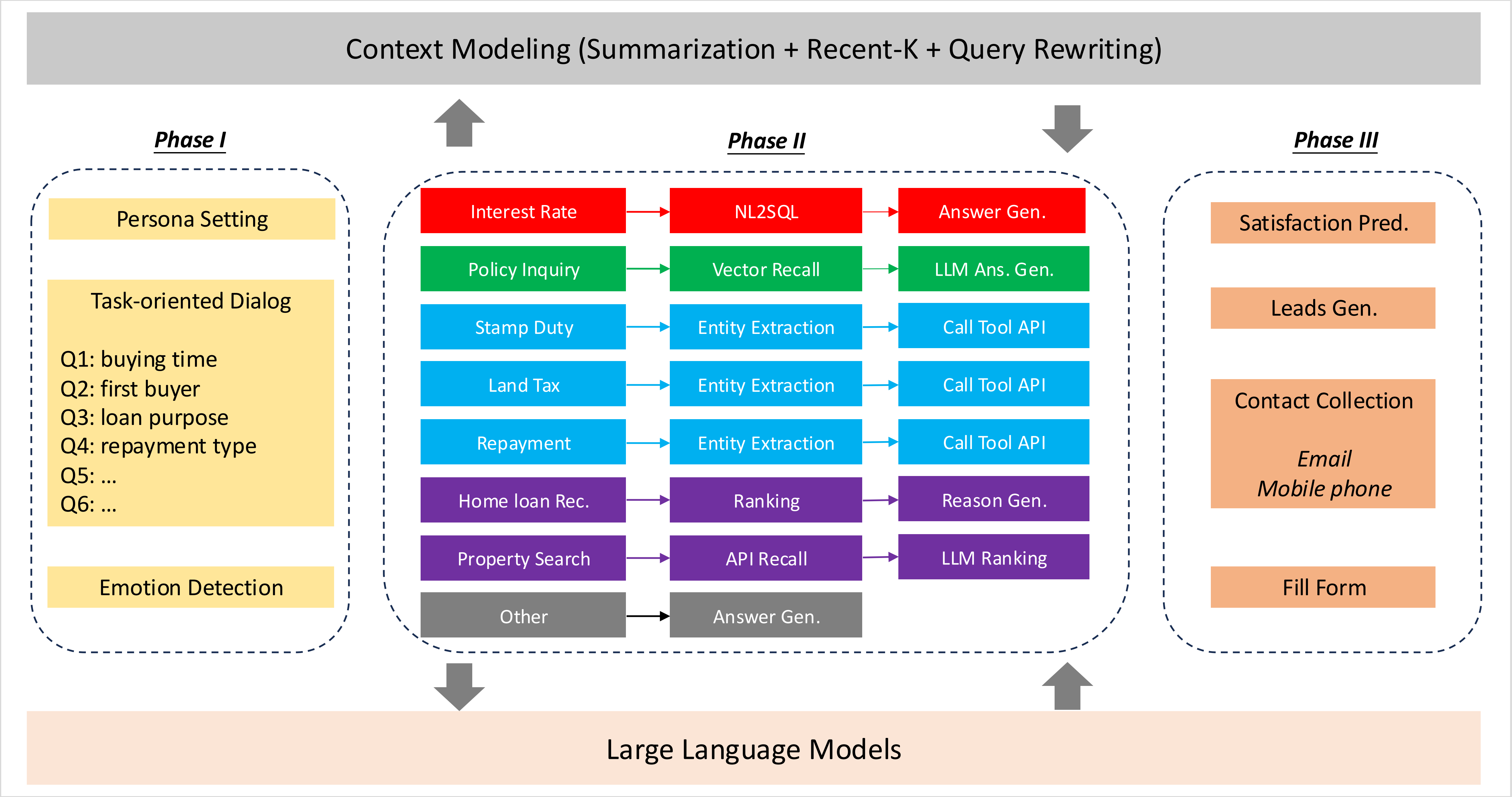
It's fantastic that you're planning to purchase your first home. The next thing I'd like to understand is your loan purpose. Are you seeking to be the owner occupier, or is this an investment?

Owner Occupier

Great, thanks for the clarification. Now, could you please tell me your preferred repayment type? Would it be principal and interest or interest only?

Have a try: https://www.yephome.com.au/chat

Architecture of AI Home Loan



Context Modeling (Summarization + Recent-K + Query Rewriting)

Phase I

Persona Setting

Task-oriented Dialog

Q1: buying time
Q2: first buyer
Q3: loan purpose
Q4: repayment type
Q5: …
Q6: …

Emotion Detection

Phase II

Interest Rate → NL2SQL → Answer Gen.
Policy Inquiry → Vector Recall → LLM Ans. Gen.
Stamp Duty → Entity Extraction → Call Tool API
Land Tax → Entity Extraction → Call Tool API
Repayment → Entity Extraction → Call Tool API
Home loan Rec. → Ranking → Reason Gen.
Property Search → API Recall → LLM Ranking
Other → Answer Gen.

Phase III

Satisfaction Pred.

Leads Gen.

Contact Collection

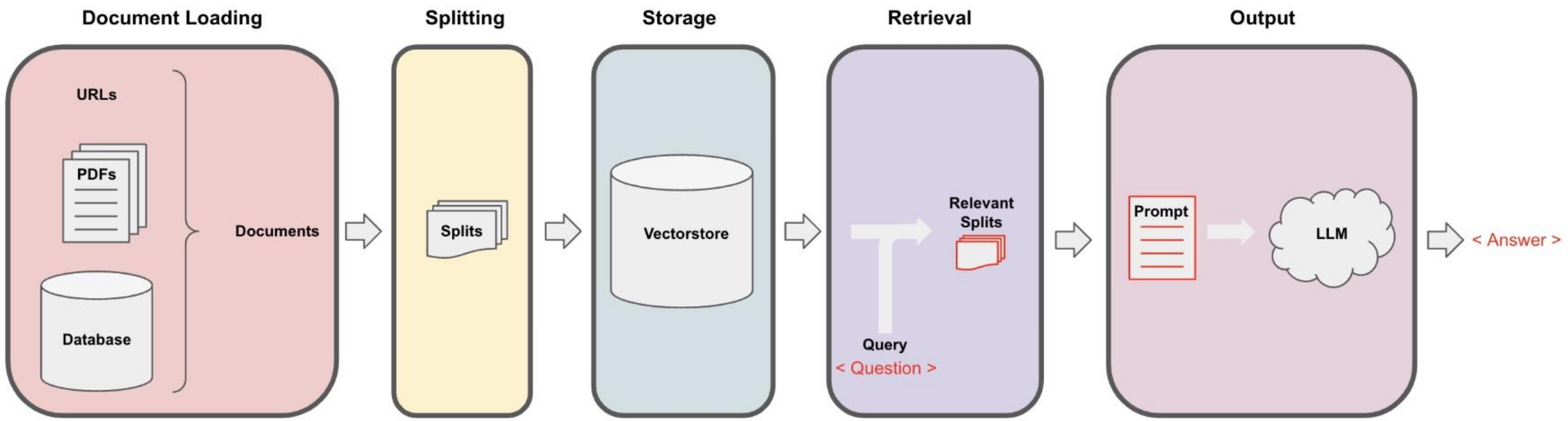
Email
Mobile phone

Fill Form

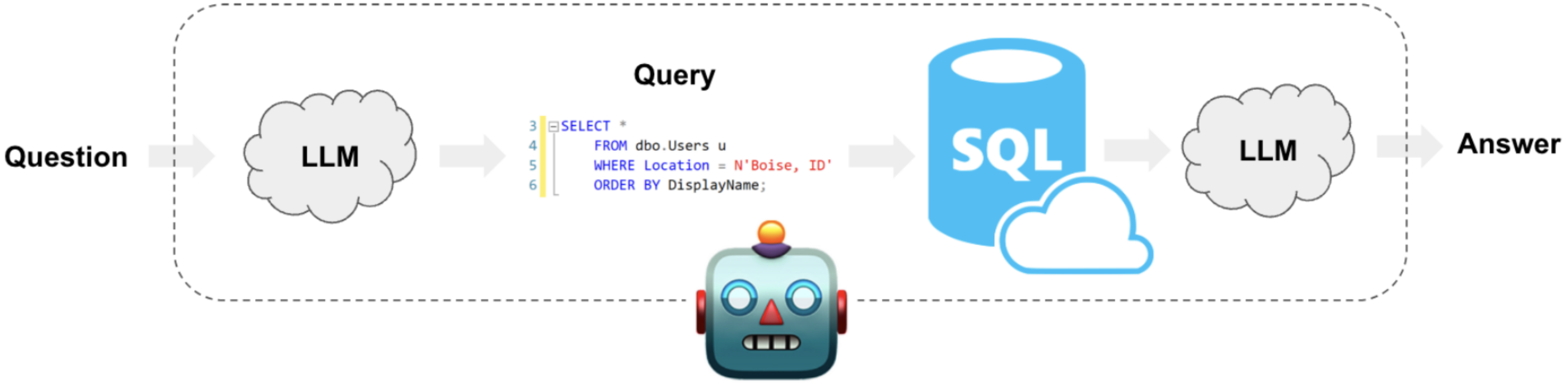
Large Language Models

Core Engine: RAG & NL2SQL

Question & Answering over unstructured document

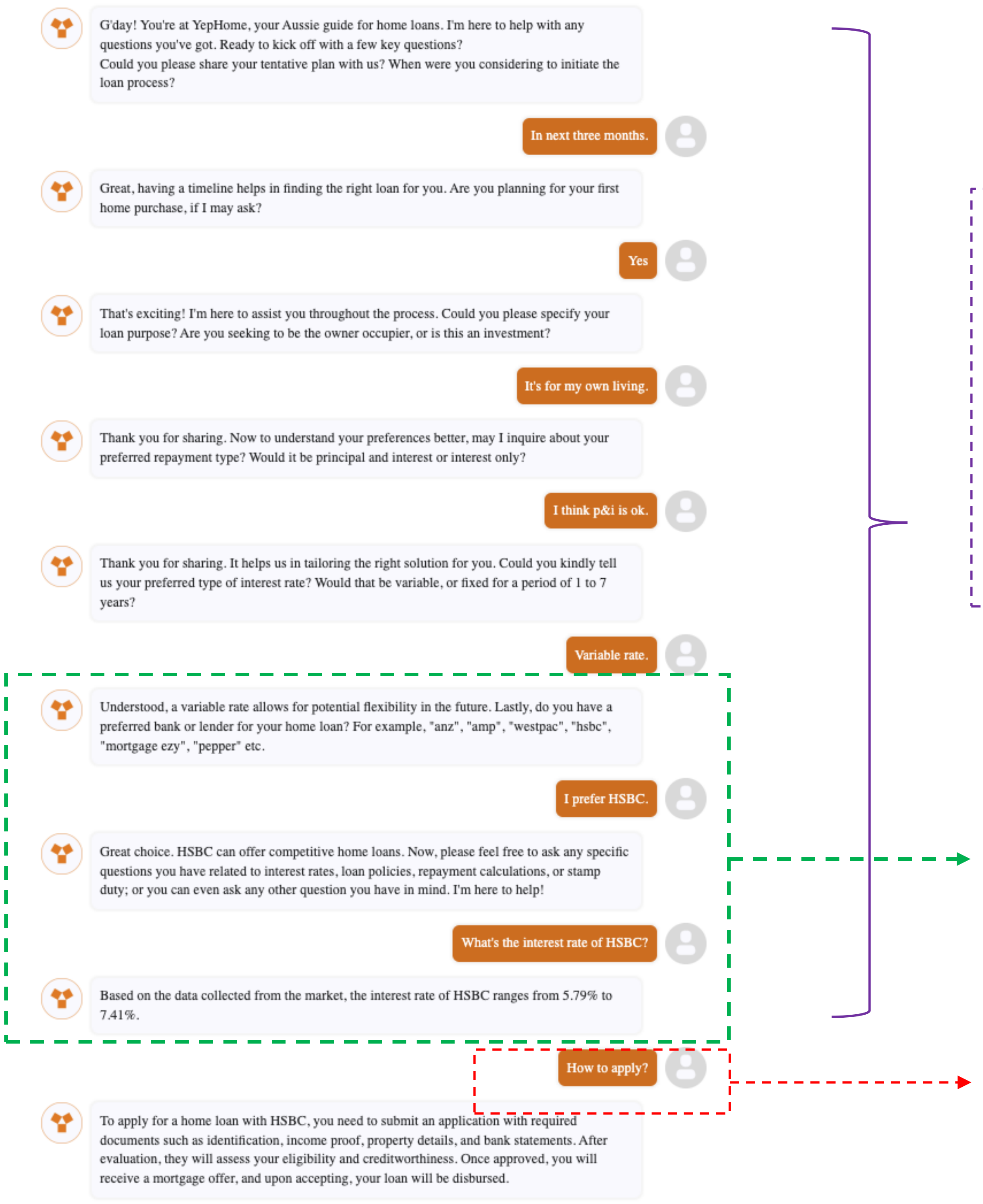


Question & Answering over structured database



```sql
SELECT *
    FROM dbo.Users u
    WHERE Location = N'Boise, ID'
    ORDER BY DisplayName;
```

Context Modeling



G'day! You're at YepHome, your Aussie guide for home loans. I'm here to help with any questions you've got. Ready to kick off with a few key questions?
Could you please share your tentative plan with us? When were you considering to initiate the loan process?

In next three months.

Great, having a timeline helps in finding the right loan for you. Are you planning for your first home purchase, if I may ask?

Yes

That's exciting! I'm here to assist you throughout the process. Could you please specify your loan purpose? Are you seeking to be the owner occupier, or is this an investment?

It's for my own living.

Thank you for sharing. Now to understand your preferences better, may I inquire about your preferred repayment type? Would it be principal and interest or interest only?

I think p&i is ok.

Thank you for sharing. It helps us in tailoring the right solution for you. Could you kindly tell us your preferred type of interest rate? Would that be variable, or fixed for a period of 1 to 7 years?

Variable rate.

Understood, a variable rate allows for potential flexibility in the future. Lastly, do you have a preferred bank or lender for your home loan? For example, "anz", "amp", "westpac", "hsbc", "mortgage ezy", "pepper" etc.

I prefer HSBC.

Great choice. HSBC can offer competitive home loans. Now, please feel free to ask any specific questions you have related to interest rates, loan policies, repayment calculations, or stamp duty; or you can even ask any other question you have in mind. I'm here to help!

What's the interest rate of HSBC?

Based on the data collected from the market, the interest rate of HSBC ranges from 5.79% to 7.41%.

How to apply?

To apply for a home loan with HSBC, you need to submit an application with required documents such as identification, income proof, property details, and bank statements. After evaluation, they will assess your eligibility and creditworthiness. Once approved, you will receive a mortgage offer, and upon accepting, your loan will be disbursed.

Summarization:
The dialogue is between a user and a home loan assistant. The assistant asks the user about their tentative plan to initiate the loan process in the next three months. The user confirms they are planning for their first home purchase. The assistant asks about the loan purpose, and the user states it is for their own living. The assistant then asks about the preferred repayment type, and the user chooses principal and interest. The assistant asks about the preferred interest rate type, and the user chooses a variable rate. Lastly, the user states a preference for HSBC as the bank or lender. The assistant offers to answer any specific questions the user may have and provides a general interest rate range for HSBC.

Recent-K utterances

Query rewriting: What is the procedure for home loan application with HSBC?

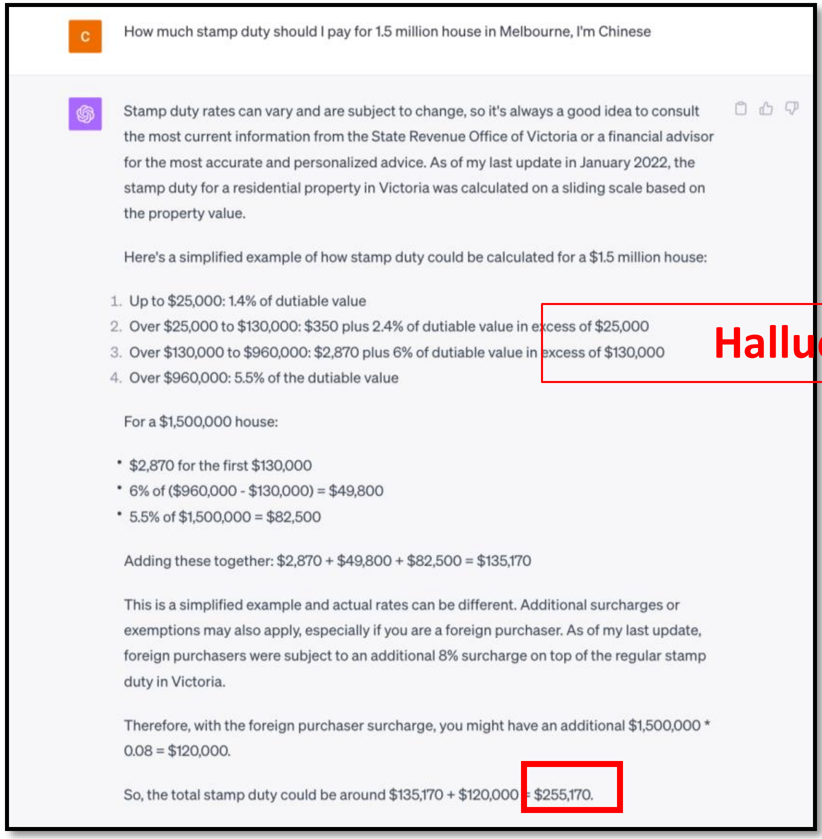
Case Study 1: Calculation

Question: How much stamp duty should I pay for 1.5 million house in Melbourne, I'm Chinese.

GPT-4: $255,170 😭

Anthropic: $122,500 😭

YepGPT: $202,500 😎



How much stamp duty should I pay for 1.5 million house in Melbourne, I'm Chinese

Stamp duty rates can vary and are subject to change, so it's always a good idea to consult the most current information from the State Revenue Office of Victoria or a financial advisor for the most accurate and personalized advice. As of my last update in January 2022, the stamp duty for a residential property in Victoria was calculated on a sliding scale based on the property value.

Here's a simplified example of how stamp duty could be calculated for a $1.5 million house:

1. Up to $25,000: 1.4% of dutiable value
2. Over $25,000 to $130,000: $350 plus 2.4% of dutiable value in excess of $25,000
3. Over $130,000 to $960,000: $2,870 plus 6% of dutiable value in excess of $130,000
4. Over $960,000: 5.5% of the dutiable value

For a $1,500,000 house:

- $2,870 for the first $130,000
- 6% of ($960,000 - $130,000) = $49,800
- 5.5% of $1,500,000 = $82,500

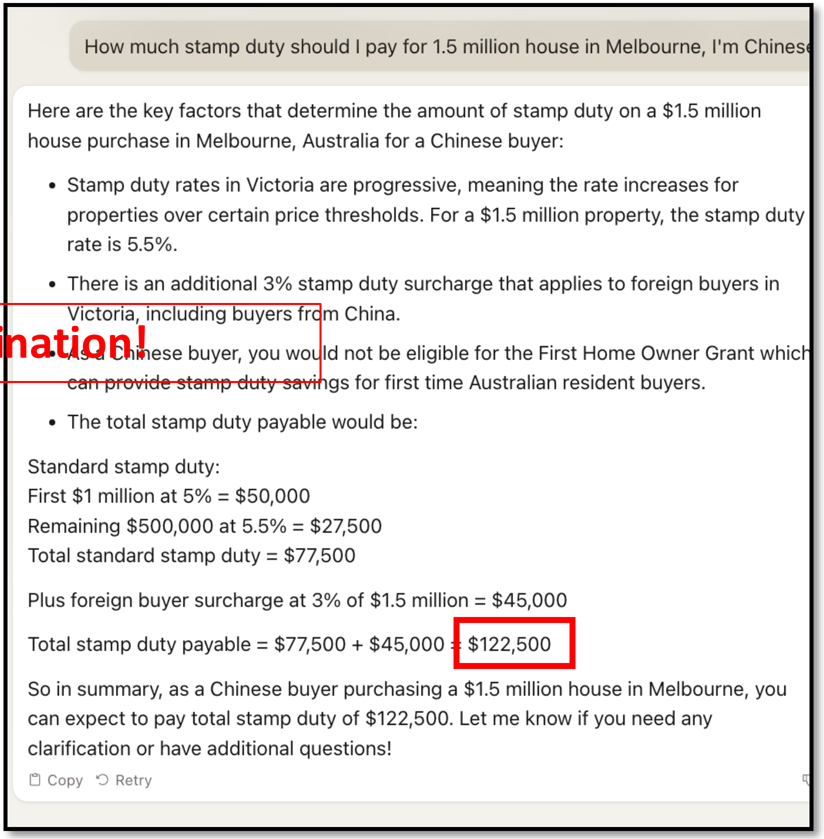
Adding these together: $2,870 + $49,800 + $82,500 = $135,170

This is a simplified example and actual rates can be different. Additional surcharges or exemptions may also apply, especially if you are a foreign purchaser. As of my last update, foreign purchasers were subject to an additional 8% surcharge on top of the regular stamp duty in Victoria.

Therefore, with the foreign purchaser surcharge, you might have an additional $1,500,000 * 0.08 = $120,000.

So, the total stamp duty could be around $135,170 + $120,000 = $255,170.

Hallucination!



How much stamp duty should I pay for 1.5 million house in Melbourne, I'm Chinese

Here are the key factors that determine the amount of stamp duty on a $1.5 million house purchase in Melbourne, Australia for a Chinese buyer:

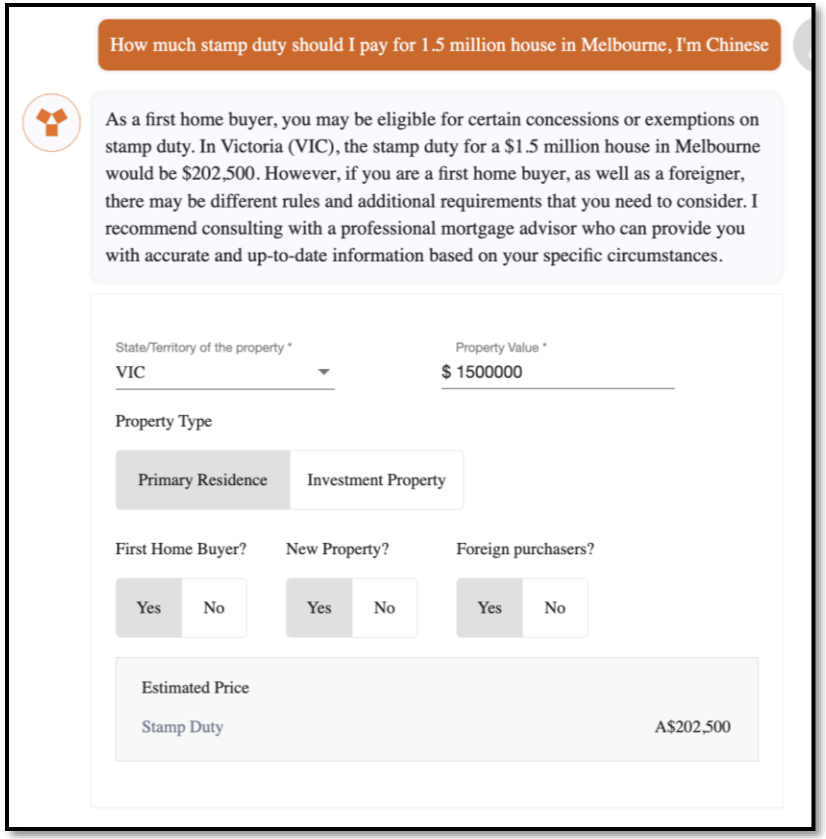
- Stamp duty rates in Victoria are progressive, meaning the rate increases for properties over certain price thresholds. For a $1.5 million property, the stamp duty rate is 5.5%.
- There is an additional 3% stamp duty surcharge that applies to foreign buyers in Victoria, including buyers from China.
- As a Chinese buyer, you would not be eligible for the First Home Owner Grant which can provide stamp duty savings for first time Australian resident buyers.
- The total stamp duty payable would be:

Standard stamp duty:
First $1 million at 5% = $50,000
Remaining $500,000 at 5.5% = $27,500
Total standard stamp duty = $77,500

Plus foreign buyer surcharge at 3% of $1.5 million = $45,000

Total stamp duty payable = $77,500 + $45,000 = $122,500

So in summary, as a Chinese buyer purchasing a $1.5 million house in Melbourne, you can expect to pay total stamp duty of $122,500. Let me know if you need any clarification or have additional questions!



How much stamp duty should I pay for 1.5 million house in Melbourne, I'm Chinese

As a first home buyer, you may be eligible for certain concessions or exemptions on stamp duty. In Victoria (VIC), the stamp duty for a $1.5 million house in Melbourne would be $202,500. However, if you are a first home buyer, as well as a foreigner, there may be different rules and additional requirements that you need to consider. I recommend consulting with a professional mortgage advisor who can provide you with accurate and up-to-date information based on your specific circumstances.

State/Territory of the property *: VIC
Property Value *: $ 1500000

Property Type: [✓ Primary Residence] Investment Property

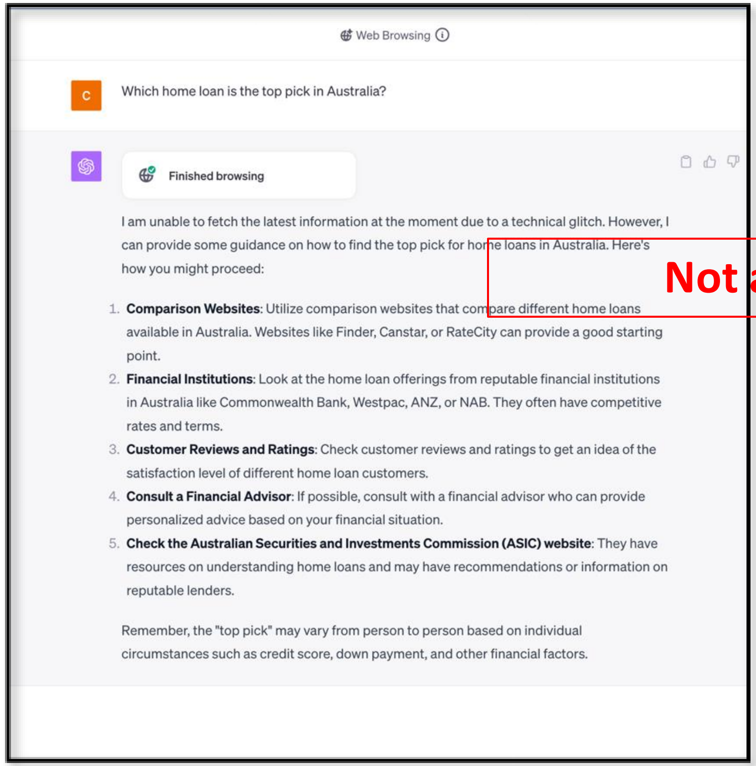
First Home Buyer? [✓ Yes] No
New Property? [✓ Yes] No
Foreign purchasers? [✓ Yes] No

Estimated Price
Stamp Duty: A$202,500

Case Study 2: Recommendation

Question: Which home loan is the top pick in Australia?

GPT-4: No answer 😭



Which home loan is the top pick in Australia?

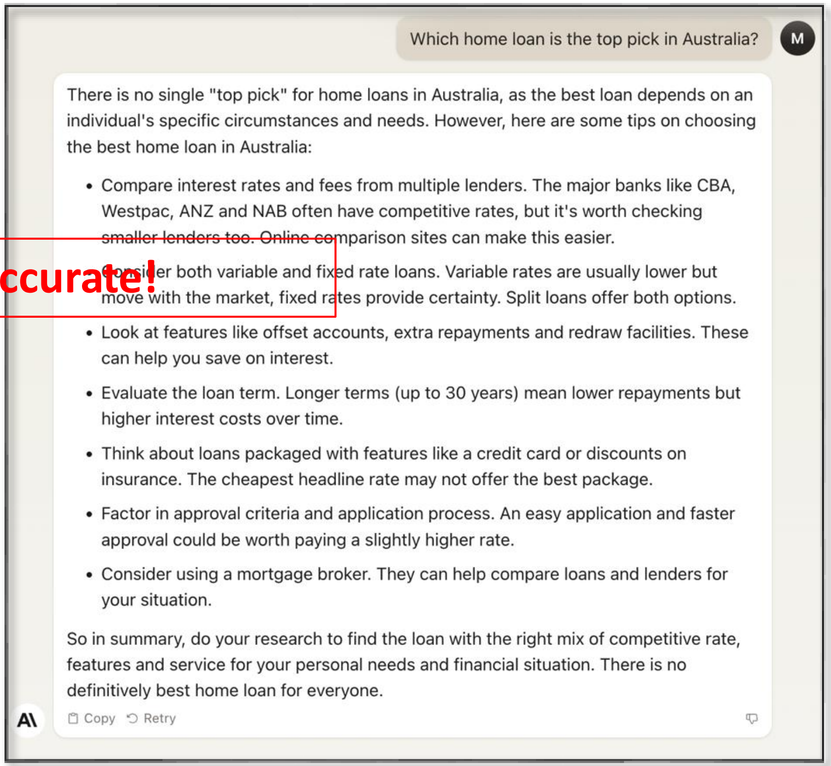
I am unable to fetch the latest information at the moment due to a technical glitch. However, I can provide some guidance on how to find the top pick for home loans in Australia. Here's how you might proceed:

1. **Comparison Websites**: Utilize comparison websites that compare different home loans available in Australia. Websites like Finder, Canstar, or RateCity can provide a good starting point.
2. **Financial Institutions**: Look at the home loan offerings from reputable financial institutions in Australia like Commonwealth Bank, Westpac, ANZ, or NAB. They often have competitive rates and terms.
3. **Customer Reviews and Ratings**: Check customer reviews and ratings to get an idea of the satisfaction level of different home loan customers.
4. **Consult a Financial Advisor**: If possible, consult with a financial advisor who can provide personalized advice based on your financial situation.
5. **Check the Australian Securities and Investments Commission (ASIC) website**: They have resources on understanding home loans and may have recommendations or information on reputable lenders.

Remember, the "top pick" may vary from person to person based on individual circumstances such as credit score, down payment, and other financial factors.

Not accurate!

Anthropic: No answer 😭



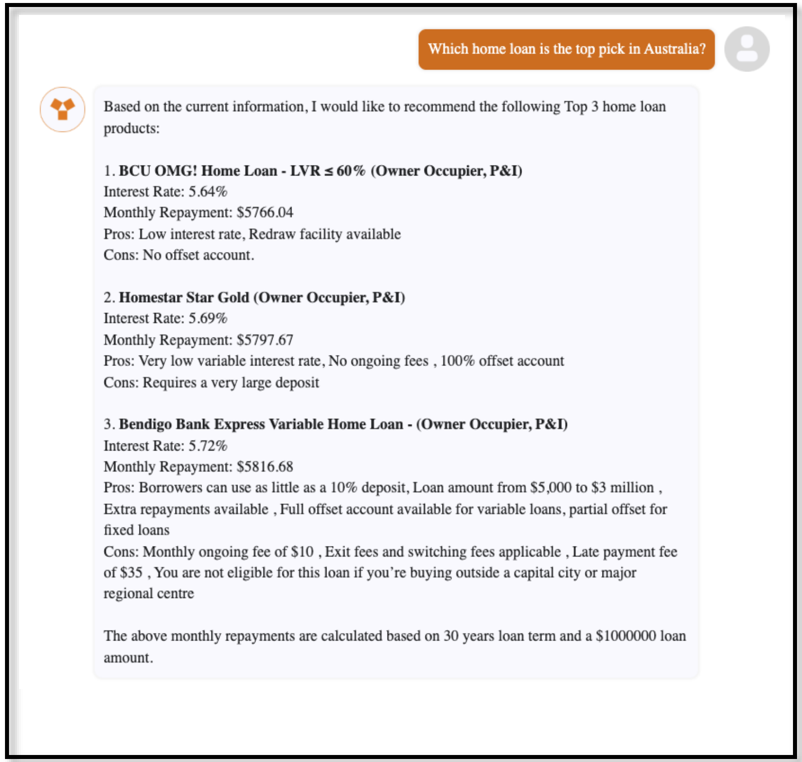
Which home loan is the top pick in Australia?

There is no single "top pick" for home loans in Australia, as the best loan depends on an individual's specific circumstances and needs. However, here are some tips on choosing the best home loan in Australia:

- Compare interest rates and fees from multiple lenders. The major banks like CBA, Westpac, ANZ and NAB often have competitive rates, but it's worth checking smaller lenders too. Online comparison sites can make this easier.
- Consider both variable and fixed rate loans. Variable rates are usually lower but move with the market, fixed rates provide certainty. Split loans offer both options.
- Look at features like offset accounts, extra repayments and redraw facilities. These can help you save on interest.
- Evaluate the loan term. Longer terms (up to 30 years) mean lower repayments but higher interest costs over time.
- Think about loans packaged with features like a credit card or discounts on insurance. The cheapest headline rate may not offer the best package.
- Factor in approval criteria and application process. An easy application and faster approval could be worth paying a slightly higher rate.
- Consider using a mortgage broker. They can help compare loans and lenders for your situation.

So in summary, do your research to find the loan with the right mix of competitive rate, features and service for your personal needs and financial situation. There is no definitively best home loan for everyone.

YepGPT: Done! 😎



Which home loan is the top pick in Australia?

Based on the current information, I would like to recommend the following Top 3 home loan products:

1. **BCU OMG! Home Loan - LVR ≤ 60% (Owner Occupier, P&I)**
Interest Rate: 5.64%
Monthly Repayment: $5766.04
Pros: Low interest rate, Redraw facility available
Cons: No offset account.

2. **Homestar Star Gold (Owner Occupier, P&I)**
Interest Rate: 5.69%
Monthly Repayment: $5797.67
Pros: Very low variable interest rate, No ongoing fees , 100% offset account
Cons: Requires a very large deposit

3. **Bendigo Bank Express Variable Home Loan - (Owner Occupier, P&I)**
Interest Rate: 5.72%
Monthly Repayment: $5816.68
Pros: Borrowers can use as little as a 10% deposit, Loan amount from $5,000 to $3 million , Extra repayments available , Full offset account available for variable loans, partial offset for fixed loans
Cons: Monthly ongoing fee of $10 , Exit fees and switching fees applicable , Late payment fee of $35 , You are not eligible for this loan if you're buying outside a capital city or major regional centre

The above monthly repayments are calculated based on 30 years loan term and a $1000000 loan amount.

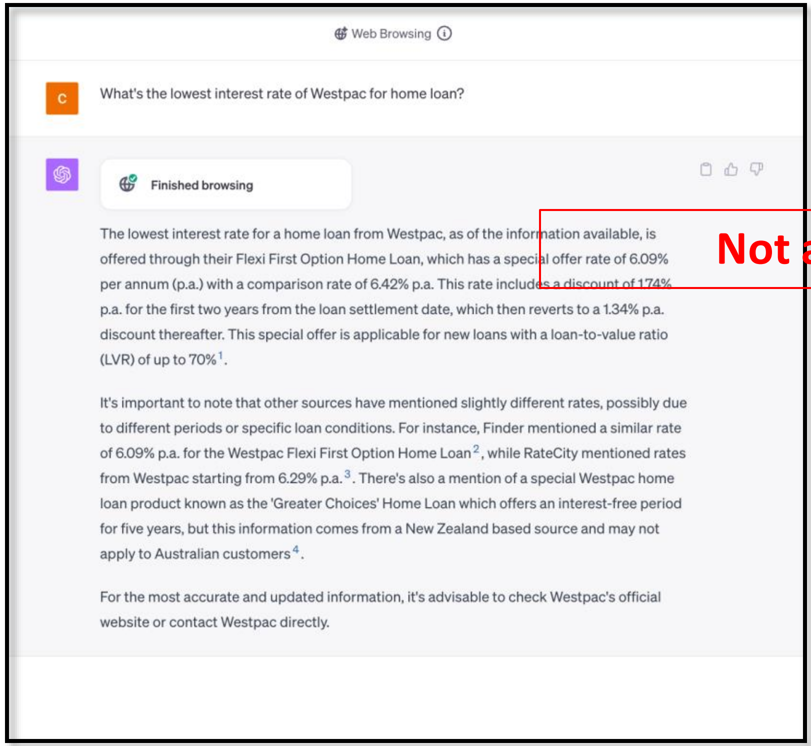
Case Study 3: Info-Seeking

Question: What's the lowest interest rate of Westpac for home loan?

GPT-4: 6.09% 🥺

Anthropic: N/A 😭

YepGPT: $5.79% 😎



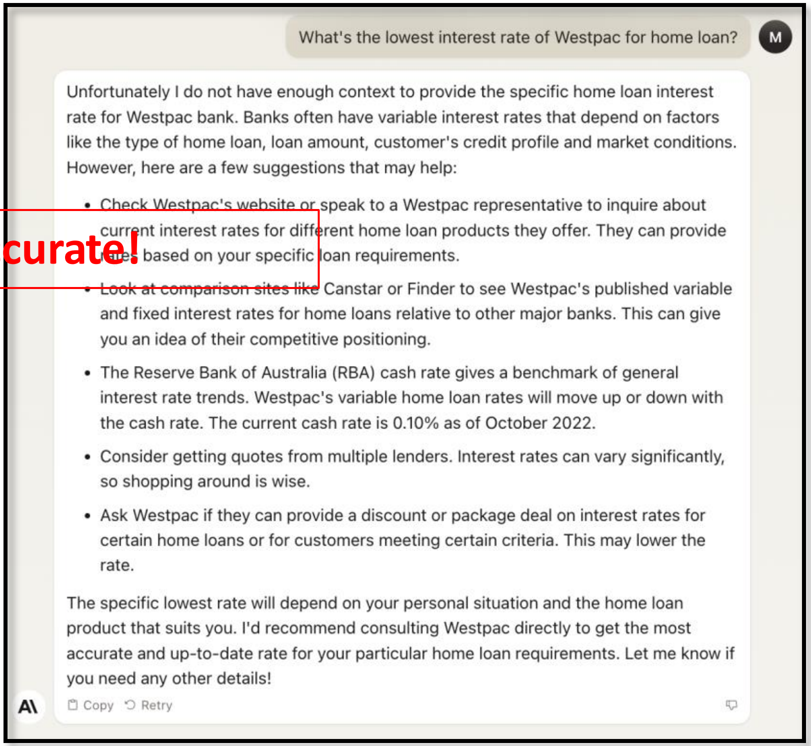
What's the lowest interest rate of Westpac for home loan?

The lowest interest rate for a home loan from Westpac, as of the information available, is offered through their Flexi First Option Home Loan, which has a special offer rate of 6.09% per annum (p.a.) with a comparison rate of 6.42% p.a. This rate includes a discount of 1.74% p.a. for the first two years from the loan settlement date, which then reverts to a 1.34% p.a. discount thereafter. This special offer is applicable for new loans with a loan-to-value ratio (LVR) of up to 70%[1].

It's important to note that other sources have mentioned slightly different rates, possibly due to different periods or specific loan conditions. For instance, Finder mentioned a similar rate of 6.09% p.a. for the Westpac Flexi First Option Home Loan[2], while RateCity mentioned rates from Westpac starting from 6.29% p.a.[3]. There's also a mention of a special Westpac home loan product known as the 'Greater Choices' Home Loan which offers an interest-free period for five years, but this information comes from a New Zealand based source and may not apply to Australian customers[4].

For the most accurate and updated information, it's advisable to check Westpac's official website or contact Westpac directly.

Not accurate!

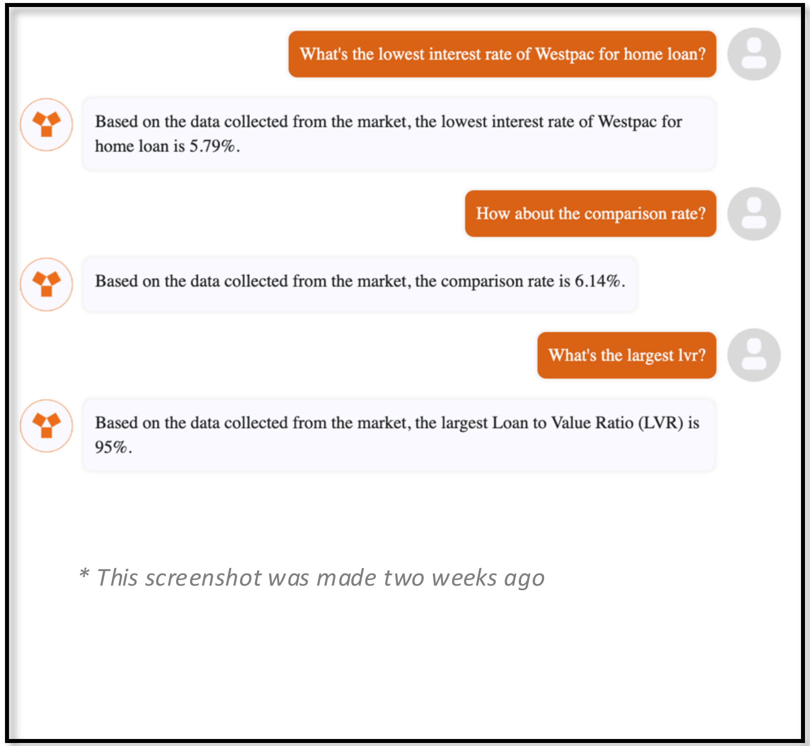


What's the lowest interest rate of Westpac for home loan?

Unfortunately I do not have enough context to provide the specific home loan interest rate for Westpac bank. Banks often have variable interest rates that depend on factors like the type of home loan, loan amount, customer's credit profile and market conditions. However, here are a few suggestions that may help:

- Check Westpac's website or speak to a Westpac representative to inquire about current interest rates for different home loan products they offer. They can provide rates based on your specific loan requirements.
- Look at comparison sites like Canstar or Finder to see Westpac's published variable and fixed interest rates for home loans relative to other major banks. This can give you an idea of their competitive positioning.
- The Reserve Bank of Australia (RBA) cash rate gives a benchmark of general interest rate trends. Westpac's variable home loan rates will move up or down with the cash rate. The current cash rate is 0.10% as of October 2022.
- Consider getting quotes from multiple lenders. Interest rates can vary significantly, so shopping around is wise.
- Ask Westpac if they can provide a discount or package deal on interest rates for certain home loans or for customers meeting certain criteria. This may lower the rate.

The specific lowest rate will depend on your personal situation and the home loan product that suits you. I'd recommend consulting Westpac directly to get the most accurate and up-to-date rate for your particular home loan requirements. Let me know if you need any other details!



What's the lowest interest rate of Westpac for home loan?

Based on the data collected from the market, the lowest interest rate of Westpac for home loan is 5.79%.

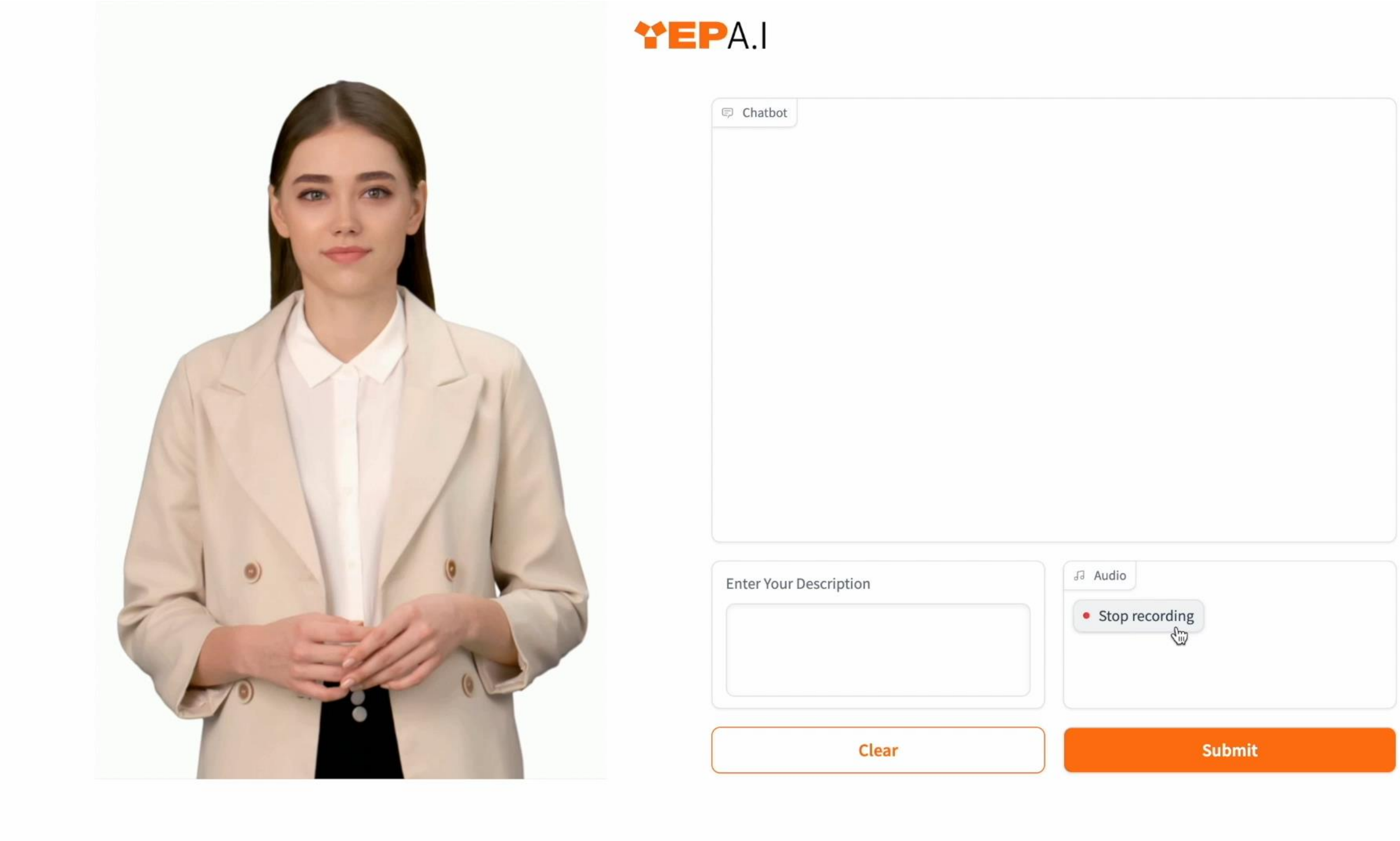
How about the comparison rate?

Based on the data collected from the market, the comparison rate is 6.14%.

What's the largest lvr?

Based on the data collected from the market, the largest Loan to Value Ratio (LVR) is 95%.

** This screenshot was made two weeks ago*


YEPA.I
Chatbot
Enter Your Description
Audio
Stop recording
Clear
Submit

Take-aways

- The core components of language understanding can be effectively reconstructed based on Large Language Models (LLMs), which significantly simplifies the design process of dialogue systems.
- Efficiently incorporating external domain knowledge into LLMs can mitigate the hallucination problem to a certain degree.
- A multimodal dialogue system, enhanced by an AI avatar, speech interaction, and LLMs, holds immense commercial potential for the future.

Q&A

Thanks!

Email: chenmengdx@gmail.com
Home page: https://chenmengdx.github.io/
Visit: https://www.yepai.com.au/