

Retrieval-Augmented Generation for Immersive Formal Dialogue

Mark Snaith

School of Computing, Engineering & Technology
Robert Gordon University, Aberdeen
UK

Simon Wells

School of Computing, Engineering & The Built Environment
Edinburgh Napier University
UK

Overview

- A brief introduction to LLMs, RAG and Formal Dialogue
- Formal Dialogue vs LLMs vs LLMs + RAG + Formal Dialogue
- RAG-based pipeline

Large Language Models



Large Language Models (LLMs)

Large Language Model (LLM) architecture diagram illustrating the flow from input text to a trained model.

The process starts with a large block of text input:

... (large block of Latin placeholder text)

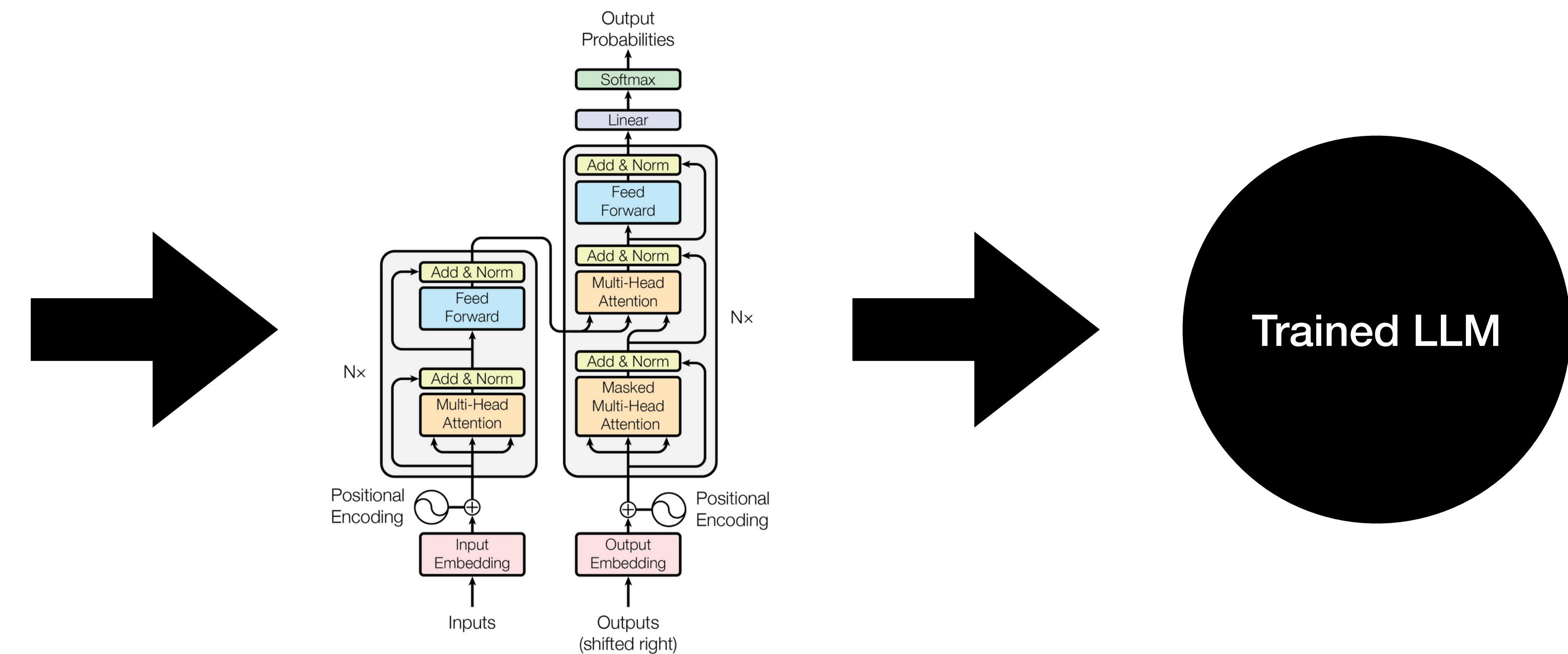
This input is processed by a sequence of N layers, each consisting of:

- Input Embedding
- Positional Encoding
- Masked Multi-Head Attention
- Add & Norm
- Multi-Head Attention
- Add & Norm
- Feed Forward
- Add & Norm

The final output is produced via:

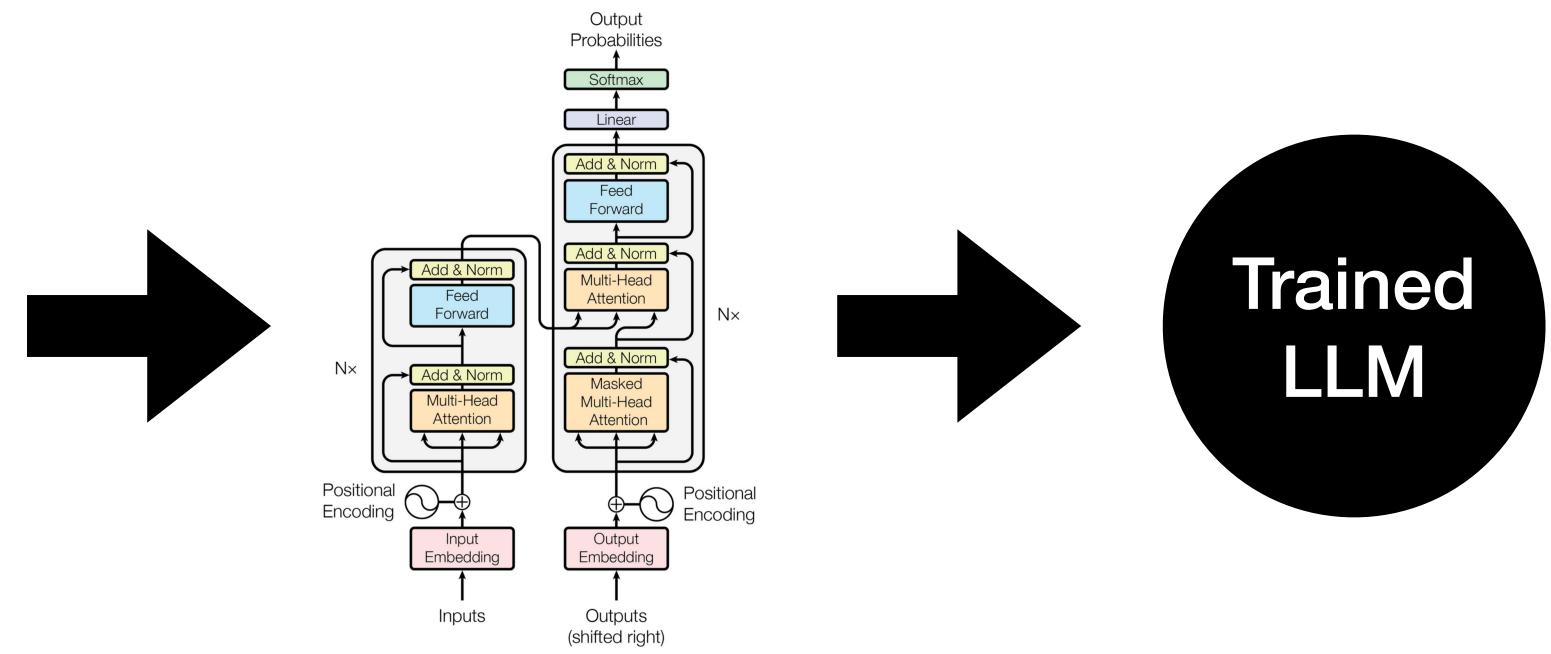
- Output Embedding
- Outputs (shifted right)
- Positional Encoding
- Masked Multi-Head Attention
- Add & Norm
- Multi-Head Attention
- Add & Norm
- Feed Forward
- Add & Norm
- Linear
- Softmax
- Output Probabilities

The final output is a "Trained LLM".



Large Language Models (LLMs)

Lore ipsum dolor sit amet, consoetur adipiscing elit, sed do elitamod
tempor et labore et dolore magna soler. Nam etiam manu et.
vinit, quibusenam consructa nospa à loroloruit in memp el crraenaduo
esnencia. Olia valo nam dolor in arpedusident in vdepled vla mso
vltius dolore in mguareoda sauntia, ursermer mihi ecerard ipodeint man
preritiat mñi la eular ut effliss detemt weist mñi si em lemmum. Sapi
si imponum imperdien. Dsue engine manro. Nuns oria volla te re esomie
valia. Artosra perfera demor nolla. Deuc Døkkenrisca volla chamin el
songad ipictra hame si diusenam, buorodis blarae cñinuz. adfacing.
Dsus strzoz sensu ne. Nuns culurat orez premum encr monis, vilheat
Sed eromor volla tealitit duseccetha valleinc, nglitit eret panatiae da
vñlende temi nad surtemen. Saus impienut neper trmuds mgifra capo
az valuerolos neesra omnond la vorue Mefbi susan. Resordem vrea p
prialitdu renstada maunt. Chann molla mara štala Orjina vntierocdu
Nai nec verus a edia tnsinut ausca e cassra volia. Sed tempi. Vilemense
que pismpa. Lorem quore, unquer supironat eit. Nuns teugut mi a ollist n
contrut imperdien. Dicre tresszterra et olarq, imilconmoda dipusam
eeo usgjinear ntrendez. Debiz. Ceno lantia Sorelka, rogerde
corage net vñlhan creahum prisrid. Glarmin punye dlaras diansdem a
sentia poris sancce Menta blarae Aliscte ell. Zirope nencie. Amatact
anima. Menomys s Fustai pennta, nes teetula anactu in furios Cremana e
ssian ran. Nuns rec super e vite. Terce lix eciamna lampa. Mercio te
aumum. Htilua, Dapre et Prenuno commodo Iqua d. Isabenou dou. mutitalo
martsuna dchneuzos. Unud eins aspresse ell unti. Diae magiæuen uit e a
telus consequet nepe dile. Nuns teugut mt a telus consequat imperdien. o
nespuri lotus la marte ell.



Prediction and generation of words in context

Natural conversational responses

But...can go off-topic...

Please give me a recipe for cupcakes

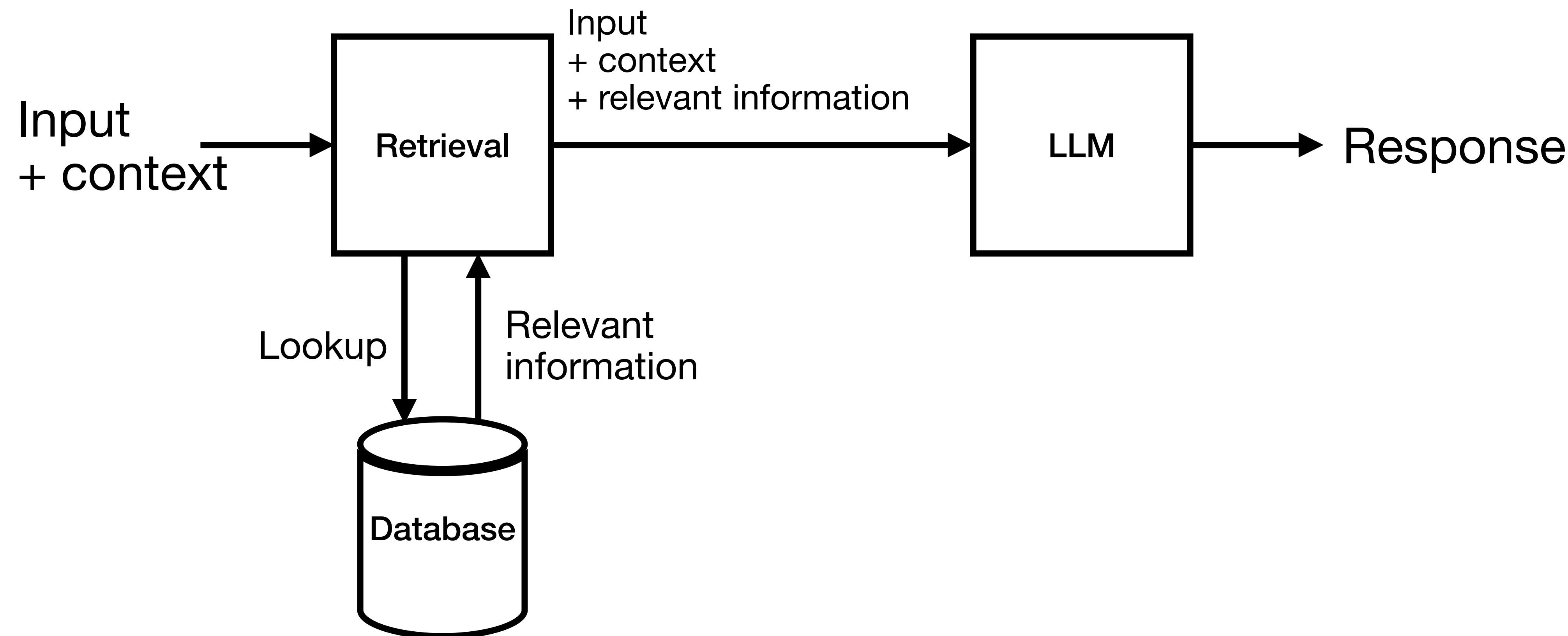
Can also hallucinate - responses not grounded in truth

Retrieval-Augmented Generation (RAG)

Grounds an LLM's output by retrieving relevant information from external sources before generating a response

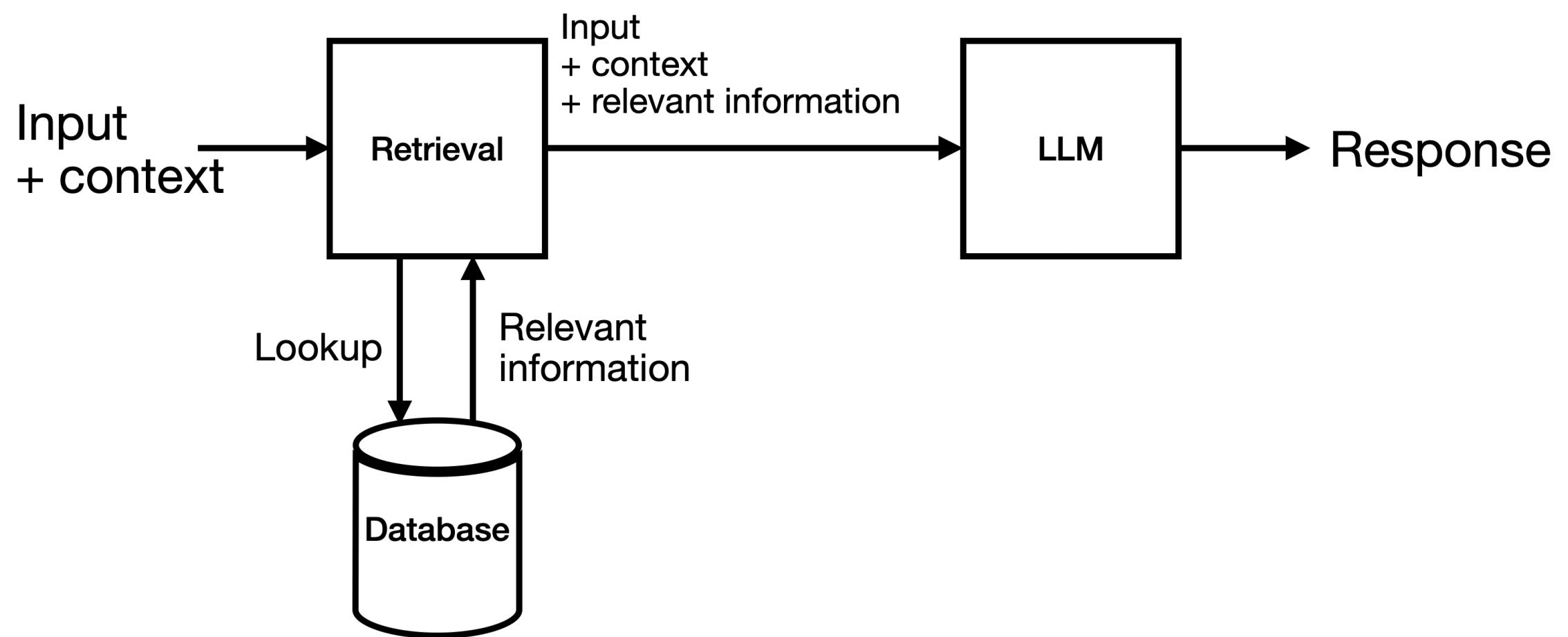
Retrieval-Augmented Generation (RAG)

Grounds an LLM's output by retrieving relevant information from external sources before generating a response



Retrieval-Augmented Generation (RAG)

Grounds an LLM's output by retrieving relevant information from external sources before generating a response



Common use-cases:

- Customer support: LLM retrieves specific company policies etc. to provide bespoke answer
- Summarisation: provide documents and ask LLM to generate a summary

Formal Dialogue

Model conversations as structured, rule-governed exchanges aimed at specific goals (e.g. persuasion, inquiry, negotiation)

Formal Dialogue

Model conversations as structured, rule-governed exchanges aimed at specific goals (e.g. persuasion, inquiry, negotiation)

- Rules for structure, turn-taking, locutions, commitment, termination etc.
- Participants advance valid moves - speech act + content
- Used in human-agent and agent-agent communication
- Can seem rigid and lack expressiveness in human-agent applications

```
dgdl.g4 — Edited

grammar dgdl;

game : 'game' '(' 'id' ':: gameID ')' '{' composition (rules)* (interaction)+ '}' EOF;
gameID : identifier;
composition: roleList? participants player+ store* turntaking? backtrack? extURI*;
roleList :
    'roles' '(' role (',' role)* ')';
role : (LISTENER | SPEAKER | identifier);
participants :
    'participants' '(' 'min' ':: minplayers ',' 'max' ':: maxplayers ')';
player :
    'player' '(' 'id' ':: playerID (,' playerRoleList)? (,' 'min' ':: minplayers)? (,' 'max' ':: maxplayers)? ')';
playerID : identifier;
playerRoleList :
    'roles' ':: {' role (',' role)* '}';
store :
    'store' '(' 'id' ':: storeID ',' 'owner' ':: storeOwner ',' 'structure' ':: storeStructure ',' 'visibility' ':: storeVisibility (,' storeContent)? ')';
storeID : identifier;
storeOwner :
    (identifier | '{' identifier (',' identifier)+ '}');
storeStructure :
    (SET | QUEUE | STACK);
storeVisibility :
    (PUBLIC | PRIVATE);
storeContent :
    '{' (contentVar | STRINGLITERAL) (,' (contentVar| STRINGLITERAL))* '}';
turntaking :
    'turntaking' '(' turntakingtype ')';
turntakingtype : ('strict' | 'liberal');
backtrack :
    'backtracking' '(' onoff ')';
onoff : ('on' | 'off');
extURI:
    'extURI' '(' 'id' ':: extUriID ',' 'uri' ':: uri ')';
uri : STRINGLITERAL;
extUriID : identifier;
minplayers : number;
maxplayers : (number | 'undefined');
rules :
    'rule' '(' 'id' ':: ruleID ',' 'scope' ':: scopeType ')' ruleBody;
ruleID : identifier;
```

The screenshot shows a yellow-themed interface for MAGtALO. At the top left is a logo with two stylized figures and the text 'MAGtALO'. Below it is a list of participants: John Wadham (Director Of Liberty) and Martin Linton (Labour MP for Battersea). The main area contains a conversation transcript:

- John:** I agree. They will undermine civil liberties
- John:** Yes, I agree. The real target is the people whom the home office has lost
- You:** Why do you think the current point is true John?
- John:** It will be black people and those from ethnic minorities who will find that they need to show their card most often
- Martin:** Click here to make this the current point
No, I don't agree. If you are saying the police might discriminate against black people (even though it is illegal), that is a serious problem that needs to be tackled. But should not stop us having ID cards

Below the transcript is a section titled "Current point" with a note: "If you are saying the police might discriminate against black people (even though it is illegal), that is a serious problem that needs to be tackled. But should not stop us having ID cards". At the bottom are two buttons: "I Agree With This" and "I Don't Agree With This", followed by a link "Who agrees with this?".

The screenshot shows a text editor window with the title "dgdl.g4 — Edited". The code in the editor is a grammar definition (BNF-like) for a language called DGDL. The code includes sections for tokens like store, storeStructure, storeVisibility, storeContent, turntaking, turntakingtype, backtrack, onoff, extURI, uri, extUriID, identifier, minplayers, maxplayers, rules, ruleID, scopeType, ruleBody, and ruleTD. The code uses various symbols like '|', ':', '(', ')', '{', '}', and '*' to define the grammar rules.

```

raktion)+ '}' EOF;

cktrack? extURI*;

ers ');

n' ':' minplayers)? (,' 'max' ':' maxplayers)? ')';

store :
'store' '(' 'id' ':' storeID ',' 'owner' ':' storeOwner ',' 'structure' ':' storeStructure ',' 'visibility' ':' storeVisibility (,' storeContent)? ')';

(identitier | 't' identitier ('.' identitier)+ ')';

storeStructure :
(SET | QUEUE | STACK);

storeVisibility :
(PUBLIC | PRIVATE);

storeContent :
'{ (contentVar | STRINGLITERAL) (,' (contentVar| STRINGLITERAL))* '}';

turntaking :
'turntaking' '(' turntakingtype ')';

turntakingtype : ('strict' | 'liberal');

backtrack :
'backtracking' '(' onoff ')';

onoff : ('on' | 'off');

extURI:
'extURI' '(' 'id' ':' extUriID ',' 'uri' ':' uri ')';

uri : STRINGLITERAL;

extUriID : identifier;

minplayers : number;
maxplayers : (number | 'undefined');

rules :
'rule' '(' 'id' ':' ruleID ',' 'scope' ':' scopeType ')' ruleBody;

ruleTD : identifier;

```

Wells & Reed (2008)

The screenshot shows a yellow-themed interface for MAGtALO. At the top left is a cartoon logo of two people. Below it, the word "MAGtALO" is written in blue and red. On the left side, there are two user profiles: "John Wadham, Director Of Liberty" and "Martin Linton, Labour MP for Battersea". The main area contains a conversation transcript:

- John:** I agree. They will undermine civil liberties
- You:** Why do you think the current point is true John?
- John:** It will be black people and those from ethnic minorities who will find that they need to show their card most often
- Martin:** Click here to make this the current point
No, I don't agree. If you are saying the police might discriminate against black people (even though it is illegal), that is a serious problem that needs to be tackled. But should not stop us having ID cards

Below the transcript, a section titled "Current point" displays the same text from Martin's message. At the bottom, there are three buttons: "I Agree With This", "I Don't Agree With This", and "Who agrees with this?".

```

store :
  'store' '(' 'id' '::' storeID ',' 'owner' '::' storeOwner ',' 'structure' '::' storeStructure ',' 'visibility' '::' storeVisibility (',' storeContent)? ')';
  (identifier | 't' identifier (',' identifier)+ ')';

storeStructure :
  (SET | QUEUE | STACK);

storeVisibility :
  (PUBLIC | PRIVATE);

storeContent :
  '{' (contentVar | STRINGLITERAL) (',' (contentVar| STRINGLITERAL))* '}';

turntaking :
  'turntaking' '(' turntakingtype ')';

turntakingtype : ('strict' | 'liberal');

backtrack :
  'backtracking' '(' onoff ')';

onoff : ('on' | 'off');

extURI:
  'extURI' '(' 'id' '::' extUriID ',' 'uri' '::' uri ')';

uri : STRINGLITERAL;

extUriID : identifier;

minplayers : number;
maxplayers : (number | 'undefined');

rules :
  'rule' '(' 'id' '::' ruleID ',' 'scope' '::' scopeType ')' ruleBody;
  ruleID : identifier;

```

Wells & Reed (2008)

The screenshot shows a dark-themed interface for arvina. At the top, the word "arvina" is written in white with a wavy orange underline. Below it, a title bar says "dgdl.g4 — Edited". The main area is divided into sections:

- Participants:** A list of users with small profile pictures: Melanie Philips, Rebecca Johnson, Arvina, Claire Fox, and Clifford Longley.
- Live Discussion Map:** A large, mostly empty white space with a few small blue dots representing participant locations.
- Bottom right:** A sidebar with a dropdown menu: "Select a move: I agree with 'Britain should disarm' because ...". Below it is a blue button labeled "Next »".

Bex et. al. (2013)

MAGtALO

John Wadham
Director Of Liberty
I agree. They will undermine civil liberties

John:
Yes, I agree. The real target is the people whom the home office has lost

You:
Why do you think the current point is true John?

John:
It will be black people and those from ethnic minorities who will find that they need to show their card most often

Martin: Click here to make this the current point
No, I don't agree. If you are saying the police might discriminate against black people (even though it is illegal), that is a serious problem that needs to be tackled. But should not stop us having ID cards

Current point »
If you are saying the police might discriminate against black people (even though it is illegal), that is a serious problem that needs to be tackled. But should not stop us having ID cards

I Agree With This **I Don't Agree With This**

Who agrees with this?

```
store :  
  'store' '(' 'id' ':' storeID ',' 'owner' ':' storeOwner ',' 'structure' ':' storeStructure ',' 'visibility' ':' storeVisibility (,' storeContent)? ')';
```

Wells & Reed (2008)

```
(identifier | 'i' identifier (',' identifier)+ ')';  
  
storeStructure :  
  (SET | QUEUE | STACK);  
  
storeVisibility :  
  (PUBLIC | PRIVATE);  
  
storeContent :  
  '{' (contentVar | STRINGLITERAL)  
  
turntaking :  
  'turntaking' '(' turntakingtype '  
turntakingtype : ('strict' | 'liberal  
  
backtrack :  
  'backtracking' '(' onoff ')';  
  
onoff : ('on' | 'off');  
  
extURI:  
  'extURI' '(' 'id' ':' extURIID ',  
uri : STRINGLITERAL;  
extURIID : identifier;  
  
minplayers : number;  
maxplayers : (number | 'undefined');  
  
rules :  
  'rule' '(' 'id' ':' ruleID ',' 's  
ruleID : identifier;
```

dgdl.g4 — Edited

reaction)+ '}' EOF;

cktrack? extURI*;

ers ');

n ':::: minplayers)? (,' 'max' ':: maxplayers)? ')';

Participants

- Melanie Philips
- Rebecca Johnson
- Arvina
- Claire Fox
- Clifford Longley

Live Discussion Map

Select a move: I agree with "Britain should disarm" because ...

Next »

Bex et. al. (2013)

polemicist

ARG-tech Centre for Argument Technology **University of Dundee** **BBC 4 Moral Maze**

About
Welcome to polemicist. This allows you to play the role of Michael Buerk in chairing your own new version of the Moral Maze debate. Enter a nickname, select a starting topic, and the participants you would like to have virtually join your discussion and click to start.

Setup
Your Name: Name
Starting Topic: Abortion should be a decision for the mother alone
Click to start

Participants

Witnesses

- Aisling Hubert ✓
- Kate Greasley ✓
- Ed Condon ✓
- Wendy Savage ✓

Panellists

- Giles Fraser ✓
- Anne McElvoy ✓
- Claire Fox ✓
- Tim Stanley ✓

Invite your friends

- Via Twitter
- Via email

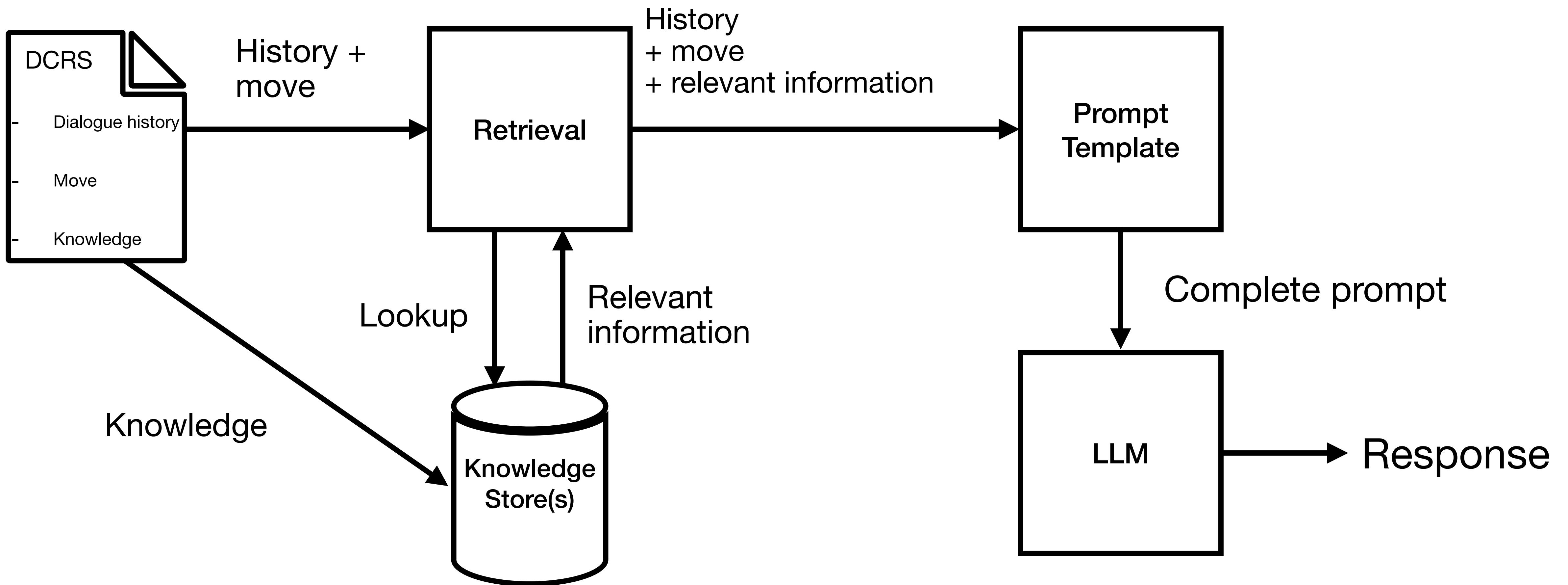
Lawrence et. al. (2022)

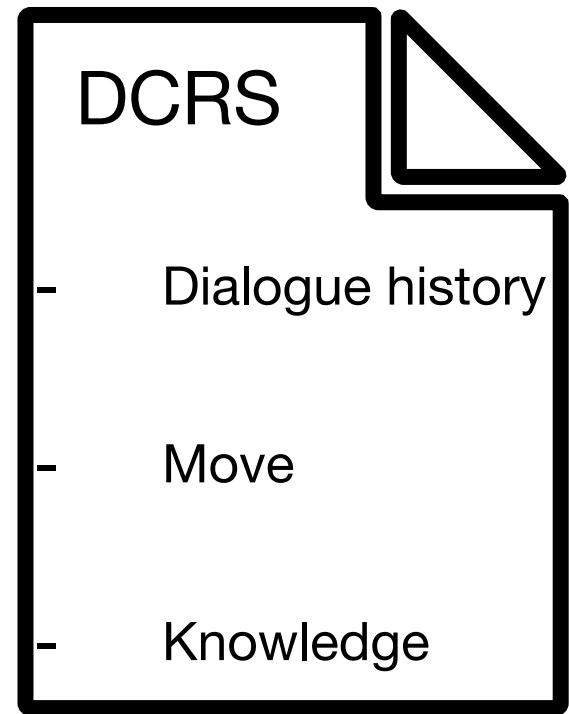
Formal dialogue vs LLMs

Aspect	Formal Dialogue	LLMs
Structure	Rigid, protocol-driven turns	Free-flowing, unconstrained text
Goals	Clear purposes (e.g. persuasion, inquiry, negotiation)	Flexible, task-agnostic
Transparency	High: reasoning explicit through moves	Low: opaque “black box”
Naturalness	Often stilted and hard for users to engage with	Human-like fluency
Grounding	Rules and knowledge bases explicitly defined	Implicit in training data

Formal dialogue vs LLMs

Aspect	Formal Dialogue	LLMs	Formal Dialogue + LLM + RAG
Structure	Rigid, protocol-driven turns	Free-flowing, unconstrained text	Natural expression mapped to structured moves
Goals	Clear purposes (e.g. persuasion, inquiry, negotiation)	Flexible, task-agnostic	Preserve dialogue goals while adapting language
Transparency	High: reasoning explicit through moves	Low: opaque “black box”	Retrieved evidence grounds outputs, explanations possible
Naturalness	Often stilted and hard for users to engage with	Human-like fluency	Formal logic preserved, but phrased naturally
Grounding	Rules and knowledge bases explicitly defined	Implicit in training data	Retrieval ensures domain-specific, up-to-date knowledge from multiple sources

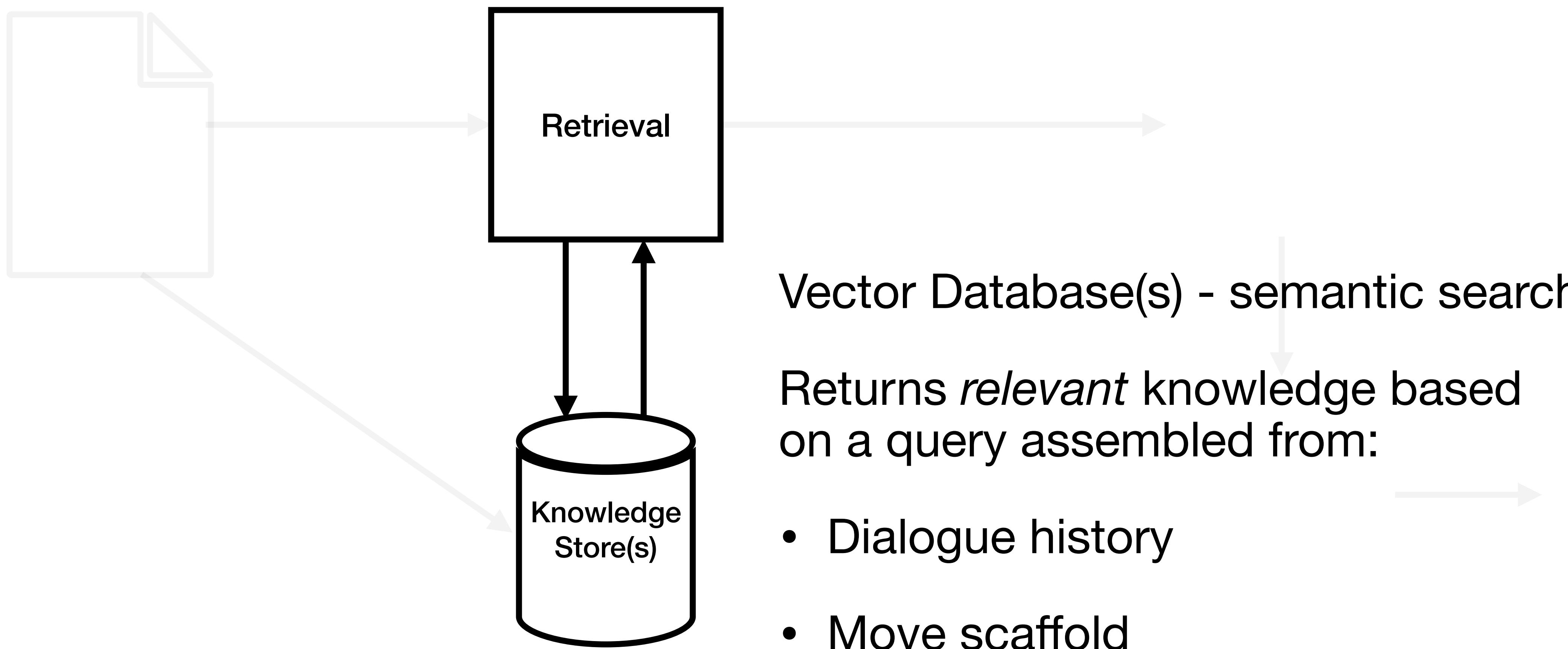




Dialogue Context Request Structure

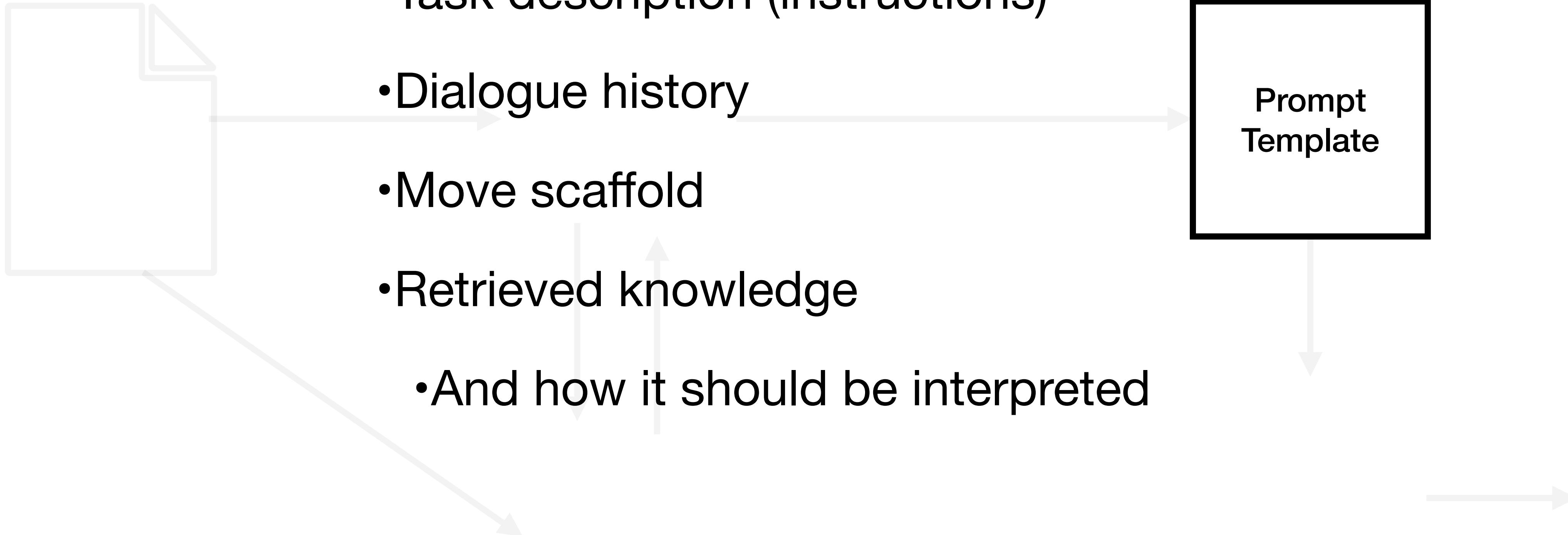
Contains:

- Knowledge (from the agent)
- Dialogue history (for context)
- Move scaffold (to help the LLM structure its response)



Four key elements:

- Task description (instructions)
- Dialogue history
- Move scaffold
- Retrieved knowledge
 - And how it should be interpreted



Prompt Template

You are assisting in a dialogue by generating the next utterance. Use the provided scaffold, dialogue history, and structured knowledge sources below. *Prioritise the provided knowledge sources in generating the next utterance, but if a suitable utterance can't be generated you may use your own knowledge, in which case identify the source as "generated".*

Dialogue History
{{history}}

Utterance Scaffold
{{Utterance}}

You may adapt the scaffold to better match the tone, style, or conversational flow of the dialogue history. The goal is to maintain coherence and sound natural within the ongoing dialogue.

Knowledge Sources

The knowledge provided below includes different structures. Each structure has a name, a formal representation, and a description of how to interpret it.

{{knowledge}}

Prompt Template

You are assisting in a dialogue by generating the next utterance. Use the provided scaffold, dialogue history, and structured knowledge sources below. ***Prioritise the provided knowledge sources in generating the next utterance, but if a suitable utterance can't be generated you may use your own knowledge, in which case identify the source as "generated".***

Dialogue History
{{history}}

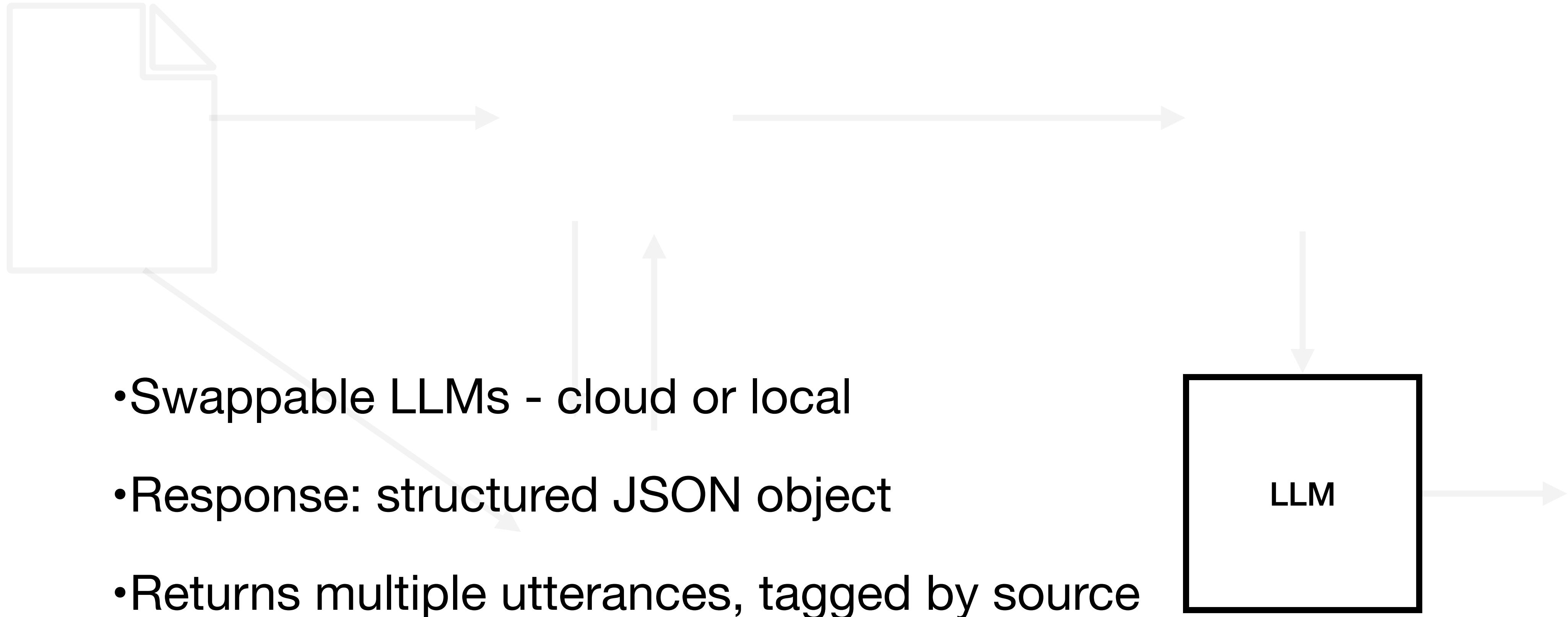
Utterance Scaffold
{{Utterance}}

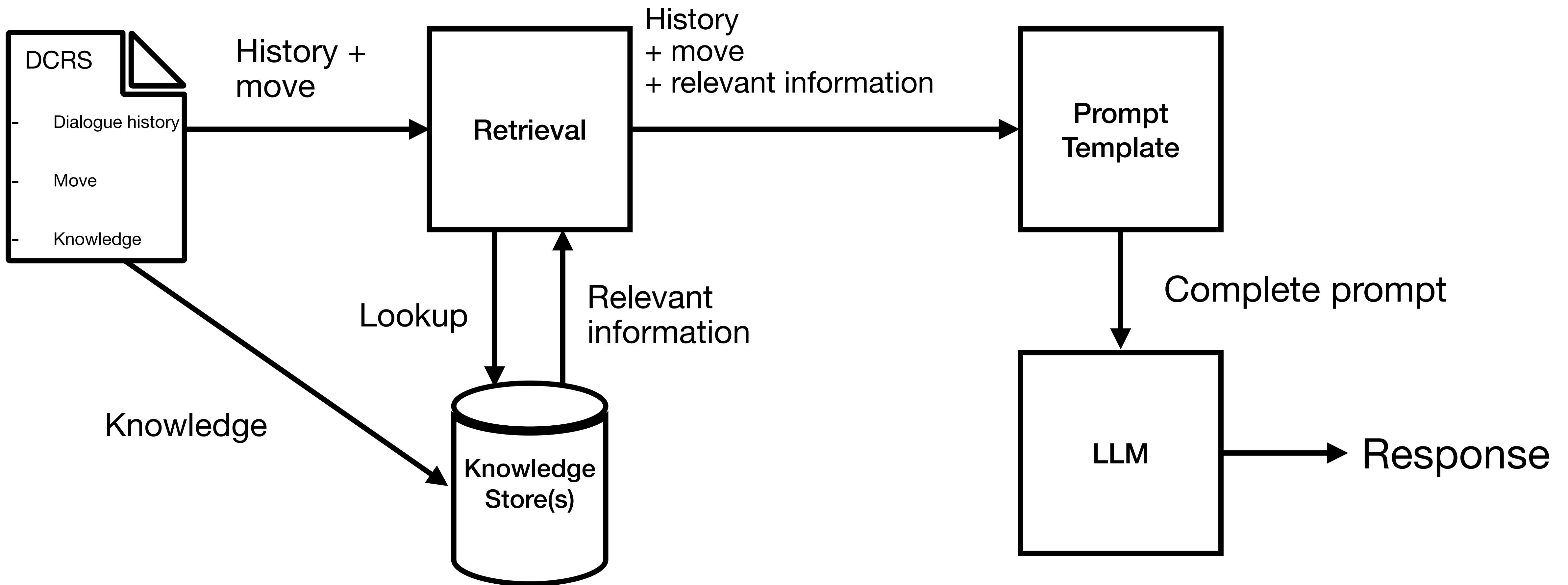
You may adapt the scaffold to better match the tone, style, or conversational flow of the dialogue history. The goal is to maintain coherence and sound natural within the ongoing dialogue.

Knowledge Sources

The knowledge provided below includes different structures. Each structure has a name, a formal representation, and a description of how to interpret it.

{{knowledge}}





We should go to
the cinema

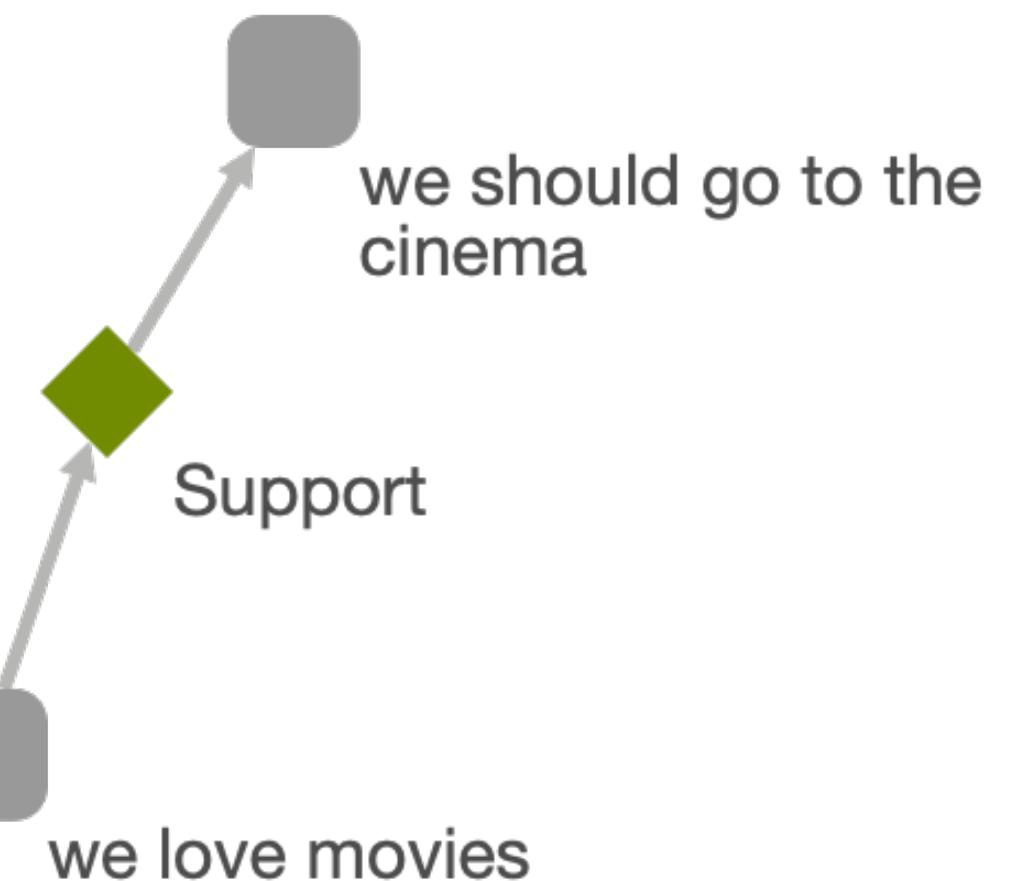
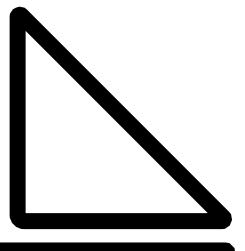
Why?

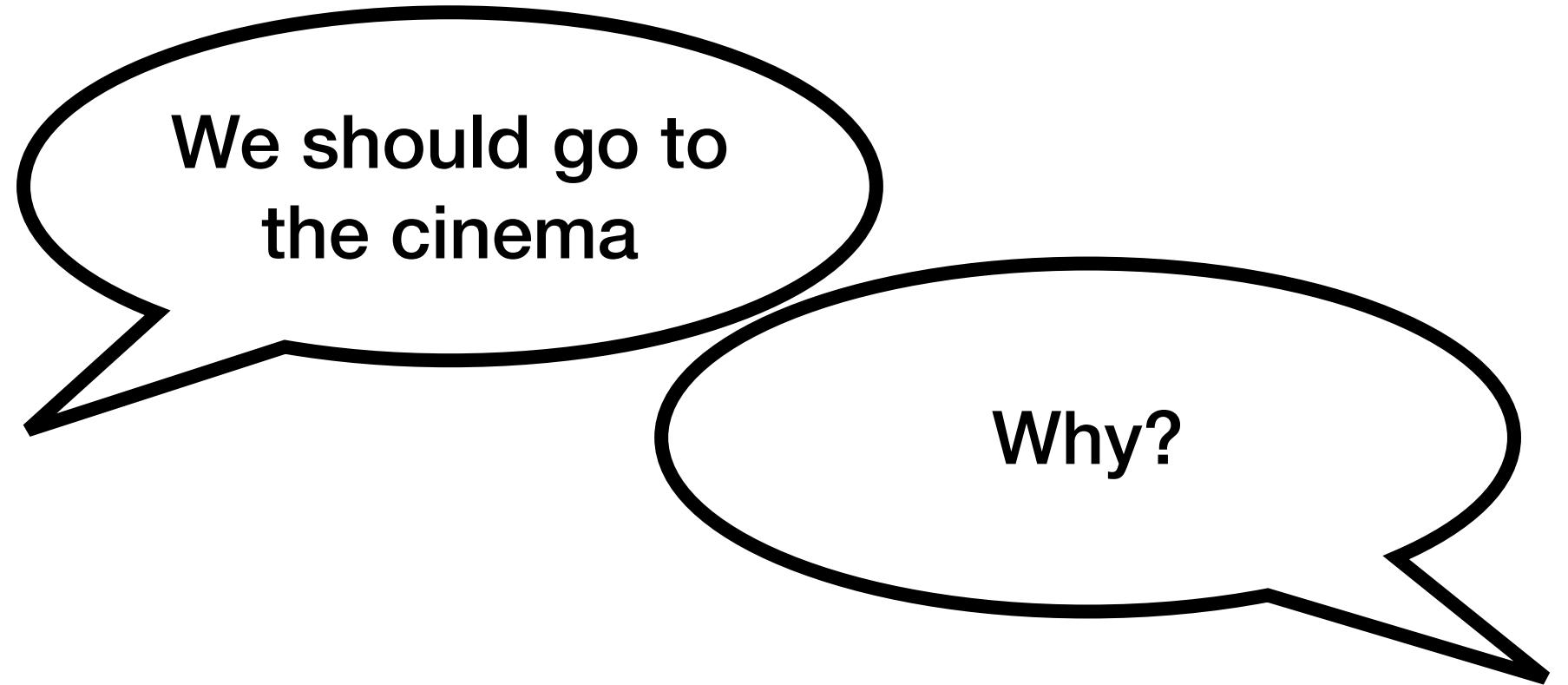
We should go to
the cinema

Why?

Knowledge

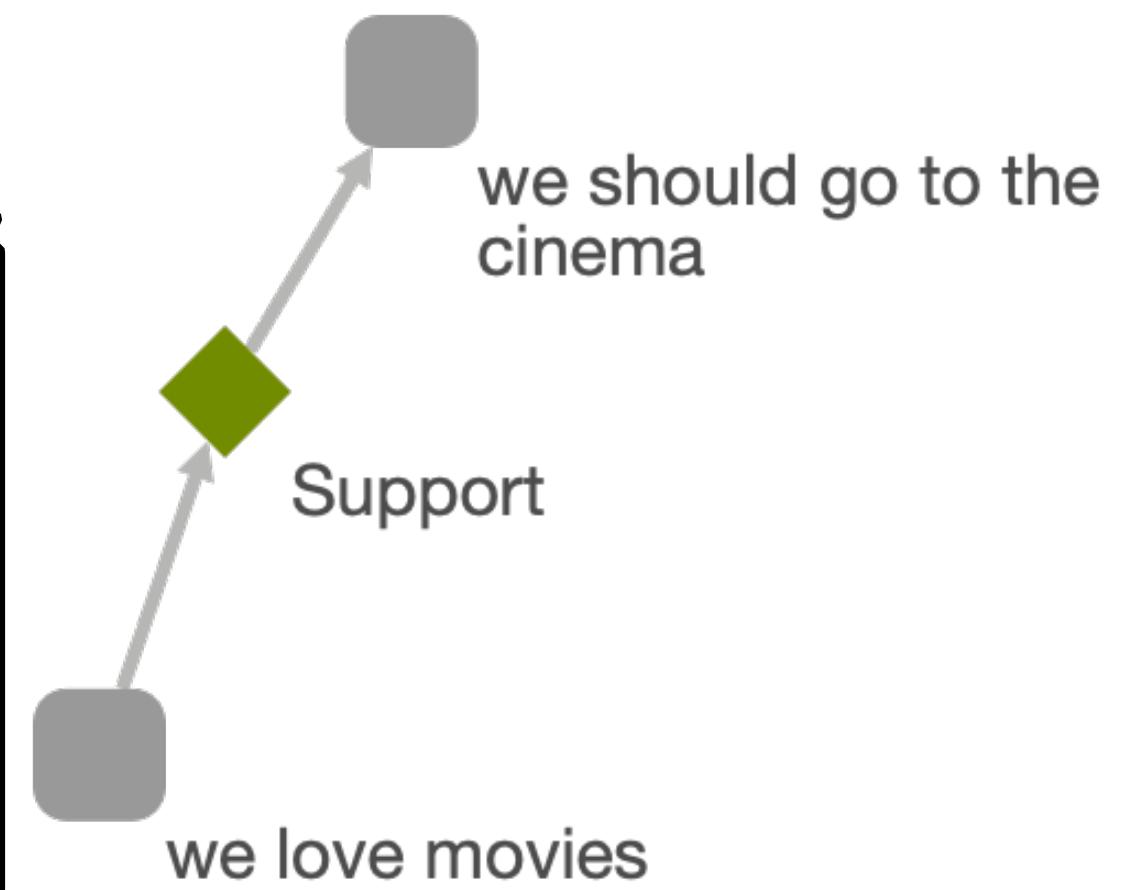
Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...





Knowledge

Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...



Dialogue history

Alice: We should go to the cinema [assert]

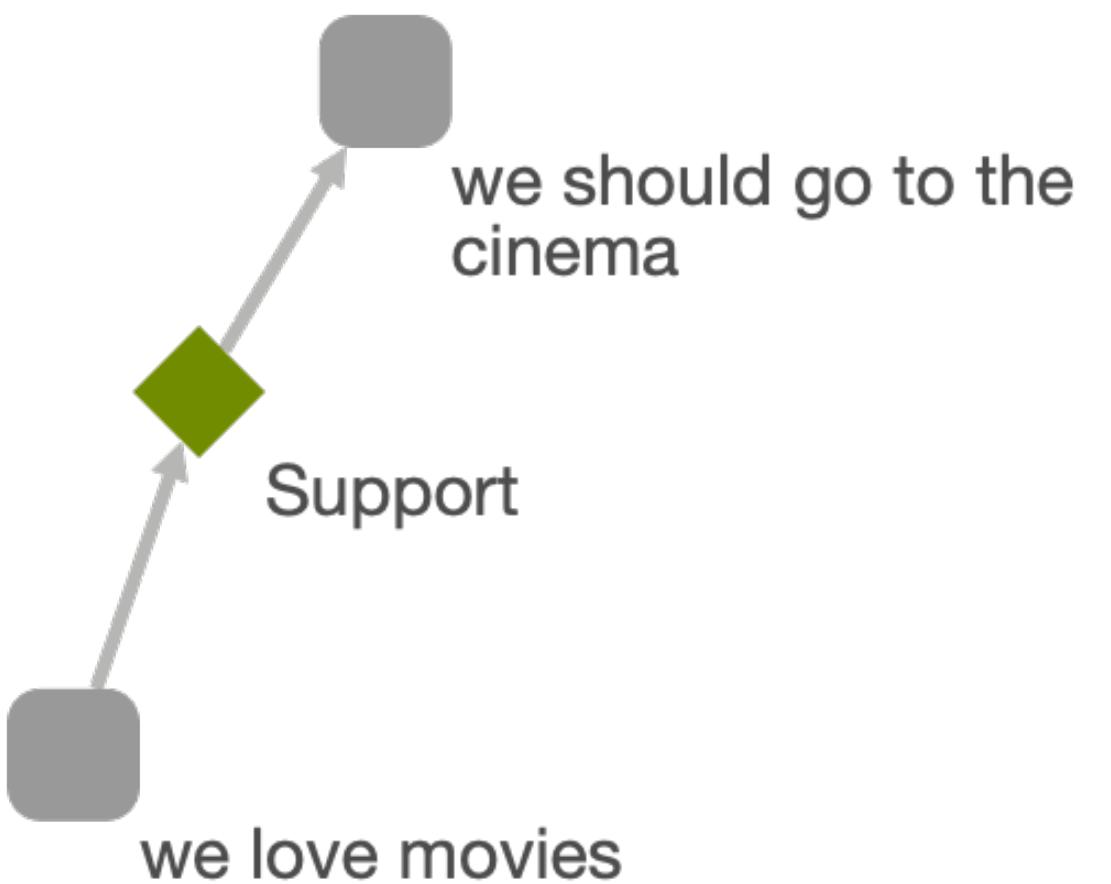
Bob: Why? [challenge]

We should go to
the cinema

Why?

Knowledge

Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...

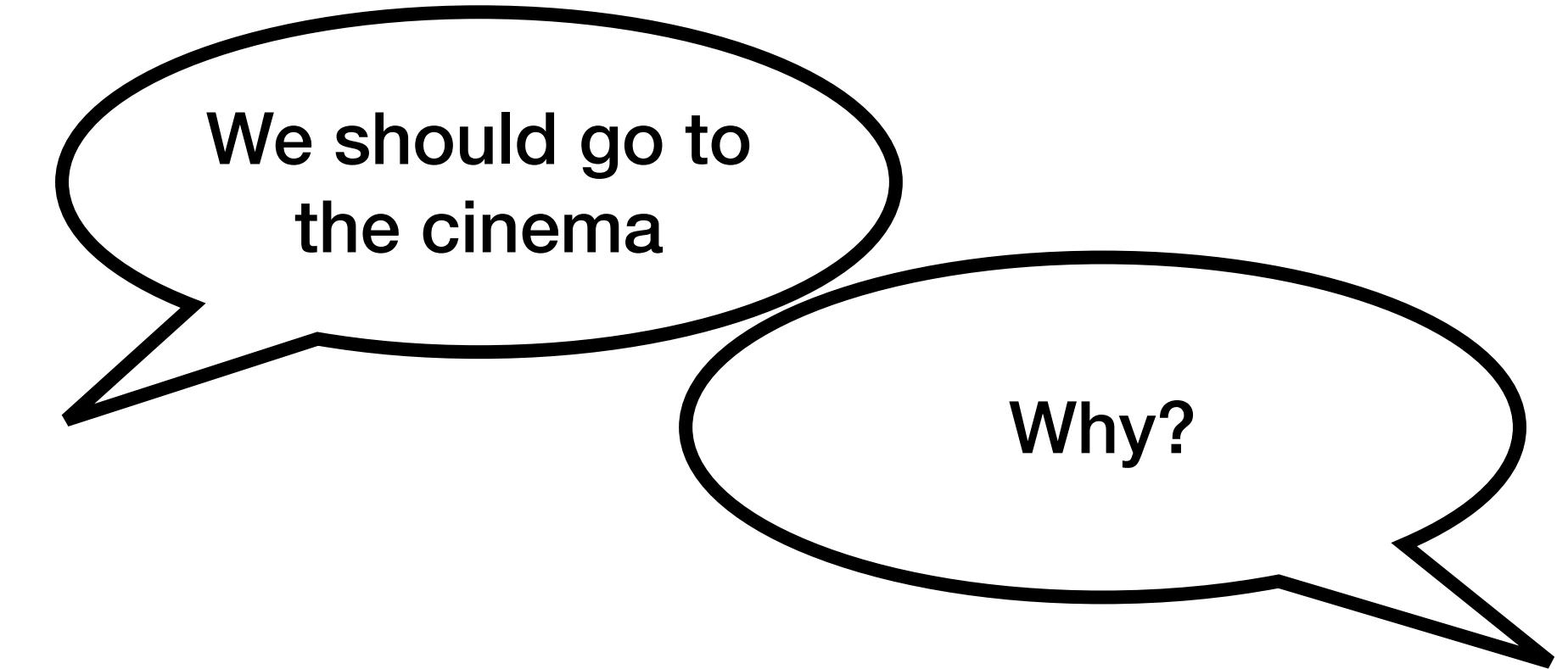


Dialogue history

Alice: We should go to the cinema [assert]
Bob: Why? [challenge]

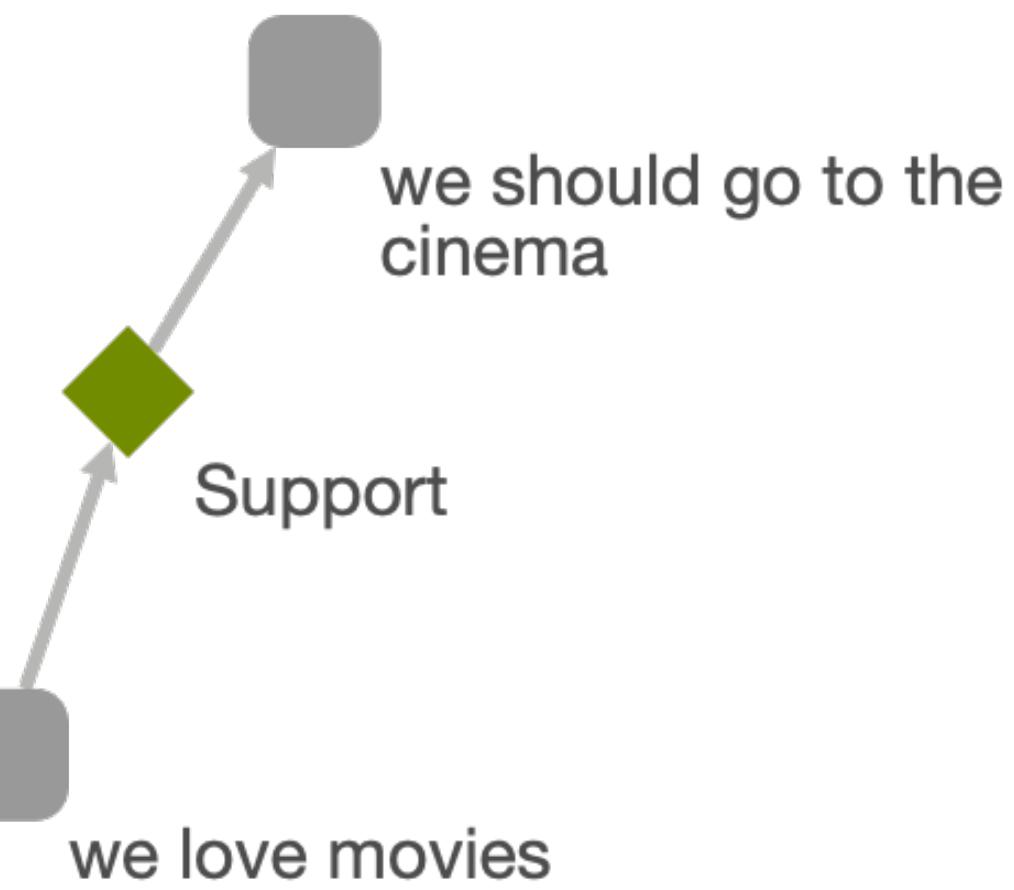
Scaffold

"We should go to the cinema because \$p"



Knowledge

Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...



Dialogue history

Alice: We should go to the cinema [assert]
Bob: Why? [challenge]

Natural language text to interpret. Rephrase if needs be to make short responses using this source.

Scaffold

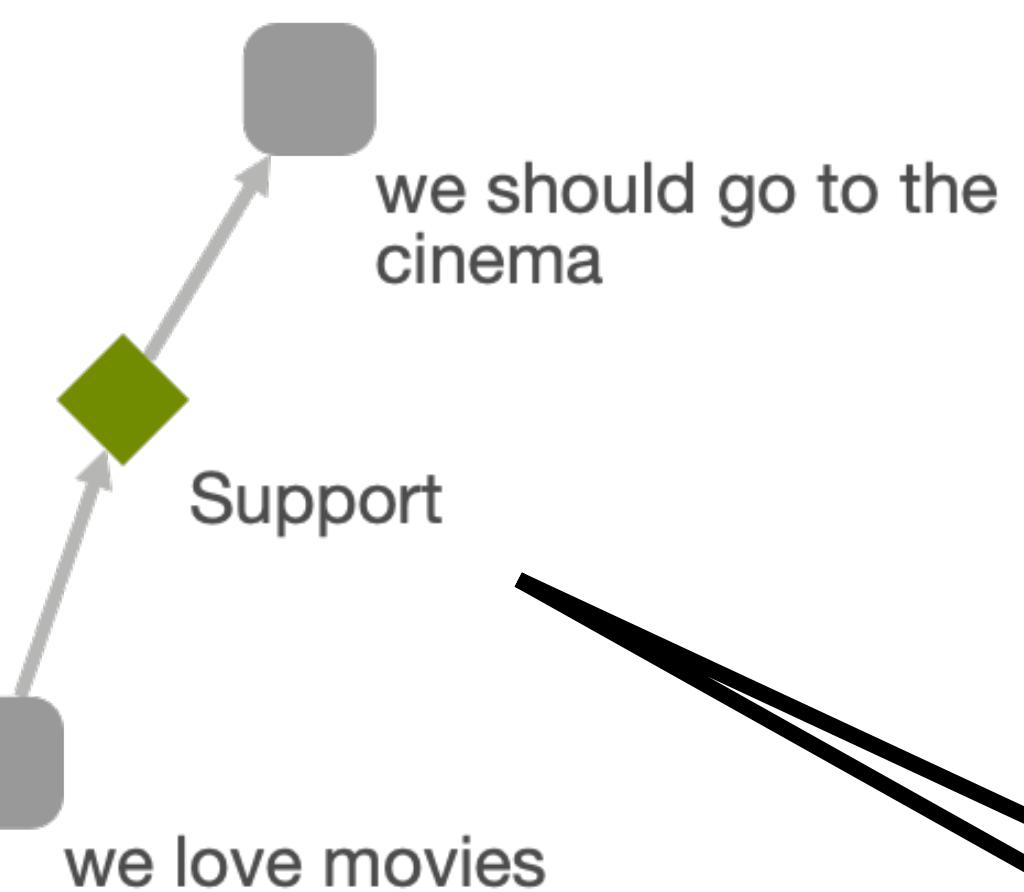
"We should go to the cinema because \$p"

We should go to
the cinema

Why?

Knowledge

Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...



Dialogue history

Alice: We should go to the cinema [assert]
Bob: Why? [challenge]

Natural language text to interpret. Rephrase if needs be to make short responses using this source.

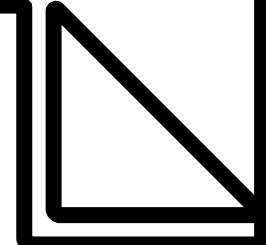
- sadface-inference
Format: $p_1, \dots, p_n \Rightarrow q$
Description: p_1, \dots, p_n are premises that together support the conclusion q

Scaffold

"We should go to the cinema because \$p"

Know

Showing this week is the brand-new Fantastic Four movie, and it's one you really shouldn't miss. The film reintroduces Marvel's original superhero team with a fresh energy that balances action, humour, and...



```
{  
  "responses": [  
    {  
      "response": "We should go to the cinema because the brand-new Fantastic  
Four movie is showing this week, with great effects and a fun team dynamic  
that's best enjoyed on the big screen.",  
      "source": "text"  
    },  
    {  
      "response": "We should go to the cinema because we love movies, and  
it's a perfect chance to enjoy one together.",  
      "source": "sadface"  
    }  
  ]  
}
```

fold

*to the
use \$p"*

Ongoing and future work

- Evaluation - Two types:
 - Is the LLM giving “good” responses?
 - How “good” are the responses from different LLMs?
- Combining knowledge sources in a single response
 - Currently limited to one source per response
- Strategic support - LLM judgement of the “best” response to put forward

DiSCoAI - Dialogue-based Structured Conversational AI

2025-2028

- Automated implementation of computational dialogue games from real (human-human) examples
- Underpinning conversational systems with formal dialogue without the design and development overhead
- When combined with RAG pipeline = rapid development of rich and immersive systems across a wide variety of domains



Engineering and
Physical Sciences
Research Council



<https://gtr.ukri.org/projects?ref=220>

Summary

- LLM and RAG-based pipeline for generated responses in formal dialogue
- Knowledge, history and move scaffold used to ground the response
- Bridges the vocabulary gap to make formal dialogue responses more natural

Thank you