Dialogue Modelling

Raquel Fernández
Institute for Logic, Language and Computation (ILLC)

NLP1 guest lecture, December 2024

Plan for today

Part 1:

- What is dialogue modelling?
- NLP methods to model text-based dialogue
    - Modular statistical approaches
    - End-to-end encoder-decoder models
    - Generative (decoder-only) large language models

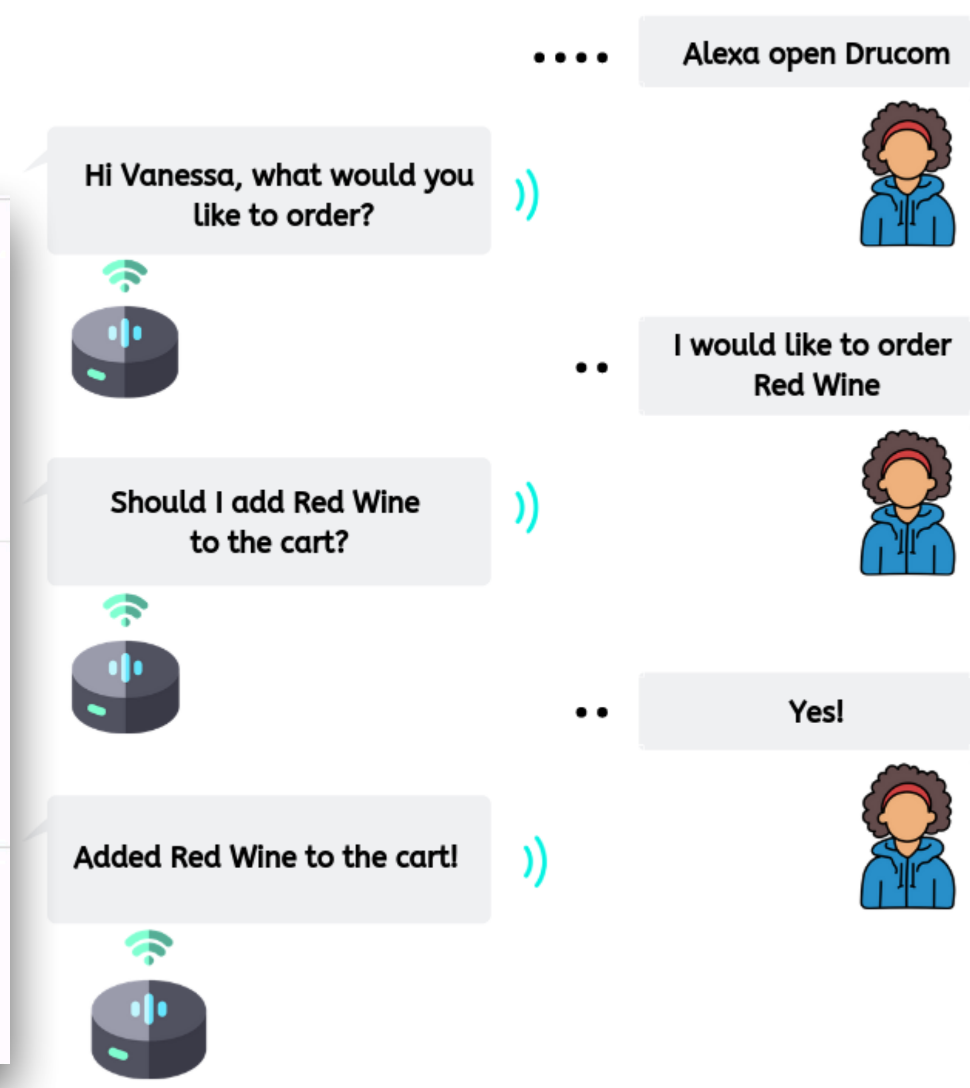
Part 2:

- Face-to-face dialogue
    - Modelling speech and gestures

Dialogue

What is it and why do we care

- Using language for cross-speaker communication and interaction
- Primary form of language use and language learning



Dialogue modelling

Modelling a dialogue agent involves:

- ☑ Understanding the utterances by the dialogue partner.
- ☑ Keeping track of the dialogue history.
- ☑ Deciding what to say.
- ☑ Generating an utterance that conveys the agent’s intend.

This requires many complex abilities: common sense reasoning, theory of mind, planning, …

Dialogue

What is it and why do we care

It is convenient to distinguish between

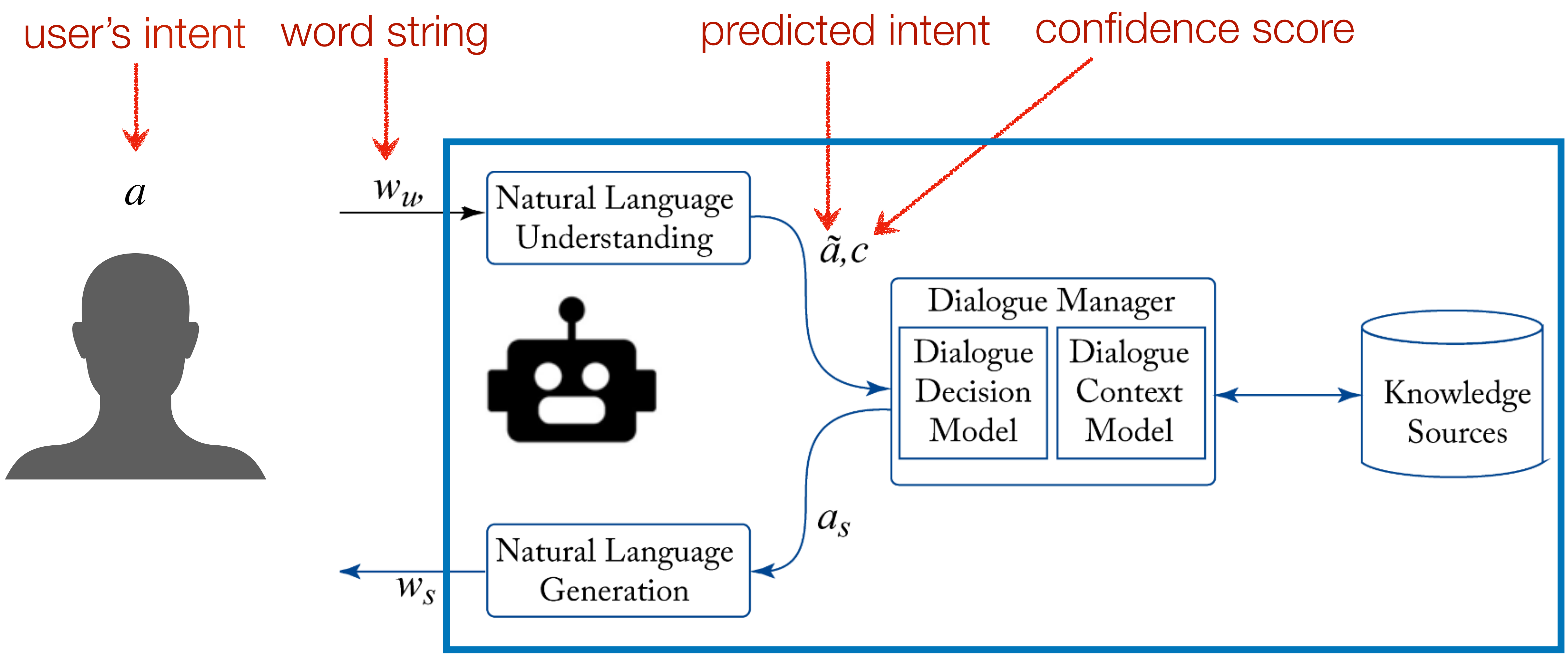
- Social chit-chat dialogue
- Task-oriented dialogue

A: What's your favorite holiday?
B: I'm a big fan of Christmas.
A: Is that so? Mine is Holloween.
B: I also like Holloween. But I like Christmas most.

PC: Alexa, open plan my trip.
ALEXA: Where are you planning to go?
PC: I'm going to Portland.
ALEXA: What city are you leaving from?
PC: Seattle.
ALEXA: What date are you flying out?
PC: Next Thursday.
ALEXA: This will be fun. You go from Seattle to Portland on April 27th, 2017.

Modular dialogue agents

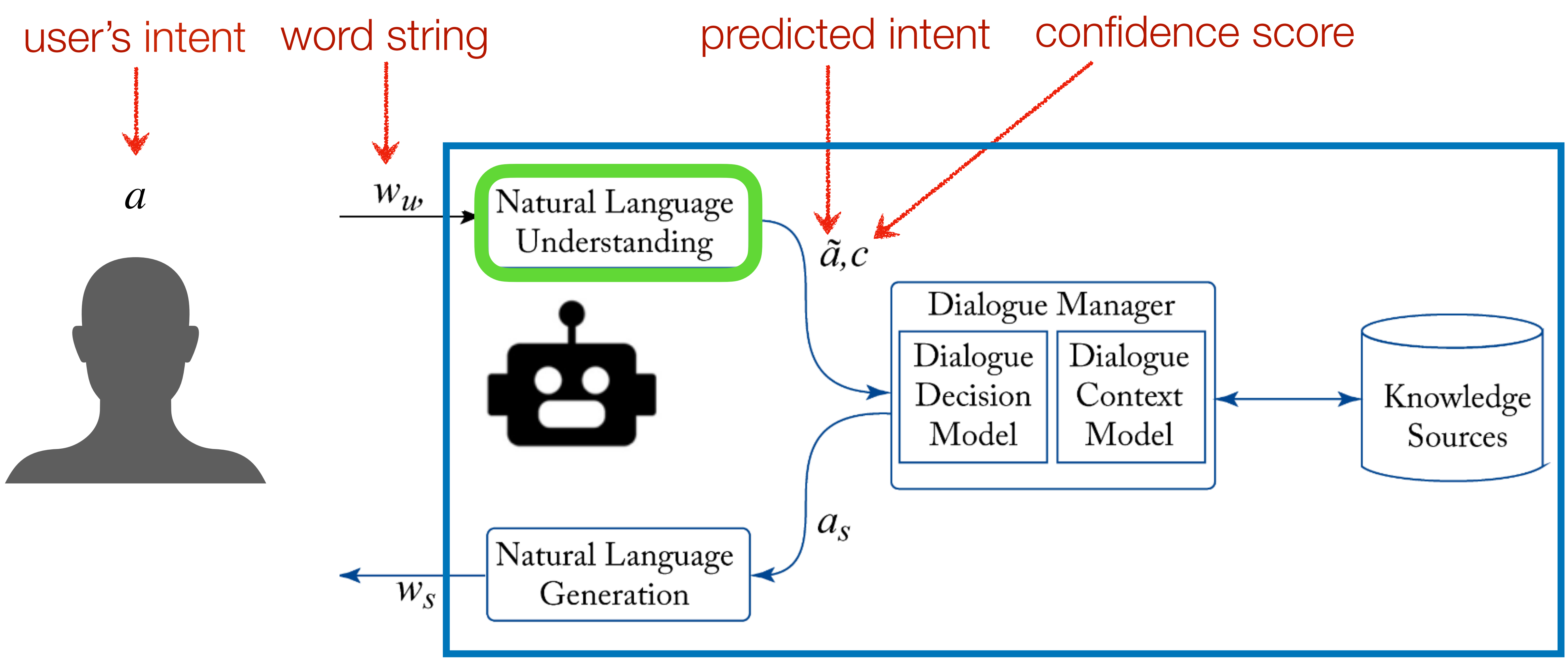
Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently



(Based on McTear, 2020)

Modular dialogue agents

Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently



(Based on McTear, 2020)

NLU

Intent prediction: Why is it difficult?

Speech act or **dialogue act**: the function of (or the action performed by) an utterance. The intention of the speaker.

- *statement, question, answer, agreement, request*, ….

- There isn't a one-to-one mapping between form and function (between the word string and the dialogue act)

  The gun is loaded. *Threat? Warning? Statement?*

- It may require inference (e.g., computing a "conversational implicature"):

  A: Are you going to Paul's party?
  B: I have to work.
  (=> I'm not going — *negative answer*)

NLU

Intent prediction: in practice

Predict a **meaning representation** given the word string.

In task-oriented dialogue, these are usually “frames” consisting of:

- Domain of the conversation (if not pre-defined)
- Each domain, has a set of possible user intents (task goals).
- Each intent, has a set of possible slots and slot values.

What are possible morning flights from Boston to SF on Tuesday?

```
DOMAIN:        AIR-TRAVEL
INTENT:        SHOW-FLIGHTS
ORIGIN-CITY:   Boston
ORIGIN-DATE:   Tuesday
ORIGIN-TIME:   morning
DEST-CITY:     San Francisco
```

Wake me tomorrow at six.

```
DOMAIN:   ALARM-CLOCK
INTENT:   SET-ALARM
TIME:     2017-07-01 0600-0800
```

NLU

Intent prediction: in practice

- Many of the NLP techniques you have seen in this course are relevant for intent prediction in dialogue:

  - word embeddings, POS tagging, syntactic parsing, compositional semantics, etc.

- This approach requires **annotated dialogue datasets** where utterances are annotated with meaning representations.

What are possible morning flights from Boston to SF on Tuesday?

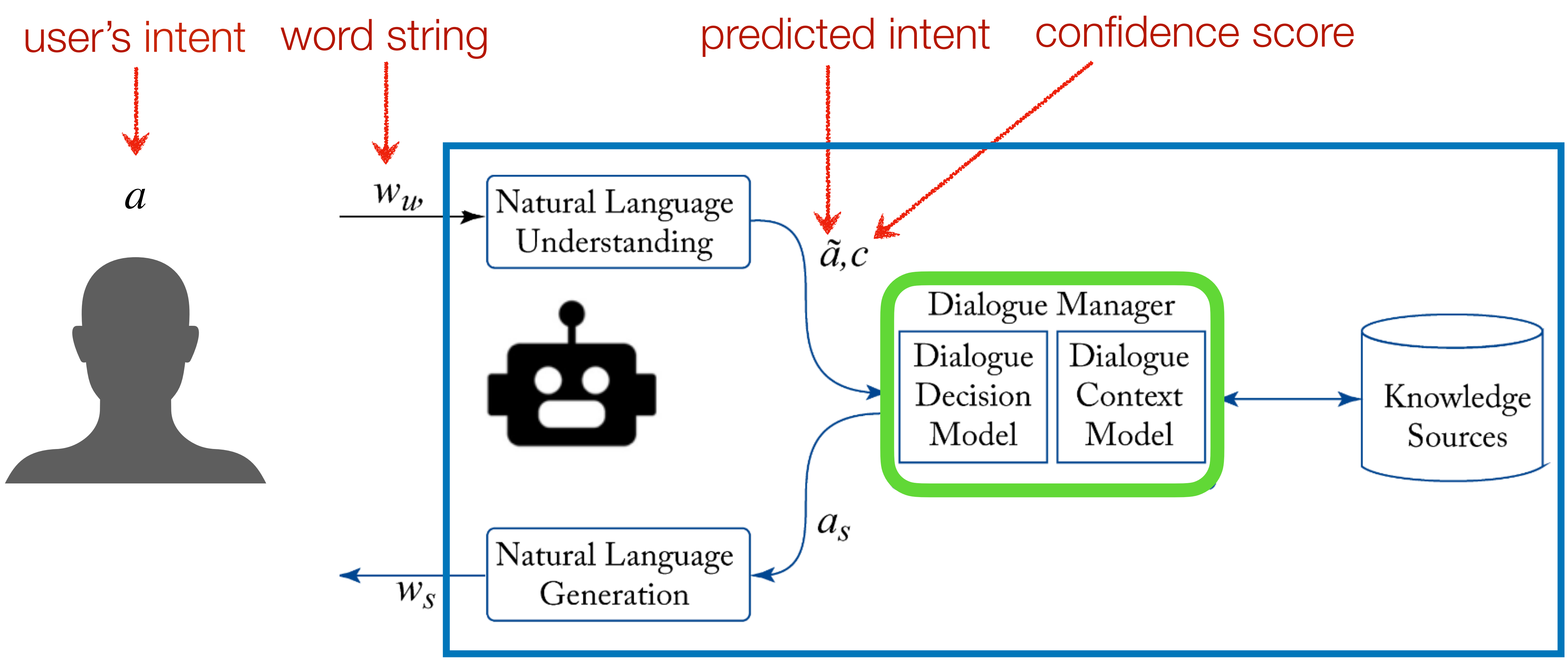
```
DOMAIN:       AIR-TRAVEL
INTENT:       SHOW-FLIGHTS
ORIGIN-CITY:  Boston
ORIGIN-DATE:  Tuesday
ORIGIN-TIME:  morning
DEST-CITY:    San Francisco
```

Wake me tomorrow at six.

```
DOMAIN:   ALARM-CLOCK
INTENT:   SET-ALARM
TIME:     2017-07-01 0600-0800
```

Modular dialogue agents

Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently



(Based on McTear, 2020)

Dialogue management

- The relevant slots may be filled across multiple dialogue turns—the **dialogue context / history** keeps track of this information.
- The **dialogue decision model / policy:** predict the next system action given dialogue context (e.g., slots that are still missing).
  - System intent with the highest probability given the context.

U: Show me morning flights to SF.

DOMAIN:	AIR-TRAVEL
INTENT:	SHOW-FLIGHTS
ORIGIN-CITY:	[ ]
ORIGIN-DATE:	[ ]
ORIGIN-TIME:	morning
DEST-CITY:	San Francisco

→

DOMAIN:	AIR-TRAVEL
INTENT:	REQUEST(ORIGIN-CITY)

S: Where are you flying from?

Dialogue management

Confirmation and rejection

- How likely is the system to have understood the user?
- We can exploit NLU confidence scores to decide on a confirmation/rejection policy:

$< \alpha$	low confidence	reject
$\geq \alpha$	above the threshold	confirm explicitly
$\geq \beta$	high confidence	confirm implictly
$\geq \gamma$	very high confidence	don't confirm at all

`CONFIRM_EXPLICIT(ORIGIN-CITY)`

S: Which city do you want to leave from?
U: Baltimore.
S: **Do you want to leave from Baltimore?**
U: Yes.

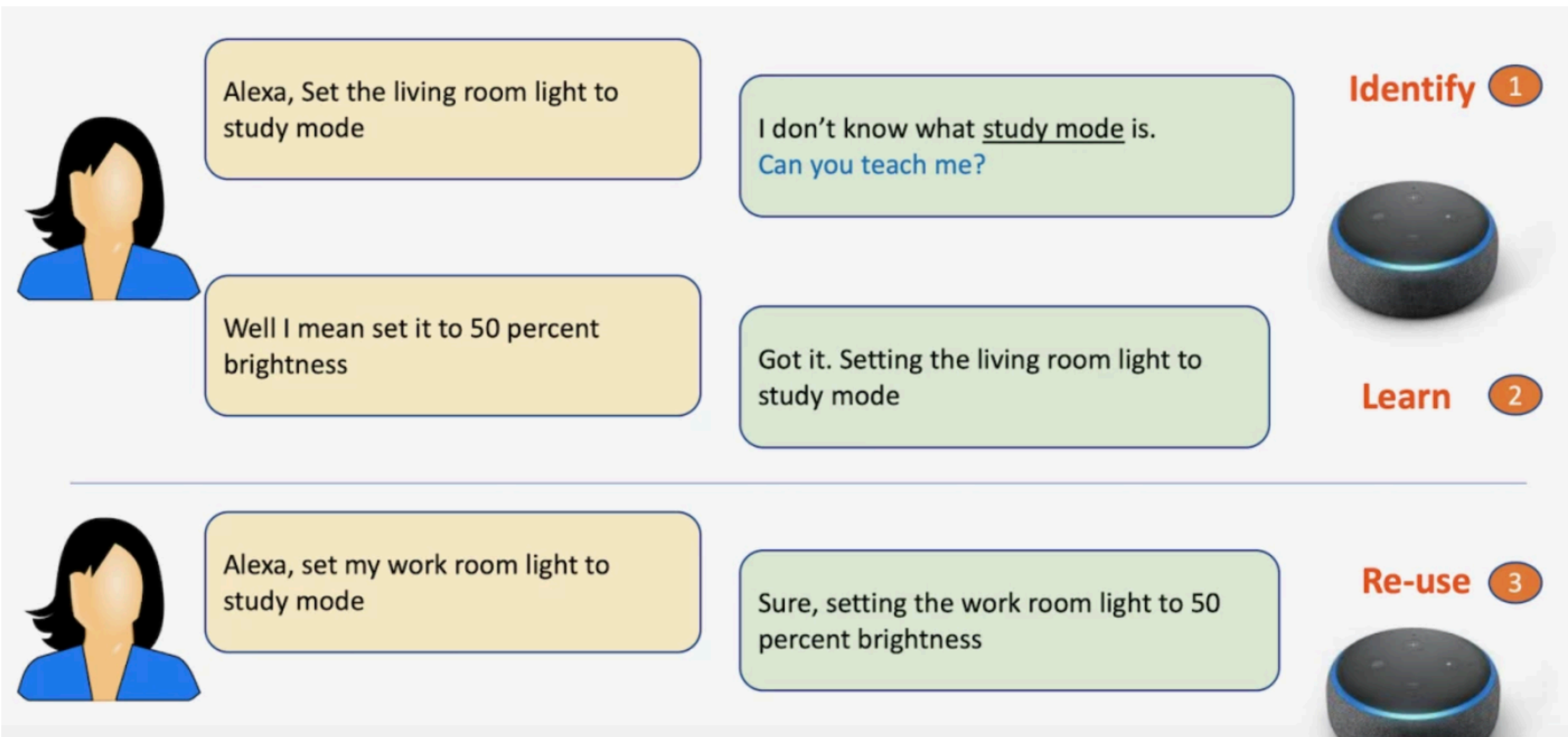
`CONFIRM_IMPLICIT(DEST-CITY)`

U: I want to travel to Berlin
S: **When do you want to travel to Berlin?**

Dialogue management

Learning and generalisation

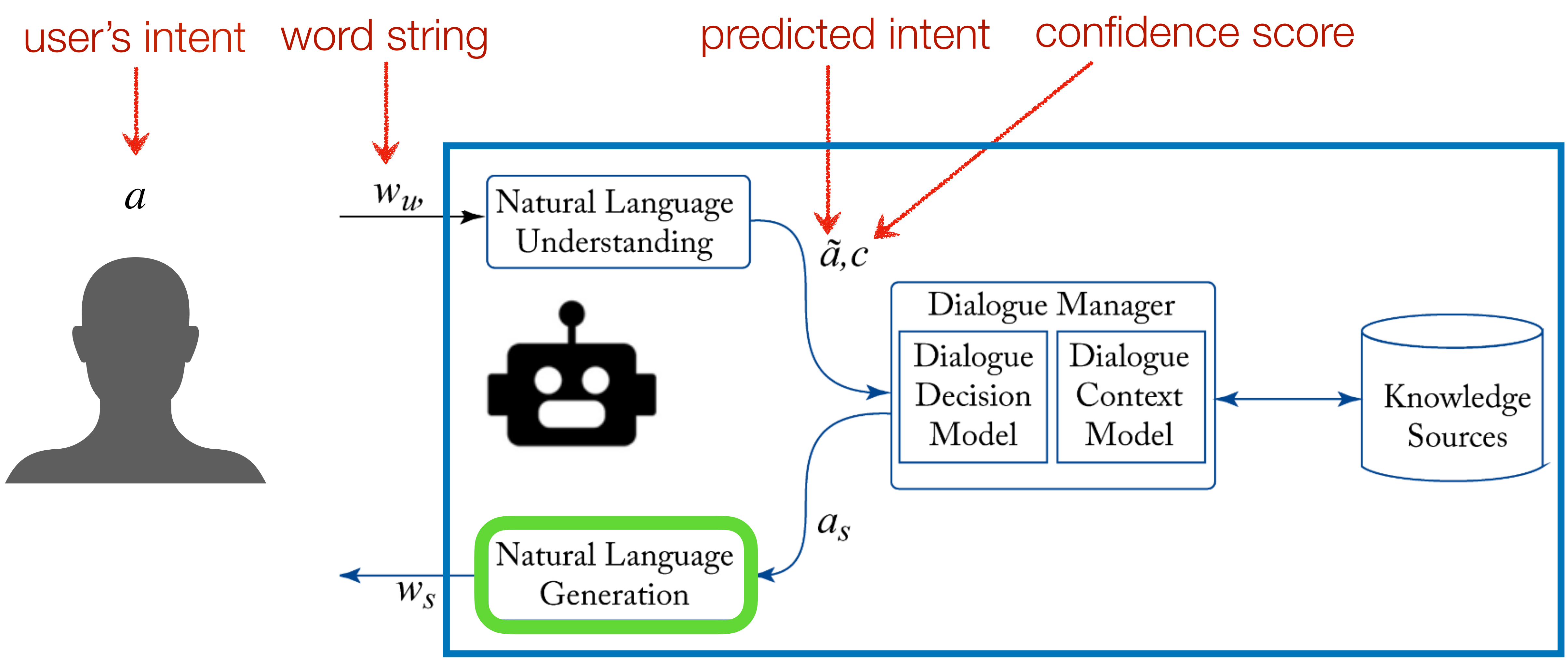
- Confidence scores can also be exploited to identify unknown slots and learn to generalise to new situations



https://www.amazon.science/blog/new-alexa-features-interactive-teaching-by-customers

Modular dialogue agents

Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently



(Based on McTear, 2020)

NLG

- Once the DM has chosen a next system action/intent, the NLG module maps it a string of words.
- Typically, this module is trained to generate sentences using an annotated dialogue corpus with representation/sentence pairs
- Some examples:

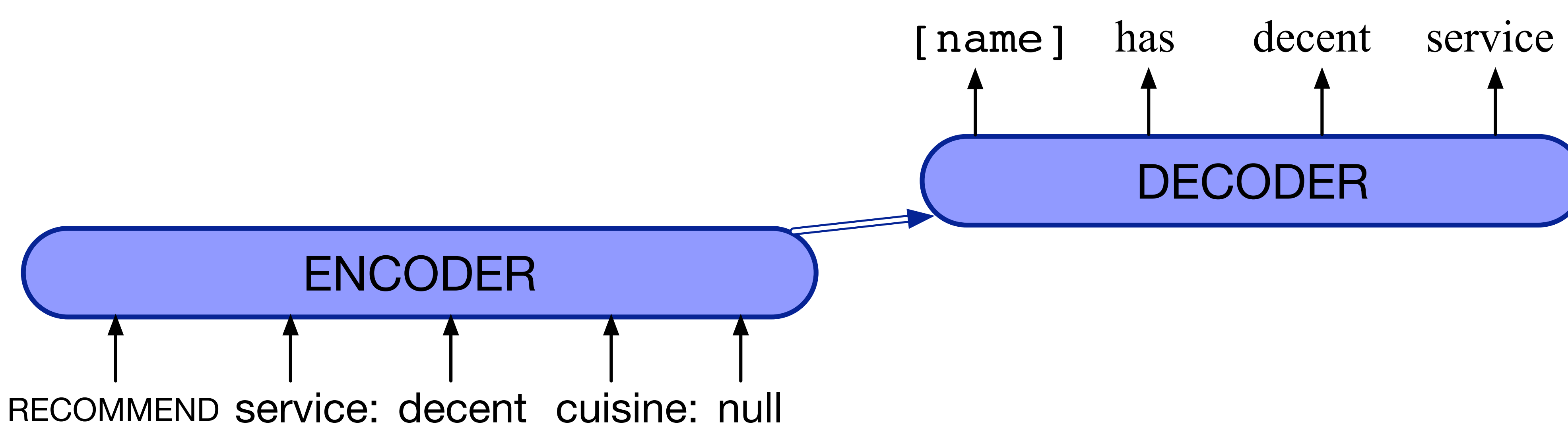
	`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french`
1	Au Midi is in Midtown and serves French food.
2	There is a French restaurant in Midtown called Au Midi.
	`recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)`
3	Loch Fyne is in the City Center and serves seafood food.
4	There is a seafood restaurant in the City Centre called Loch Fyne.

NLG

This can be modelled as **sequence-to-sequence** prediction:

- Input: linearised meaning representation
- Output: word string (system utterance)

	`recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french`
1	Au Midi is in Midtown and serves French food.
2	There is a French restaurant in Midtown called Au Midi.
	`recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)`
3	Loch Fyne is in the City Center and serves seafood food.
4	There is a seafood restaurant in the City Centre called Loch Fyne.

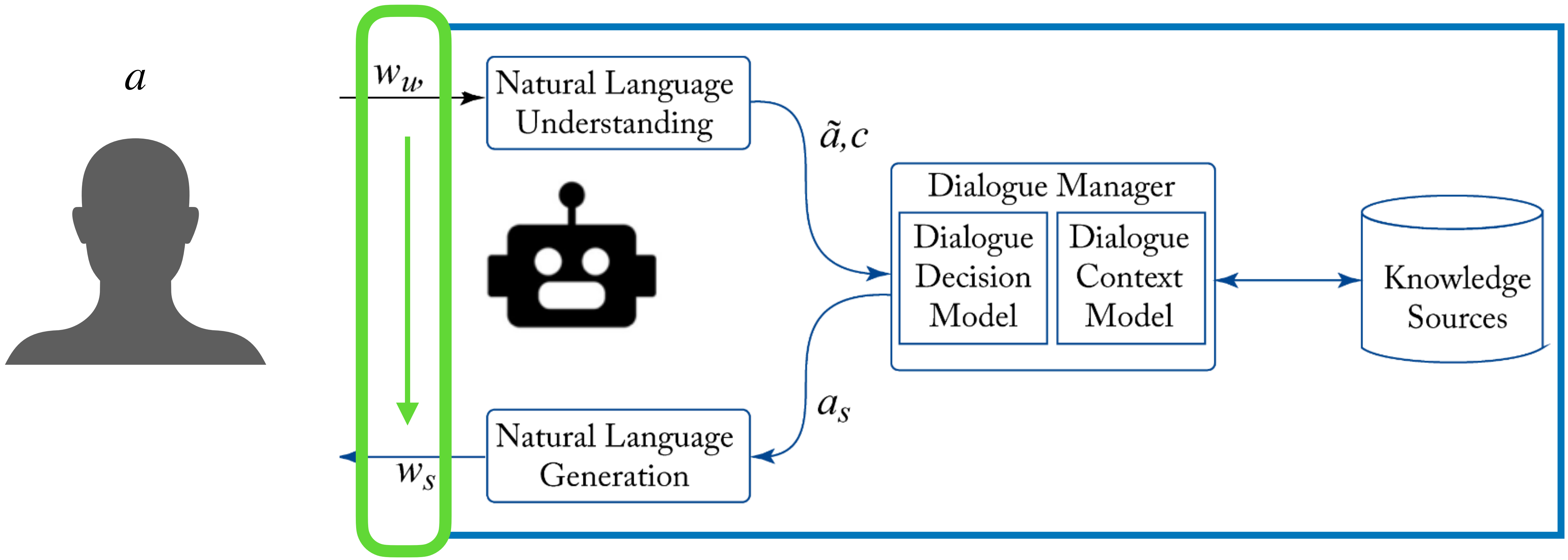


(NB: Delexicalised representation where entities are replaced with general placeholders to help with generalisation)

Modular dialogue agents

Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently

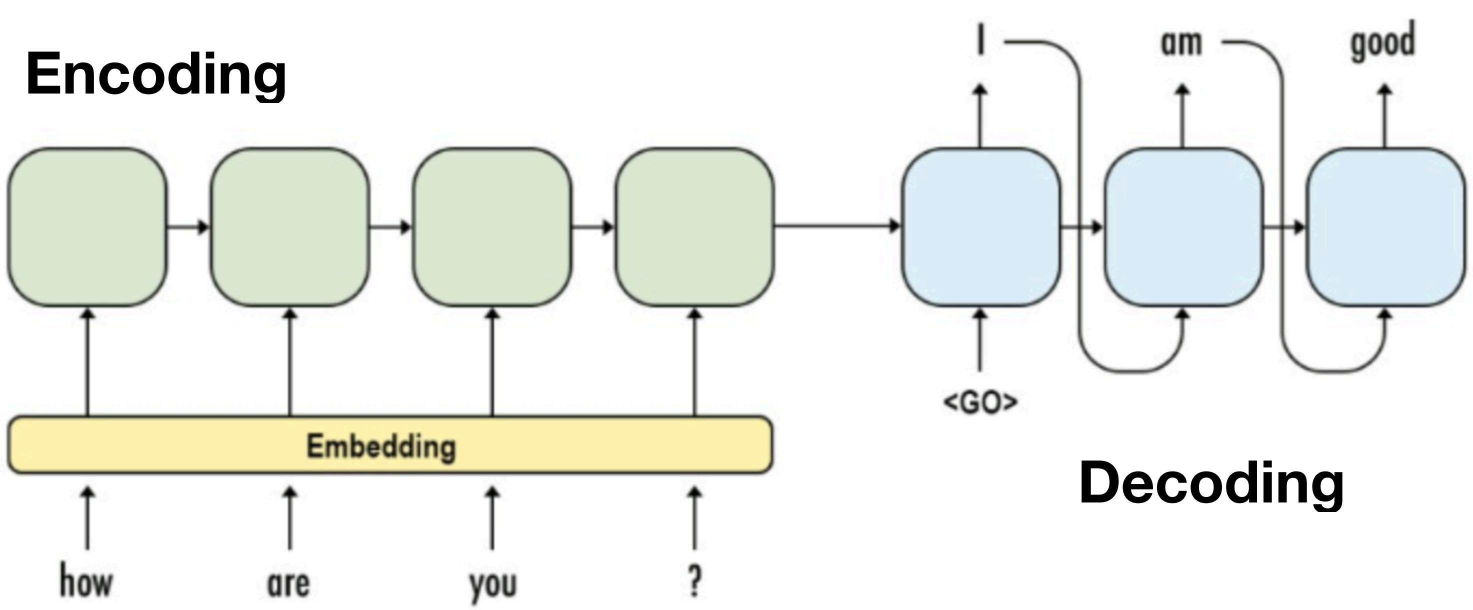
The widespread adoption of **deep learning** gave rise to far less modular approaches, without explicit internal representations



(Based on McTear, 2020)

End-to-end encoder-decoder systems

- **Sequence-to-sequence** dialogue models: an alternative to modular statistical approaches, inspired by machine translation.
- Trained on conversational datasets, without explicit meaning representations.
- Typically used to model social **chit-chat dialogue** (no need to make progress towards task completion)
- **Encoder** RNN to produce a representation of the previous turns
- **Decoder** RNN to generate the response word-by-word by conditioning on the context and the response so far



End-to-end encoder-decoder systems

PROS

- No annotations needed
- No finite, predefined set of meaning representation

CONS

- Very data-hung: trained on dialogue corpora with hundreds or millions of words
- No real understanding
- Tendency to output generic utterances (uninformative, bland, repetitive)
- Simplistic encoding of the dialogue history, leading to semantic inconsistency

U:	tell me ... how long have you had this falling sickness ?
S:	i'm not a doctor.
U:	How long does it take you guys to learn the drill?
S:	Like 2 weeks ago!!
U:	Where do you live?
S:	I live in England.
U:	In which city do you live?
S:	I live in Los Angeles.

[adapted from Jurafsky & Martin, SLP-3]

Modular dialogue agents

Task-oriented dialogue agents have traditionally been modelled using a **modular architecture**, with modules trained independently

The widespread adoption of **deep learning** gave rise to far less modular approaches, without explicit internal representations.

With the advent of **Transformers**, the field has shifted towards the use of large language models for dialogue modelling.

Systems based on generative LLMs

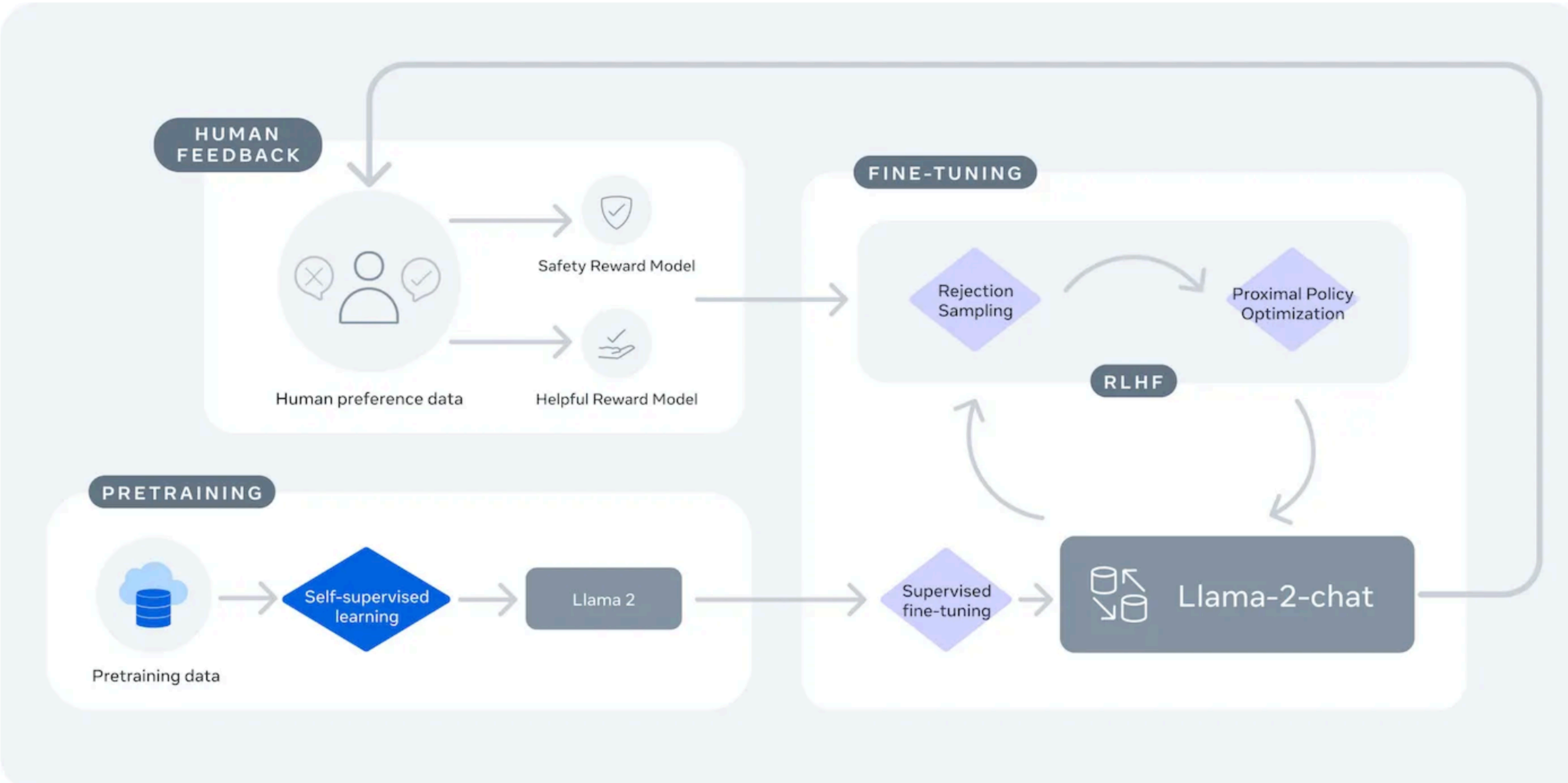
The current alternative to the encoder-decoder architectures:

- Pre-trained decoder-only **large language models** (e.g., the GPT family — see previous lecture on 26-11-2024)
- The Transformer architecture allows for better handling of the dialogue history.

How to turn a base generative LLM into a dialogue agent?

- Dialogue-related instruction tuning + Reinforcement Learning with Human Feedback (RLHF)
- E.g., ChatGPT and other publicly available models such as Llama

(See previous lecture by Pushkar Mishra on 3/12/2024)



Training Llama Chat: Llama 2 is pretrained using publicly available online data. An initial version of Llama Chat is then created through the use of supervised fine-tuning. Next, Llama Chat is iteratively refined using Reinforcement Learning from Human Feedback (RLHF),

https://www.llama.com/llama2/

Examples from the LIMA dataset for instruction tuning (Zhou et al. 2023)

[Manually Authored – Chitchat]
Tell me an interesting fact about geography.

Here are some interesting facts about geography:

- Bhutan is one of the most intriguing places in the world. The Buddhist kingdom is the only carbon-negative country on Earth, with strict environmental laws and cultural norms that preserve its stunning natural landscape. For similar reasons, Bhutan was one of the last countries to introduce television in 1999, and has no traffic lights to this day.
- There are currently few territories on Earth that are unclaimed by any sovereign state (terra nullius). One example is Bir Tawil, along the border between Egypt and Sudan, which was created by a discrepancy between borders drawn in 1899 and 1902. The territory remains unclaimed by either country, because it would entail forfeiting the much larger Halaib Triangle, which is claimed by both countries. Bir Tawil has no settled population, but the land is used by Bedouins who roam the area.
- The antipode of any spot on Earth [...]

[Manually Authored – Advice]
I'm going to NeurIPS for the first time to present the first paper I've ever published. I'm afraid I'll be alone and overwhelmed by everyone. What should I do?

Going to an academic conference for the first time may seem like a bewildering experience, but if done right, conferences are a wonderful opportunity to brainstorm with experts from around the world, present your work and get feedback, and make new friends. Here are some tips for improving your conference experience:

- Reach out in advance to people who have inspired your work. [...]
- Show curiosity about other people's work. [...]
- Sign up for student volunteer [...]
- Ask your advisor to introduce you [...]
- At the same time, try to socialize outside of your lab [...]

Ultimately, it is important to remember that every first-conference attendee is probably feeling a bit shy too, and that taking the first step in starting a conversation will almost always be met with a big smile!

[Manually Authored – Writing]
I am planning to start a book club with some friends. Can you write an email invitation for the same?

Here is an example email invitation for a book club:

Subject: Join me for a book club!

Hi [Friend's Name], how are you?

I'm starting a book club with some friends, and I'd love to have you join us!

The book club is a casual and fun way to read and discuss books that we love or want to learn more about, and also make new friends. We plan to meet more-or-less once a month.

We're thinking of having our first meeting on [Date & Time]. Could that work for you?

Would love to see you there!

Thanks,
[Your Name]

Challenges

The field of dialogue modelling has made a huge step forward thanks to generative LLMs.

- Fluent text generation
- More robustness to unseen scenarios
- Able to exploit knowledge learned during pre-training

Yet, many **challenges** remain:

- Factual and inference errors
- Hallucinations
- Many issues related to safety, social stereotyping and bias

Challenges

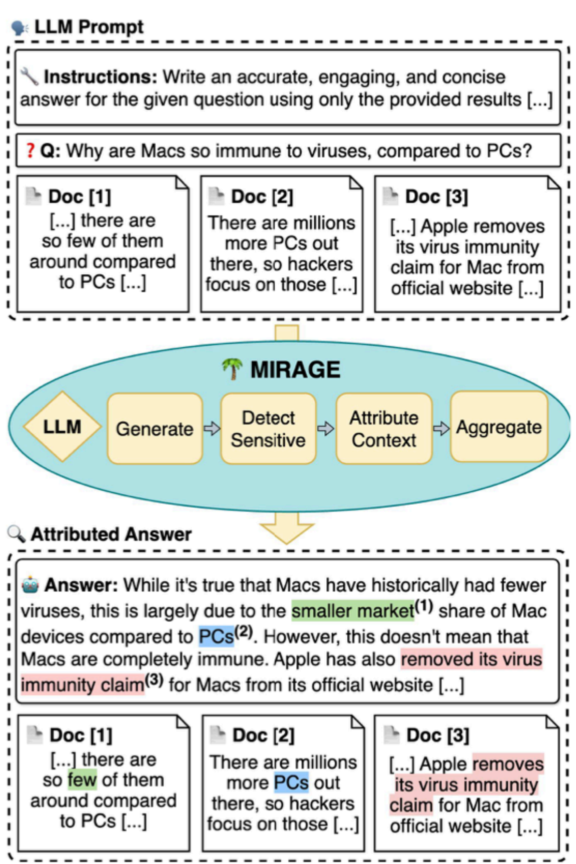
Two example of our recent work addressing some of these challenges:

Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation

Jirui Qi[1*] Gabriele Sarti[1*] Raquel Fernández[2] Arianna Bisazza[1]
[1]Center for Language and Cognition (CLCG), University of Groningen
[2]Institute for Logic, Language and Computation (ILLC), University of Amsterdam
{j.qi, g.sarti, a.bisazza}@rug.nl, raquel.fernandez@uva.nl

Abstract

Ensuring the verifiability of model answers is a fundamental challenge for retrieval-augmented generation (RAG) in the question answering (QA) domain. Recently, self-citation prompting was proposed to make large language models (LLMs) generate citations to supporting documents along with their answers. However, self-citing LLMs often struggle to match the required format, refer to non-existent sources, and fail to faithfully reflect LLMs' context usage throughout the generation. In this work, we present MIRAGE – Model Internals-based RAG Explanations – a plug-and-play approach using model internals for faithful answer attribution in RAG applications. MIRAGE detects context-sensitive answer tokens and pairs them with retrieved documents contributing to their prediction via saliency methods. We evaluate our proposed approach on a multilingual extractive QA dataset, finding high agreement with human answer attribution. On open-ended QA, MIRAGE achieves citation quality and efficiency comparable to self-citation while also allowing for a finer-grained control of attribution parameters. Our qualitative evaluation high-



MBBQ: A Dataset for Cross-Lingual Comparison of Stereotypes in Generative LLMs

Vera Neplenbroek* Arianna Bisazza† Raquel Fernández*
*Institute for Logic, Language and Computation, University of Amsterdam
†Center for Language and Cognition, University of Groningen
{v.e.neplenbroek|raquel.fernandez}@uva.nl, a.bisazza@rug.nl

Abstract

Generative large language models (LLMs) have been shown to exhibit harmful biases and stereotypes. While safety fine-tuning typically takes place in English, if at all, these models are being used by speakers of many different languages. There is existing evidence that the performance of these models is inconsistent across languages and that they discriminate based on demographic factors of the user. Motivated by this, we investigate whether the social stereotypes exhibited by LLMs differ as a function of the language used to prompt them, while controlling for cultural differences and task accuracy. To this end, we present MBBQ (Multilingual Bias Benchmark for Question-answering), a carefully curated version of the English BBQ dataset extended to Dutch, Spanish, and Turkish, which measures stereotypes commonly held across these languages. We further complement MBBQ with a parallel control dataset to measure task performance on the question-answering task independently of bias. Our results based on several open-source and proprietary LLMs confirm that some non-English languages suffer from bias more than English, even when controlling for cultural shifts. Moreover, we observe significant cross-lingual differences in bias behaviour for all except the most accurate models. With the release of MBBQ, we hope to encourage further research on bias in multilingual settings. The dataset and code are available at https://github.com/Veranep/MBBQ.

1 Introduction

Generative large language models (LLMs) have proven useful for tasks ranging from summarization, translation and writing code to answering healthcare and legal questions and taking part in open-domain dialogue (Bang et al., 2023; Zan et al., 2023; Hung et al., 2023). At the same time, a large amount of work has shown that they exhibit various harmful

Plan for today

Part 1:

- What is dialogue modelling?
- NLP methods to model text-based dialogue
  - Modular statistical approaches
  - End-to-end encoder-decoder models
  - Generative (decoder-only) large language models

Part 2:

- Face-to-face dialogue
  - Modelling speech and gestures

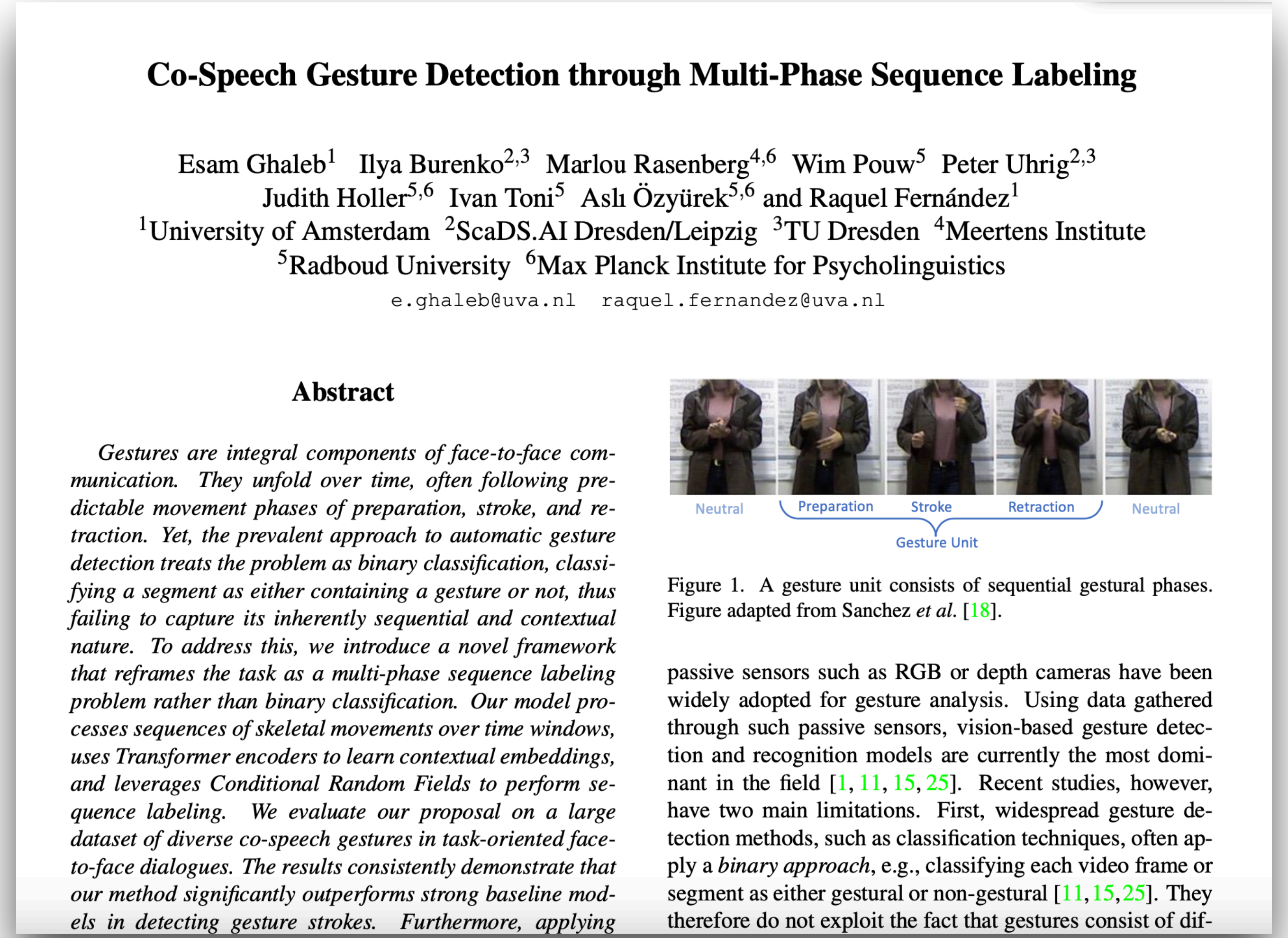
The primary form of language use is **face-to-face dialogue**

Face-to-face dialogue is **multimodal**:

- We NLP exploit a rich array signals: gestures, gaze, facial expressions — and their interplay with speech.

Work with Esam Ghaleb.

Upcoming slides by Esam.



Co-Speech Gesture Detection through Multi-Phase Sequence Labeling

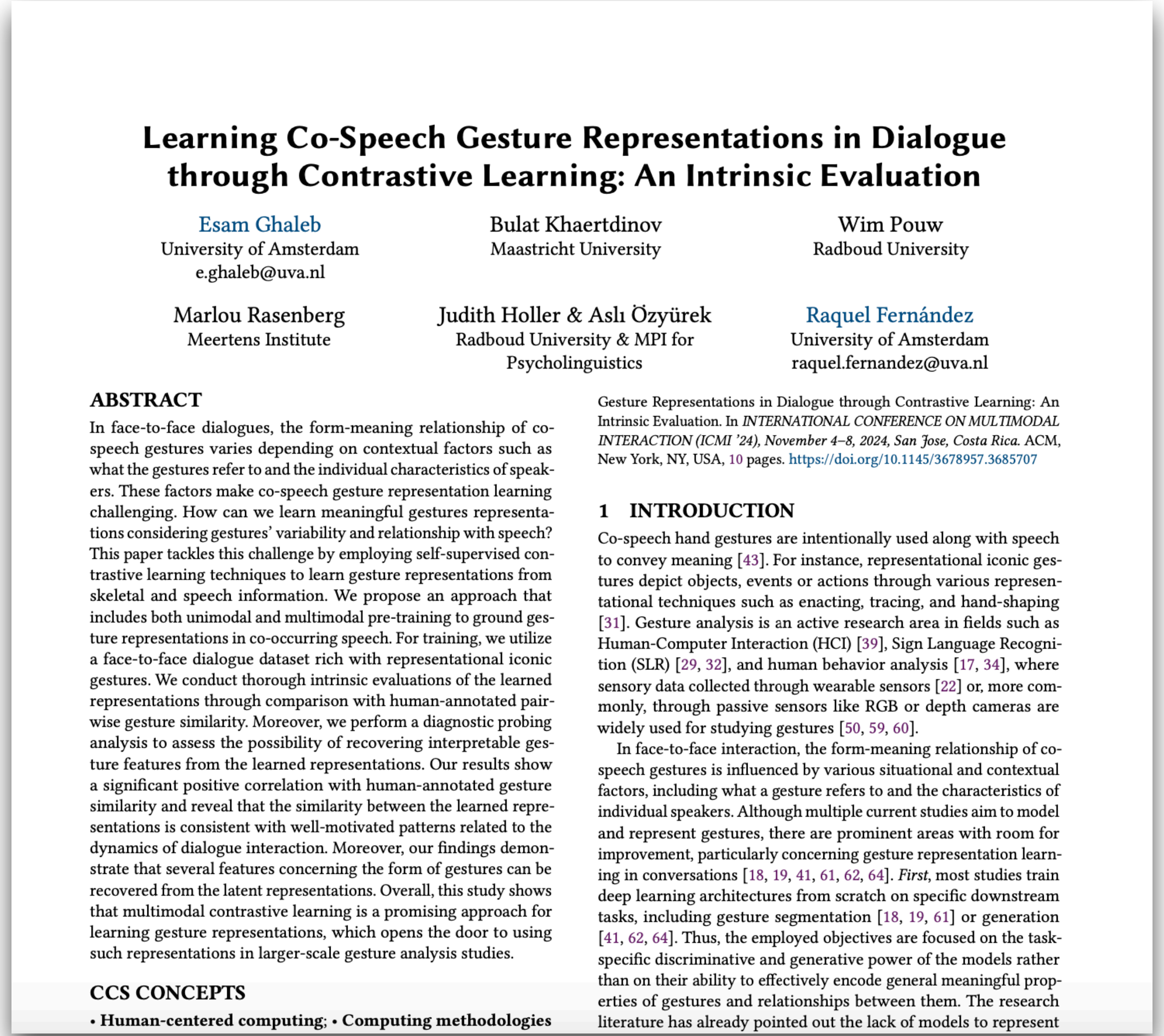
Esam Ghaleb[1] Ilya Burenko[2,3] Marlou Rasenberg[4,6] Wim Pouw[5] Peter Uhrig[2,3]
Judith Holler[5,6] Ivan Toni[5] Aslı Özyürek[5,6] and Raquel Fernández[1]
[1]University of Amsterdam [2]ScaDS.AI Dresden/Leipzig [3]TU Dresden [4]Meertens Institute
[5]Radboud University [6]Max Planck Institute for Psycholinguistics
e.ghaleb@uva.nl raquel.fernandez@uva.nl

Abstract

Gestures are integral components of face-to-face communication. They unfold over time, often following predictable movement phases of preparation, stroke, and retraction. Yet, the prevalent approach to automatic gesture detection treats the problem as binary classification, classifying a segment as either containing a gesture or not, thus failing to capture its inherently sequential and contextual nature. To address this, we introduce a novel framework that reframes the task as a multi-phase sequence labeling problem rather than binary classification. Our model processes sequences of skeletal movements over time windows, uses Transformer encoders to learn contextual embeddings, and leverages Conditional Random Fields to perform sequence labeling. We evaluate our proposal on a large dataset of diverse co-speech gestures in task-oriented face-to-face dialogues. The results consistently demonstrate that our method significantly outperforms strong baseline models in detecting gesture strokes. Furthermore, applying

Figure 1. A gesture unit consists of sequential gestural phases. Figure adapted from Sanchez *et al.* [18].

passive sensors such as RGB or depth cameras have been widely adopted for gesture analysis. Using data gathered through such passive sensors, vision-based gesture detection and recognition models are currently the most dominant in the field [1, 11, 15, 25]. Recent studies, however, have two main limitations. First, widespread gesture detection methods, such as classification techniques, often apply a *binary approach*, e.g., classifying each video frame or segment as either gestural or non-gestural [11, 15, 25]. They therefore do not exploit the fact that gestures consist of dif-



Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation

Esam Ghaleb
University of Amsterdam
e.ghaleb@uva.nl

Bulat Khaertdinov
Maastricht University

Wim Pouw
Radboud University

Marlou Rasenberg
Meertens Institute

Judith Holler & Aslı Özyürek
Radboud University & MPI for Psycholinguistics

Raquel Fernández
University of Amsterdam
raquel.fernandez@uva.nl

ABSTRACT

In face-to-face dialogues, the form-meaning relationship of co-speech gestures varies depending on contextual factors such as what the gestures refer to and the individual characteristics of speakers. These factors make co-speech gesture representation learning challenging. How can we learn meaningful gestures representations considering gestures' variability and relationship with speech? This paper tackles this challenge by employing self-supervised contrastive learning techniques to learn gesture representations from skeletal and speech information. We propose an approach that includes both unimodal and multimodal pre-training to ground gesture representations in co-occurring speech. For training, we utilize a face-to-face dialogue dataset rich with representational iconic gestures. We conduct thorough intrinsic evaluations of the learned representations through comparison with human-annotated pairwise gesture similarity. Moreover, we perform a diagnostic probing analysis to assess the possibility of recovering interpretable gesture features from the learned representations. Our results show a significant positive correlation with human-annotated gesture similarity and reveal that the similarity between the learned representations is consistent with well-motivated patterns related to the dynamics of dialogue interaction. Moreover, our findings demonstrate that several features concerning the form of gestures can be recovered from the latent representations. Overall, this study shows that multimodal contrastive learning is a promising approach for learning gesture representations, which opens the door to using such representations in larger-scale gesture analysis studies.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies**

Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685707

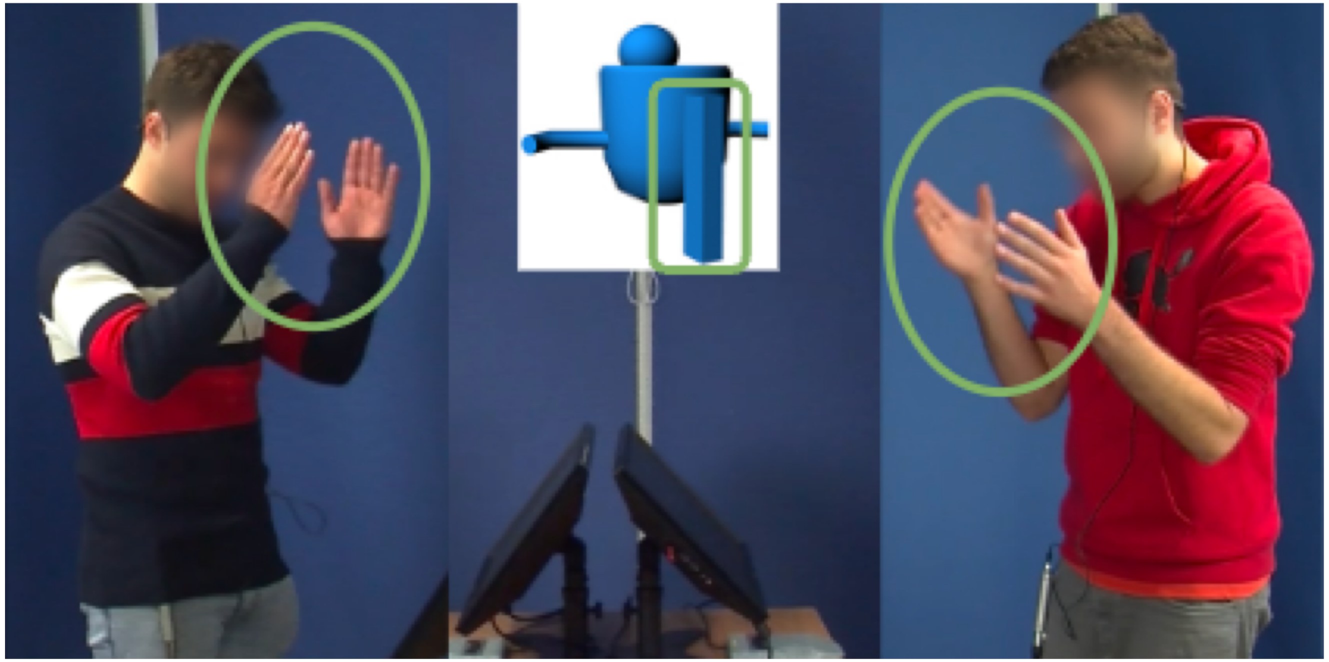
1 INTRODUCTION

Co-speech hand gestures are intentionally used along with speech to convey meaning [43]. For instance, representational iconic gestures depict objects, events or actions through various representational techniques such as enacting, tracing, and hand-shaping [31]. Gesture analysis is an active research area in fields such as Human-Computer Interaction (HCI) [39], Sign Language Recognition (SLR) [29, 32], and human behavior analysis [17, 34], where sensory data collected through wearable sensors [22] or, more commonly, through passive sensors like RGB or depth cameras are widely used for studying gestures [50, 59, 60].

In face-to-face interaction, the form-meaning relationship of co-speech gestures is influenced by various situational and contextual factors, including what a gesture refers to and the characteristics of individual speakers. Although multiple current studies aim to model and represent gestures, there are prominent areas with room for improvement, particularly concerning gesture representation learning in conversations [18, 19, 41, 61, 62, 64]. *First*, most studies train deep learning architectures from scratch on specific downstream tasks, including gesture segmentation [18, 19, 61] or generation [41, 62, 64]. Thus, the employed objectives are focused on the task-specific discriminative and generative power of the models rather than on their ability to effectively encode general meaningful properties of gestures and relationships between them. The research literature has already pointed out the lack of models to represent

Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation

- **Motivation:** Co-speech gestures in face-to-face dialogues vary widely depending on context (e.g., gesture meaning, speaker differences).



Ghaleb, E., Khaertdinov, B., Pouw, W., Rasenberg, M., Holler, J., Ozyurek, A., & Fernandez, R. (2024, November). Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation. In Proceedings of the 26th International Conference on Multimodal Interaction (pp. 274-283).

Research Questions

Grounding on spoken language

How can we learn gesture representations, grounding them with co-occuring speech?

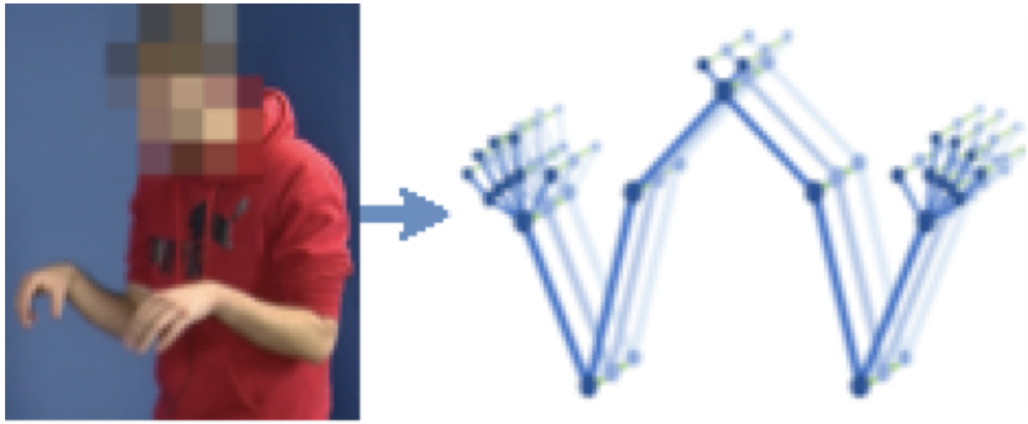
Robust meaningful representations

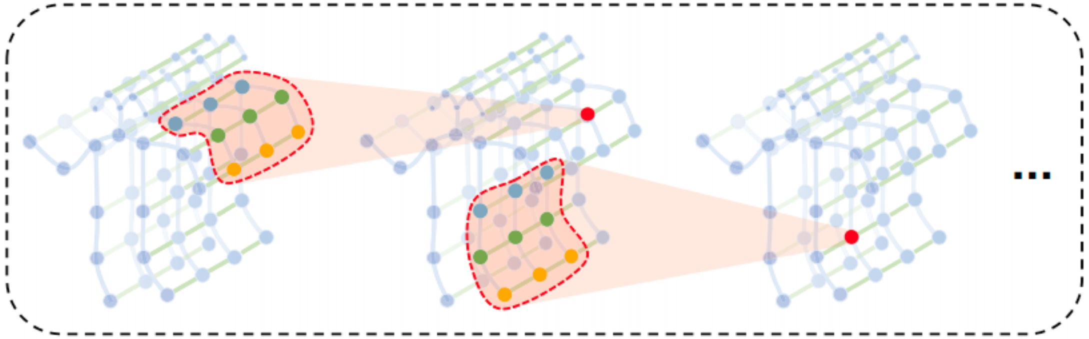
How can we develop meaningful gesture representations while considering their variability and relationship with speech?

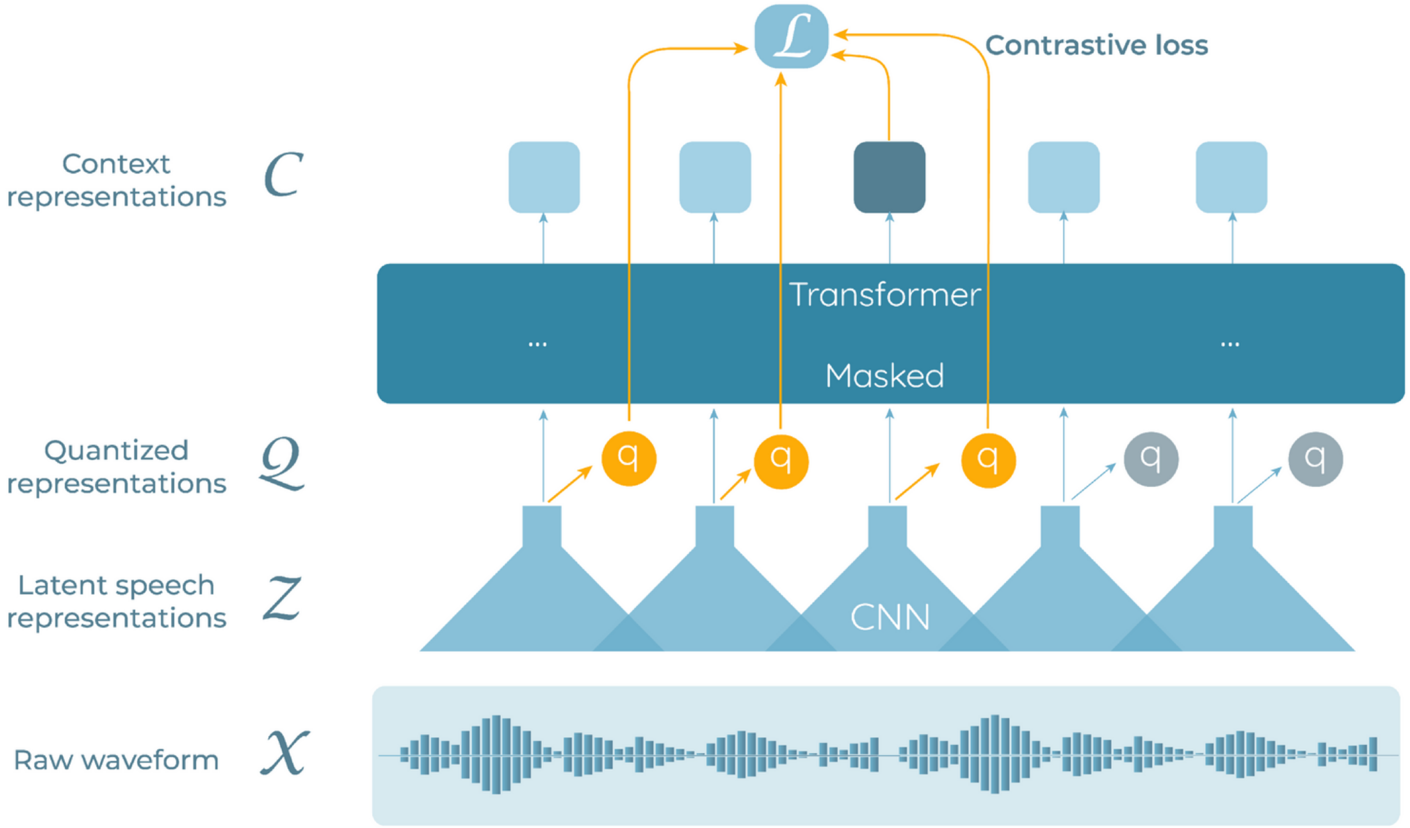
Evaluations against human experts annotations

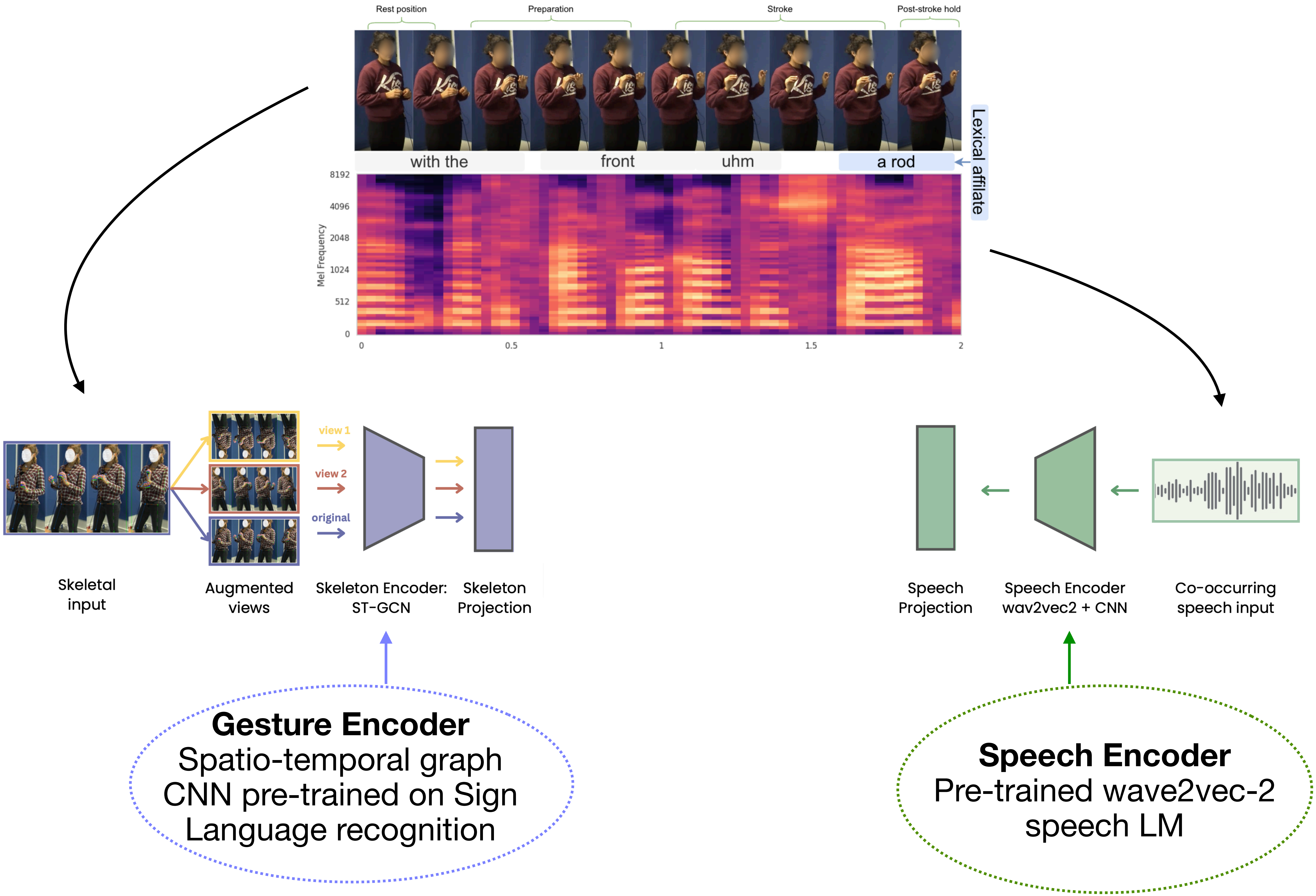
How aligned are the learned representations with expert annotations and patterns related to the dynamics of dialogue interaction?

Vision and Speech Models

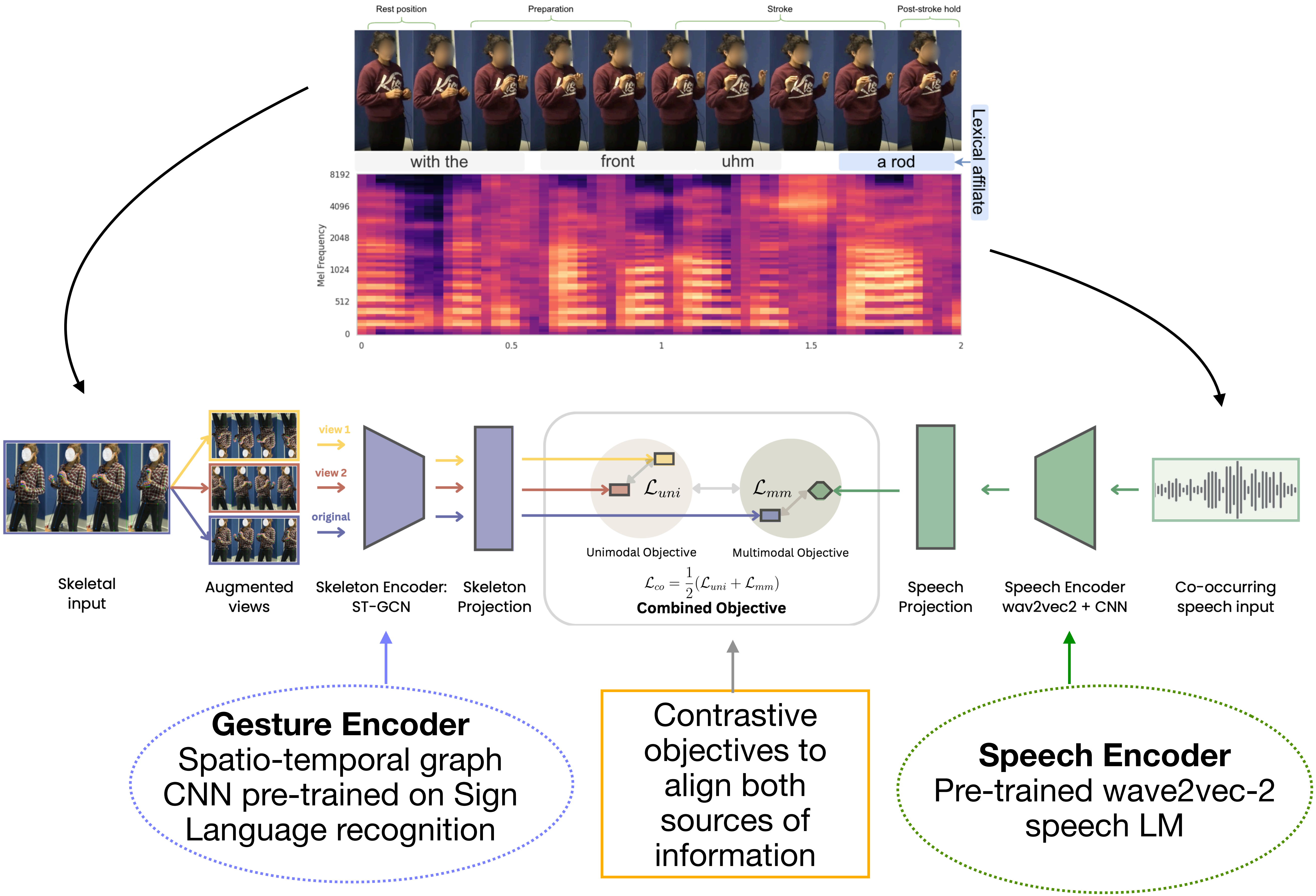








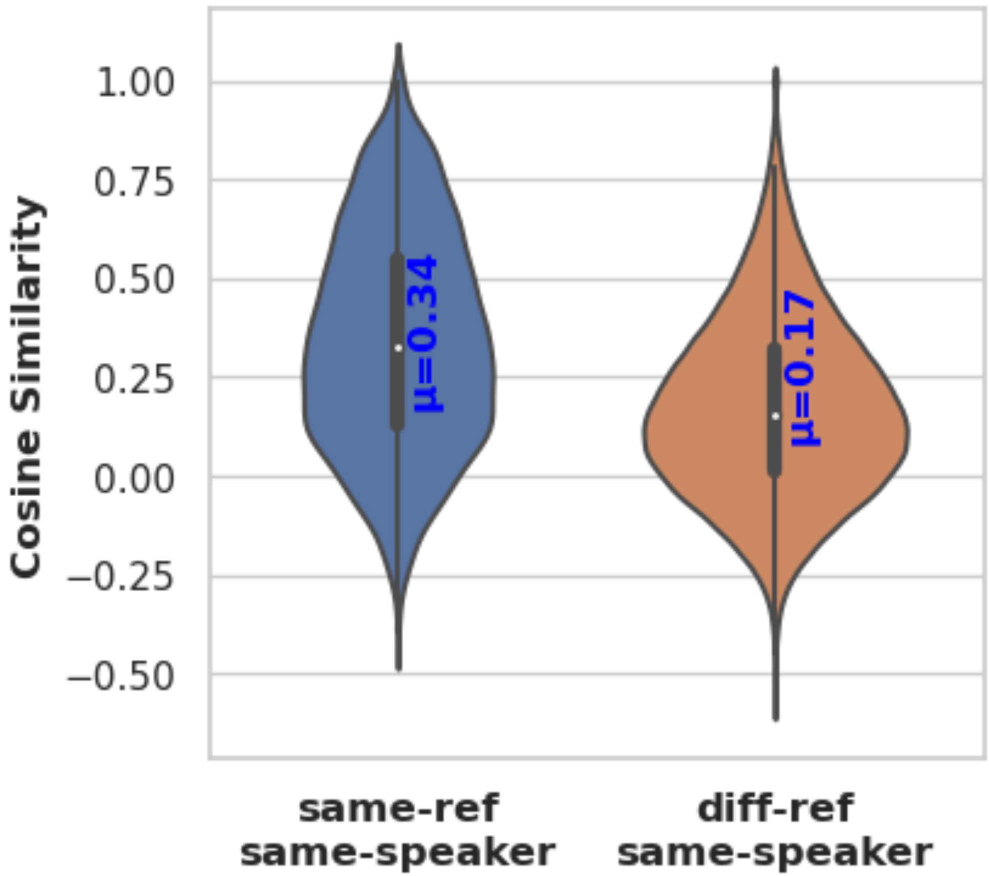
Rest position
Preparation
Stroke
Post-stroke hold
with the
front
uhm
a rod
Lexical affiliate
Mel Frequency
8192
4096
2048
1024
512
0
0
0.5
1
1.5
2
view 1
view 2
original
Skeletal input
Augmented views
Skeleton Encoder: ST-GCN
Skeleton Projection
Speech Projection
Speech Encoder wav2vec2 + CNN
Co-occurring speech input
Gesture Encoder
Spatio-temporal graph CNN pre-trained on Sign Language recognition
Speech Encoder
Pre-trained wave2vec-2 speech LM



Rest position
Preparation
Stroke
Post-stroke hold
with the
front
uhm
a rod
Lexical affiliate
Mel Frequency
8192
4096
2048
1024
512
0
0
0.5
1
1.5
2
Skeletal input
view 1
view 2
original
Augmented views
Skeleton Encoder: ST-GCN
Skeleton Projection
$\mathcal{L}_{uni}$
$\mathcal{L}_{mm}$
Unimodal Objective
Multimodal Objective
$\mathcal{L}_{co} = \frac{1}{2}(\mathcal{L}_{uni} + \mathcal{L}_{mm})$
Combined Objective
Speech Projection
Speech Encoder wav2vec2 + CNN
Co-occurring speech input
Gesture Encoder
Spatio-temporal graph CNN pre-trained on Sign Language recognition
Contrastive objectives to align both sources of information
Speech Encoder
Pre-trained wave2vec-2 speech LM

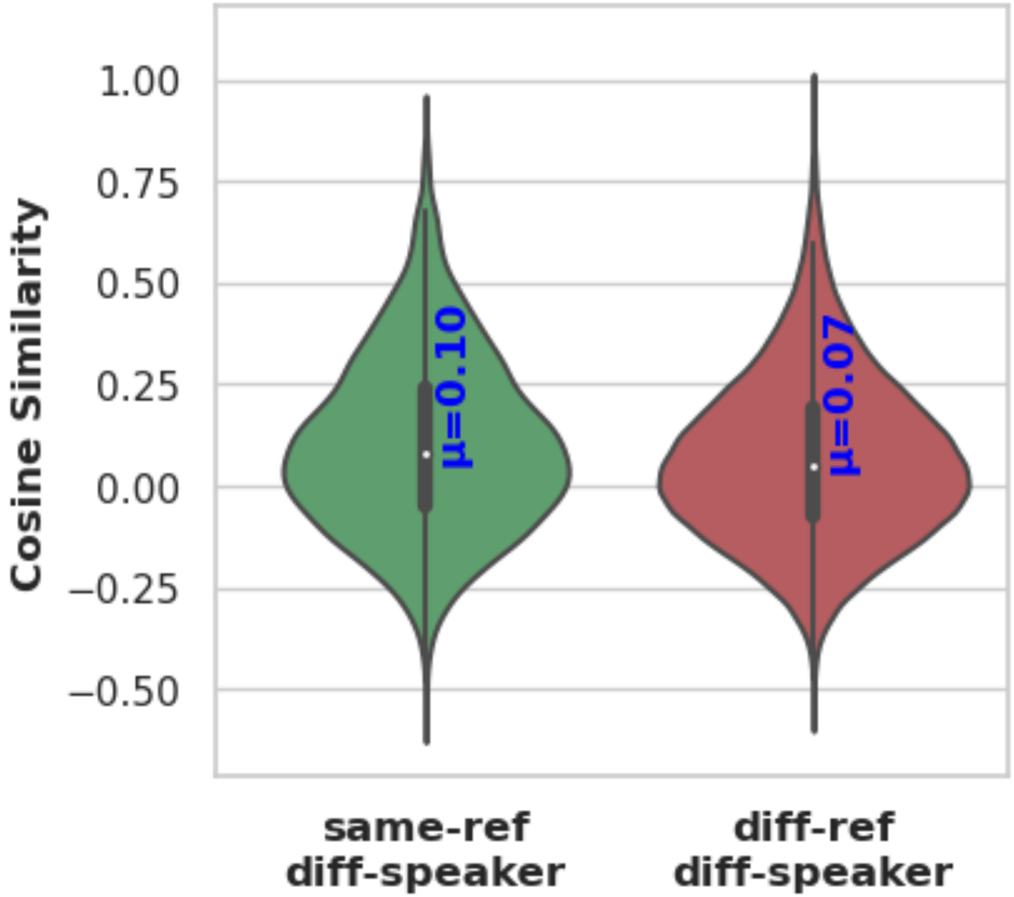
Gesture Similarity in Referential Dialogues

- Hypothesis 1:
  - Representations of gestures by the **same speaker** will be more similar if the gestures have the same referent than if they refer to different objects.



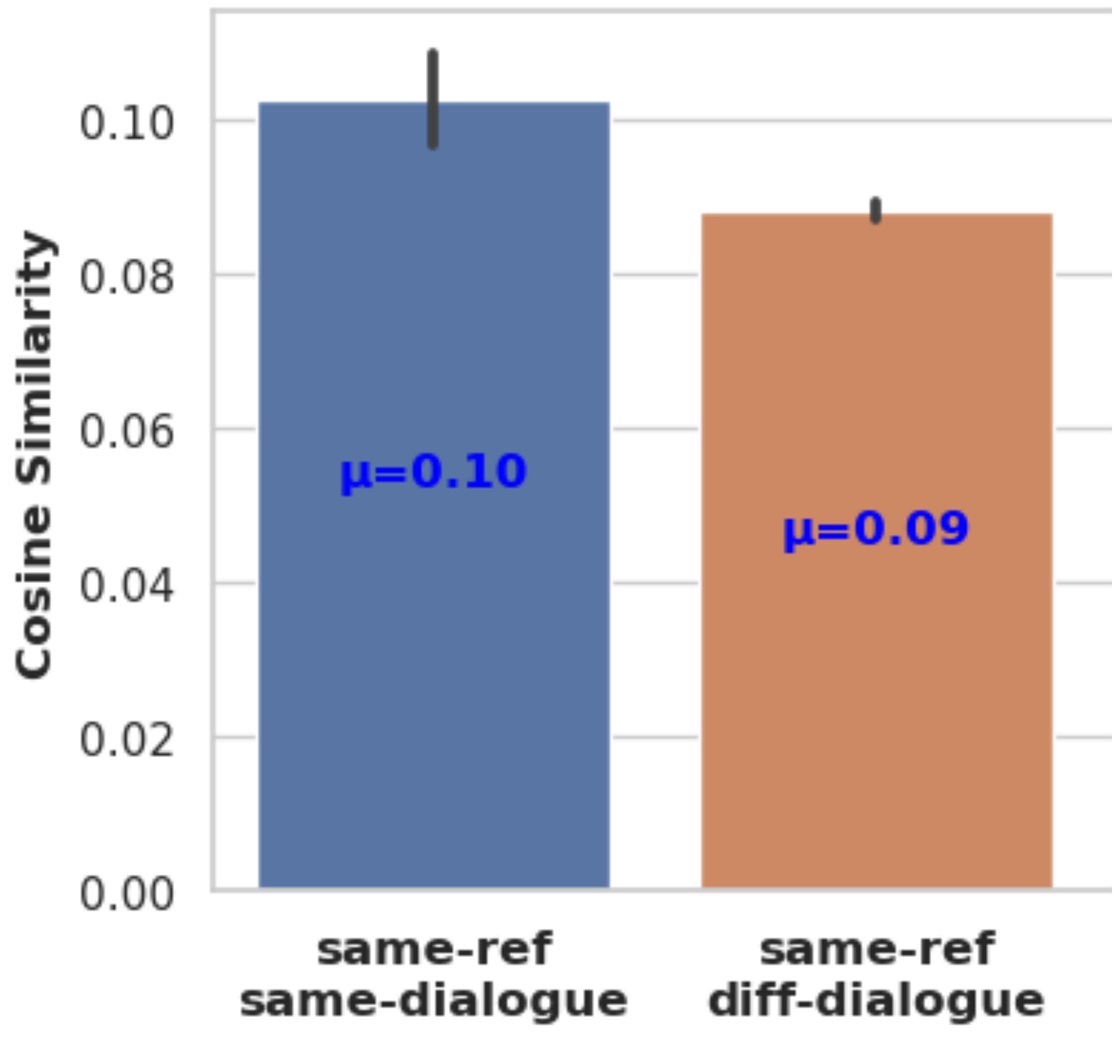
Gesture Similarity in Referential Dialogues

- Hypothesis 2:
    - Representations of gestures made by **different speakers** will be more similar if the gestures have the **same referent** than if they refer to different objects.



Gesture Similarity in Referential Dialogues

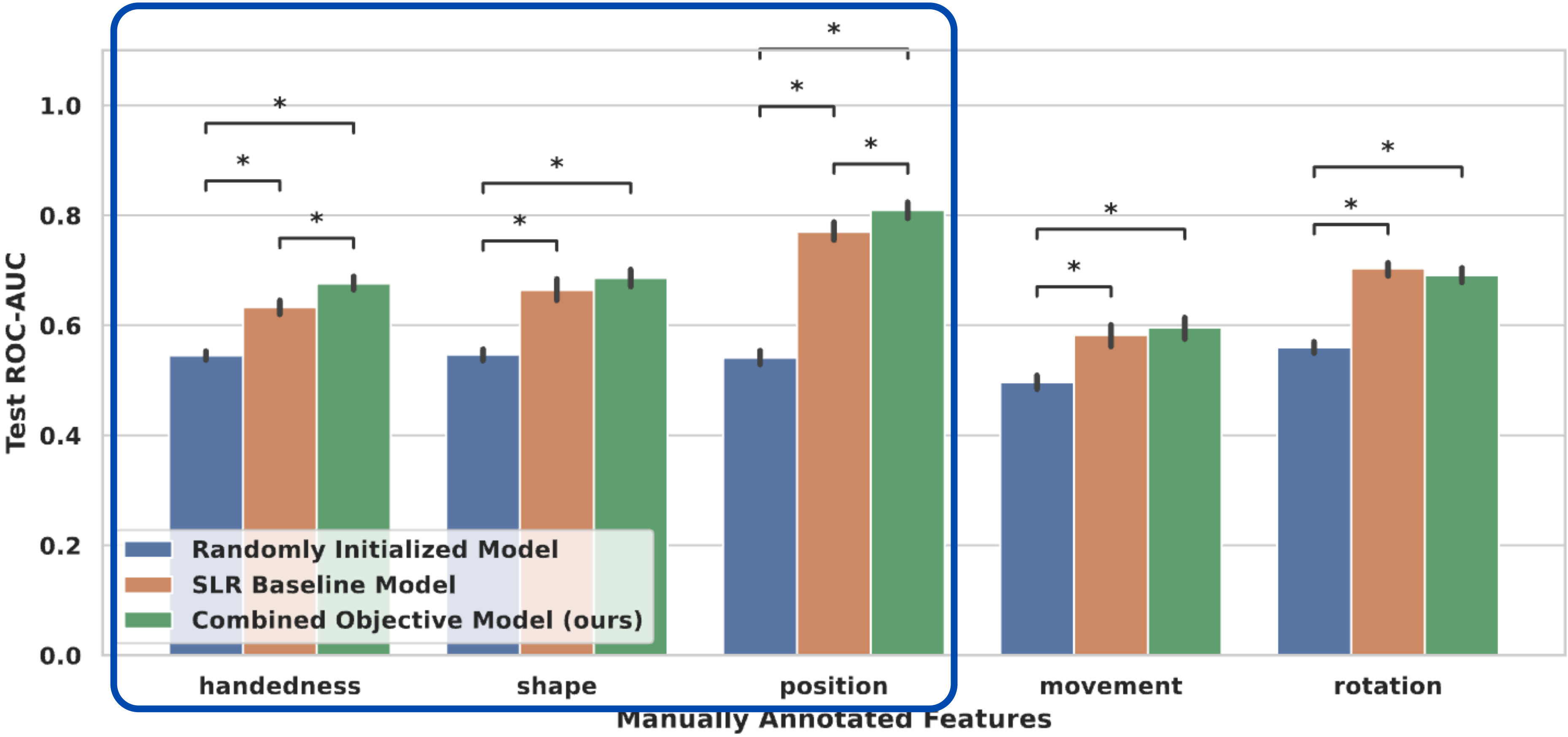
- Hypothesis 3:
  - Representations of gestures by **different speakers** will be more similar when the two speakers are interlocutors **within a dialogue** than when the speakers are from different dialogues



Probing Analysis

- To what extent the latent representations may encode interpretable features?

- Form features, particularly handedness and position, can more accurately be decoded from the latent representations learned by combined objective model



Plan for today

Part 1:

- What is dialogue modelling?
- NLP methods to model text-based dialogue
  - Modular statistical approaches
  - End-to-end encoder-decoder models
  - Generative (decoder-only) large language models

Part 2:

- Face-to-face dialogue
  - Modelling speech and gestures

Thanks!

Feel free to contact us

Amsterdam's
Dialogue Modelling Group

https://dmg-illc.github.io/dmg/