A Comparative Study of Large Language Models for Gesture Selection in Virtual Agents

Parisa Ghanad Torshizi

ghanadtorshizi.p@northeastern.edu

Northeastern University

Laura B. Hensel

University of Glasgow

Ari Shapiro

Flawless

Stacy C. Marsella

Northeastern University

Research Article

Keywords: gesture selection, virtual agents, large language models, multimodal behavior, embodied agents

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8768097/v1

Additional Declarations: No competing interests reported.

A Comparative Study of Large Language Models for Gesture Selection in Virtual Agents

Parisa Ghanad Torshizi[1*], Laura B. Hensel[2], Ari Shapiro[3], Stacy C. Marsella[1]

[1*]Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, United States.
[2]School of Psychology & Neuroscience, University of Glasgow, Glasgow, United Kingdom.
[3]Flawless, Los Angeles, California, United States.

*Corresponding author(s). E-mail(s): ghanadtorshizi.p@northeastern.edu;
Contributing authors: laura.b.hensel@gmail.com; ariyshapiro@gmail.com; s.marsella@northeastern.edu;

Abstract

Co-speech gestures convey a wide variety of meanings and play an important role in face-to-face human interaction, influencing addressees' engagement, recall, comprehension, and attitudes toward speaker. Similar effects have been observed in interactions with embodied virtual agents, making selection and animation of meaningful gestures a key focus in agent design. Automating this gesture selection process, however, remains challenging. Prior approaches range from fully data-driven techniques—which often struggle to produce contextually meaningful gestures—to more manual methods that rely on handcrafted expertise and lack generalizability. In this paper, we leverage semantic capabilities of Large Language Models (LLMs) to realize an automated gesture selection. We first illustrate information on gestures encoded into LLM, using GPT4 as a high-quality baseline. Building on this, we evaluate alternative prompting approaches for their ability to select meaningful, contextually appropriate gestures aligned to co-speech utterance. While GPT-4 demonstrates strong gesture selection performance, its inference latency makes it unsuitable for real-time interaction. To assess feasibility under real-time constraints, we then evaluate two additional models relative to this baseline: locally deployable Llama3 and a new high-capacity cloud model (GPT-5.2). Our comparative analysis highlights differences in gesture appropriateness and alignment, including GPT5.2's tendency

to over-generate gestures by selecting multiple, inappropriate gestures for a single utterance. In contrast, Llama3 achieves a more favorable balance between gesture quality and inference speed, informing model selection for interactive virtual humans. Finally, we demonstrate how this approach is integrated into a virtual agent system, enabling automated gesture selection and animation during human–agent interaction.

Keywords: gesture selection, virtual agents, large language models, multimodal behavior, embodied agents

1 Introduction

People now regularly interact with embodied facsimiles of people. Graphics-based virtual humans and social robots with anthropomorphic features and behaviors engage users using the same verbal and non-verbal behaviors that people use when interacting with each other. These technologies exploit the fact that the nonverbal behaviors of participants powerfully influence face-to-face interaction.

In this paper, we focus on an integral part of such social interactions: co-speech gestures. Gestures convey a range of meanings and have a powerful impact on face-to-face interaction [1–4], impacting the speaker's persuasiveness as well as an addressee's comprehension, recall, engagement, and trust in the speaker [5–7].

However, these impacts are dependent on the particular gestures being used and the context in which they occur [e.g., 8]. Clinicians, politicians, and comedians use different gestures because they seek to achieve different goals in different contexts [9–11]. Beyond these differences in contexts and goals, there are also considerable cultural and individual differences in the use of gestures [12, 13].

This richness makes gesture selection and animation a fundamental challenge in embodied agent research. To solve this challenge, researchers have focused on various automated approaches to select and generate virtual human gestures, often relying on data-driven or analysis-driven approaches [for an overview, see 14, 15].

With any approach, there are design factors that are critical to the realization of the virtual agent's gestures. Among these are requirements for the gestures to effectively convey the speaker's communicative intent. Further, we assume the virtual human should use gestures consistent with the context of the interaction between agent and human, the agent's role and the agent's personality that the designer intends to convey.

Thus, a virtual human who is taking on the role of a labor union executive should not gesture like a comedian or clinician, nor should it use gestures based on some statistical average over differing roles and situations.

In this work, we thus assume one is designing a virtual human to perform a specific role in some application. Leveraging recent work on exploring the ability of Large Language Models (LLMs) to predict gestures [16], we investigate the use of LLMs in gesture selection with an additional focus on implementing this approach within a virtual human architecture. We begin with standard GPT-4 and then we extend our

evaluation beyond GPT-4 and compare it with two models that represent complementary deployment and capability profiles: LLAMA3, a locally deployable model better suited for low-latency inference, and GPT-5.2 (the instant version), a newer high-capacity model. This analysis provides insight into how model capacity and inference behavior interact with embodied communication requirements, including the tendency of LLMs to over-generate gestures, which can reduce perceived realism, distract, alter emphasis transforming interpretation, and consequently degrade the interaction if left unconstrained.

What makes LLMs particularly promising as a component in gesture selection is that they are trained on very large and highly varied corpora. As we demonstrate, that leads to LLMs encoding representations of both the linguistic properties critical to conveying meaning in gesture performances, as well as the relation of those phenomena to actual gestures. Furthermore, we explore and evaluate GPT4' ability to represent rheme and theme [4, 17], metaphor [18, 19]and image schemas [20] and rhetorical structures [21]. We assess these abilities in the context of crafting a specific individual, a skilled presenter from a labor organization. The gesture selection approach is also implemented into an embodied virtual agent architecture allowing automation of gesture selection.
Our contributions are as follows:

- We investigate the capabilities of LLMs for gesture selection, addressing both *what* gestures to select and *when* to gesture, and comparing model choices against an annotated corpus of human gestures.
- We systematically evaluate multiple prompting approaches for LLM-based gesture selection, providing the model with incremental information on gestures and speaker examples.
- We conduct an expert evaluation in which gesture researchers assess the appropriateness of the selected gestures relative to the original annotated performances, allowing a direct comparison of prompting strategies, and models.
- We compare gesture selection quality and real-time performance trade-offs across GPT-4, LLAMA3, and GPT-5.2, underlying the impact of model choice on appropriateness, alignment and inference time of selected gestures.
- We develop an automated pipeline that maps LLM outputs to the behavior specifications required by a character animation system, enabling selection of concrete gesture animations from high-level model predictions.
- We integrate this gesture selection system into an embodied virtual human framework, demonstrating end-to-end automation of gesture selection and animation for a virtual agent.

2 Background

Research on gestures and their generation in virtual humans has a rich history, with gesture studies spanning millennia (e.g., [22]) and virtual gesture generation predating the turn of the century (e.g., [23]). In this section, we touch on some key aspects relevant to how we approached the problem of gesture generation. Research in human co-speech gestures has identified different categorizations of gestures, related to the

kinds of information they convey. For example, the work of McNeil (e.g., [4]) identifies four gesture categories—deictic, beat, metaphoric, and iconic. This work will largely focus on metaphoric gestures for two primary reasons: First, our research often focuses on skilled presenters, and such professional speakers tend to use these gestures frequently [24]. Second, metaphoric gestures can require deep analysis of the utterance to uncover relations to physical imagery, called image schemas, that underlie metaphoric gestures. Uncovering these relations presents a challenge for automated gesture selection.

Specifically, image schemas [25] are recurring spatiotemporal relationships grounded in our body's interaction in the world that are argued to underlie common patterns of physical and abstract reasoning, as well as motivate linguistic metaphors and metaphoric gestures. Consider this example of the image schema of container from our data set: "with workers and employers putting a little bit of contribution from each into a trust fund". Note container is not only the collection, the "trust fund", but there is also a pattern of reasoning associated with container, putting things into the container. The speaker used a pattern of metaphoric gestures associated with container, a gesture denoting the container, along with gestures depicting placing things into it.

Ideational Units: The above container example leads to an even more challenging issue. Gestures also occur in sequences where the individual gestures convey interrelated meanings. These gestural sequences are called ideational units [2] or gesture units [1]. This coupling plays important demarcative functions as well as helping to convey related meanings. These ideational units can be associated with image schemas as above but they can also establish relations between parts of an utterance, giving them a rhetorical structure. Consider a communicative intent to convey a rhetorical contrast between two abstract concepts such as strongly different political views. Metaphorically, abstract concepts can be viewed as physical objects, with physical properties such as locations and physical distance conveying conceptual differences. Thus the rhetorical contrast between political views can be conveyed metaphorically by abstract deictic gestures that convey this separation by pointing to disparate regions in space. Such gestures pose difficult challenges for gesture animation because they set up the physical space of multiple gestures in a consistent fashion that relates physical space and motion to meaning as the above two examples illustrate.

2.1 Gesture Selection

The process of automating the selection of co-speech gestures for an utterance can be broken into two tasks. The first task is determining which segments of an utterance have co-speech gestures. Since gestures emphasize as well as convey meaning in relation to parts of the utterance, identifying these segments is fundamental. We characterize this as *when to gesture*. Then there is the question of *what gesture to use*. In other words, what the gesture should convey, given the utterance and the context of the interaction.

When to gesture: Gestures are tied to what the speaker seeks to emphasize [26], which in turn transforms the meaning conveyed. Therefore, to determine when to gesture, it is critical for co-speech gesture generation to derive emphasis information

about the utterance. A common approach is to use prosodic cues [27–29]. However, this presumes that the spoken utterance that is driving gesture generation includes prosodic cues appropriate for the meaning the speaker seeks to convey. An alternative is a discourse analysis [30] that can identify the *theme* or topic of a sentence and *rheme* or focus of the sentence, which is what is being said about the topic. The rheme provides new information about the topic and, critically, tends to be associated with gesture co-occurence [4, 31].

Gesture generation systems have used rheme analysis [32] as well as prosodic analysis [29] to determine when to gesture. Assuming we choose to use rheme analysis to automate when to gesture, the question becomes how to quickly and efficiently do the analysis. In the context of the current work, we can specifically ask whether LLMs can do this kind of discourse analysis as part of a gesture selection process.

What gesture to use: There is a considerable variation in gesture types and usage frequency across individuals and contexts [2, 24, 33]. However, a key focus in this work has been on exploring the use of metaphoric gestures often used in the effective public speakers we have studied. This focus provides the impetus for exploring the use of LLMs to undertake the necessary analysis tied to the uses of these gestures. As just one example of metaphoric gestures, consider again the linguistic metaphor that abstract concepts can be physical objects [18]. All the physical properties and actions one takes on physical objects can be used to convey abstract meanings gesturally. Thus, concepts, as physical objects, can be gesturally grasped (understood), they can be thrown away (rejected), and they can be big (important). Two concepts, such as political orientation, as physical objects can have separate locations to contrast them or convey their differences. Additionally, sets of abstract concepts can have cardinality (large or small) and openness(closed or open sets). Whether a set represent abstract concepts or physical objects, the abstract operations of set theory are evidenced in gestural forms, such as adding elements, deleting, and union of sets.

This richness is not surprising given the richness of metaphors represented in our language [18] and argued to be integral to our thought processes [34]. It also represents a central challenge to approaches to gesture generation.

3 Related Work

As noted earlier, a key challenge in designing embodied conversational agents involves selecting appropriate gestures for the character, with various approaches proposed to address this challenge [35].

Rule-based methods [31, 36] rely on predefined sets of rules that select from a predefined set of gesture animations. For instance, the Behavior Expression Animated Toolkit (BEAT) is a rule-based system that generates body movements and vocal intonations based on linguistic and contextual features of the text [31].

Mixed approaches [29, 37, 38] integrate machine learning and ontologies to perform syntactic, semantic, and prosodic analysis of the utterance to infer communicative intent and then use handcrafted knowledge to map that intent to non-verbal behaviors such as head movements, gestures, and gaze.

Although these rule-based and mixed systems offer designer control over gesture selection, they are limited to a fixed relationship between the properties of the speech they infer and gestures. Essentially, there is a trade-off between the designer's more explicit control over a character's use of gesture and both the burden placed on the developer or designer and the flexibility of the approach.

Data-driven approaches aim to learn a mapping between utterances and gestural motion from annotated corpora of gestural performances. Work in this area relies on, for example, deep learning techniques. A recent review of this deep learning work [15] has identified several key limitations, including a lack of designer control over the performance and the limited ability of current approaches to realize semantically meaningful gestures such as metaphoric gestures.

Most recent works on gesture generation have focused on using diffusion-based generative models and LLMs for the gesture generation process. DiffMotion [39] combines an LSTM with a diffusion model to generate gestures. DiM-Gesture [40] utilizes a Mamba-2 diffusion architecture to enable the creation of personalized full-body gestures. ConvoFusion [41], [42] present diffusion-based approach for multi-modal gesture synthesis. [43], presents DiffuseStyleGesture, a based speech-driven gesture generation approach that utilizes cross-local attention and self-attention in the diffusion pipeline. [44]and [45] present frameworks for direct generation of audio-driven gesture videos. [46] introduces SIGGesture; a diffusion-based framework that combines rhythmical gesture generation with LLM-driven semantic augmentation. [47] introduces a framework that leverages LLMs to interpret speech and motion jointly, enabling controlled co-speech gesture synthesis. [48] leverages LLMs for a controllable stylized co-speech synthesis.

These alternatives are solving different problems in some respects. Rule-based approaches continue to be used in applications, such as health applications, where there are significant ethical concerns and the designer seeks to retain control over the interaction, whereas machine learning-based approaches allow for more general, natural-looking behavior but with less control over what the nonverbal behavior conveys.

3.1 Use of LLMs

In this paper, we explore a different approach. Unlike an end-to-end approach that maps from utterance to animation, we explore the use of LLMs for gesture selection. We therefore go beyond traditional mixed and rule-based approaches in actually proposing gestures based on a more thorough analysis of the utterance by the LLM but stop short of end-to-end machine learning-based approaches that also generate the animation. This is both a limitation and a potential benefit in that it allows for carefully crafted gestures that effectively convey desired communicative intent.

In addition, LLMs encode massive amounts of information across domains. Unlike the simpler word-based ontologies like Wordnet used previously in mixed approaches, or even the data used to train end-to-end approaches, the far more extensive data used to train LLMs allows them to encode contextual, semantic, and structural relations at the phrase, utterance and situational levels. LLMs also encode extensive amounts

of information across scientific domains. GPT4, when asked, can define rheme and theme, metaphors, image schema, and metaphoric gestures. As one example:

```
GPT4: Image schemas are recurring, dynamic patterns
of our perceptual interactions and motor programs ...
basic cognitive structures that arise from bodily
interactions with the world ... are understood to be
one of the fundamental building blocks of human thought

GPT4: Container:|definition|, Path:|definition|,
Source-Path-Goal:|definition|, Force:|definition|...
```

LLMs also have representations of different gestural forms and styles used by speakers in different roles. They have for example knowledge of clinicians, their role, and their goals, and can even capture the variations in gestural styles among well-known individuals within these professions. Moreover, they can be prompted with examples that prime and integrate with their existing knowledge to improve their ability to select gestures. We asked GPT4 to report a repertoire of gestures that a clinician and a politician would use.

```
GPT4: Here's a repertoire of hand gestures that
clinicians often use:
Palms Open and Upward: Indicates openness and
transparency, used when explaining procedures or
discussing options.
Hand to Chest: Expresses sincerity and empathy,
particularly in sensitive conversations.
Enclosed Hands or Fingers Interlocked: Demonstrates
patience and focus, typically seen during attentive
listening.

GPT4: Here's a repertoire of hand gestures that
politicians often use:
Fist Pumping: Used to express determination or
victory, energizing the audience.
Thumb and Index Finger Pinch: Known as the
"precision grip," it emphasizes specific and
important points, suggesting control.
Hand Chopping: Aligns with making strong,
decisive points, conveying clarity and decisiveness.
```

Detailing the extent of all these representations is beyond the scope of the current article. Rather, the focus of this article is to lay out an integrated approach to using LLMs to automate gesture selection and evaluate the ability to determine *when to gesture* and *what gestures to use*. Our approach builds on the work of Hensel et al. [16]

but goes further in a) exploring different prompting techniques, b) having the LLM also determine when to gesture, c) evaluating fast inferencing approaches to support real-time interaction, d) developing a prompting approach that allows the automated mapping of LLM output to a behavior specification for a character animation system, and finally e) implementing the approaches. The implementation takes utterance text as input, generates the utterance audio, and drives a virtual human's spoken dialog and gesturing.

4 Research Questions

To explore an LLM-based gesture generation framework, we lay out a set of research questions relating to gesture generation more broadly and the selection of individual gestures more specifically:

RQ1: Selection of Gestures using GPT-4. What are the impacts of alternative prompting approaches on the appropriateness of gesture selection and the speed of inference in GPT-4?

- *Appropriateness* How appropriate are the selected gestures, with regards to the context of the speech and the speaker?
- *Speed of inference* Can GPT-4 select gestures with a speed sufficient to enable real-time inference in animation systems?

RQ2: Selection of when to gesture using GPT-4. Does the rheme, identified through GPT-4-enabled rheme and theme analysis, correspond accurately to the actual gestural timings of the speaker in the data?

RQ3: Selection of gestures using faster models. What are the impacts of prompting on the appropriateness of gesture selection and the speed of inference in Llama3 and GPT5.2 instant version?

5 Analysis

To answer our research questions, we first annotated hand gestures in a specific speech. We then used these annotations as ground truths to evaluate gestures suggested by language models on the same utterances.

5.1 Data

To derive a set of ground truth gestures and utterances, we selected a video featuring Elizabeth Shuler, a labor activist, speaking at the Working Families Summit, which is available through the National Archives and Records Administration.[1] To collate Elizabeth Schuler's gesture repertoire, we annotated three segments of the video in which she was the primary speaker, resulting in a total of six minutes of annotated video. We then split these annotations into a training set (21 utterances) and a test set (20 utterances).

[1] https://youtu.be/-6NA1xl32uY?si=lOjTOkyIrQhgnsKJ

5.2 Classes of Gestures: Gestural Intents

Based on the training data set, we created a list of gestural classes based on image schemas and metaphors that were apparent in the speaker's performance. Each of these classes conveys a particular intent and we thus refer to these as *gestural intents*. Note that alternative gestures with different physical properties can realize one of these gestural intents. For example, the speaker may convey the gestural intent of progress either through a forward circling of the hands or a sweep showing the direction of progress. Below, we list the speaker's repertoire of *gestural intents*, including brief explanations.

- Progress: This gesture represents progress, advancement, or moving forward. It is part of the path group image schema.
- Regress: This gesture represents moving backward, regressing, or returning to a previous point. It is part of the path group image schema.
- Cycle: This gesture represents actions or processes that repeat in a continuous loop or follow a recurring pattern. It is an image schema.
- Collect: This gesture represents gathering, collecting, or bringing things together, into one entity. It is an image schema.
- Container: This gesture represents a boundary, a sweep, or an imaginary box holding a collection of items. This is an image schema and basis for the container metaphoric gesture.
- Oscillation: This gesture represents alternation, uncertainty, indecision, or items being out of balance. It is part of the balance group image schema.
- Temporal: There are many, culture-specific time metaphors. Here we refer to representing time as a line, with different points on the line representing past, present, and future.

5.3 Analysis on GPT4

As the baseline model, we chose GPT-4 since its performance is shown to be superior to earlier models like GPT-3 or GPT-3.5 [16]. We used the GPT-4 Chat Completion API as it is faster and more versatile in terms of parameter settings, with the following parameters: (Temperature: 0.2, Max tokens: 256, Frequency penalty: 0). The transcribed speech was split into utterances, using end of sentence punctuations.

5.3.1 Approach: What Gesture to Use

To evaluate the selection of gestures, we designed a 2X2 factorial design of alternative prompting approaches: *gestural intents repertoire explained/not explained* **X** *annotated examples given/not given*. For each of those approaches, we asked GPT-4 to produce gestures based on the utterances in the test set. Below is a detailed description of each approach.

Approach 0: In this approach, the model is not prompted with any prior information, on gestural intents or annotations. The model is just asked to report a gesture for each of the utterances, describe the physical properties of its suggested gesture, and the specific phrase in that utterance that the gesture is trying to illustrate.

Approach 1: In this approach, the model is only prompted with the above-mentioned list of gestural intents, and their descriptions. The model is then asked to report the gestural intents in the utterance, suggest a gesture for each gestural intent, the physical properties of that gesture, and the associated phrase.
Approach 2: In this approach, the model is prompted with only the annotations from the training set. The model is then asked to report the gestural intent, suggest a gesture for that gestural intent, the physical properties of that gesture, and the associated phrase.
Approach 3: In this approach, the model is prompted with the above-mentioned list of gestural intents, and the annotations from the training set. The model is then asked to report the gestural intent, suggest a gesture for that gestural intent, the physical properties of that gesture, and the associated phrase.

5.3.2 Results: What Gesture to Use

This section details the results of our experiments on GPT-4's ability to select appropriate gestures to be used by the virtual human. In this section, we investigated our RQ1, examining the appropriateness of gestures selected by alternative prompting approaches and comparing their speed of inference.

Appropriateness

To assess the semantic appropriateness of the gestures selected by the model, two experts jointly evaluated each proposed gesture for every utterance in the test set across all four approaches. Evaluators made their judgments for each gesture and then immediately compared them. If there was disagreement, the two discussed where this disagreement arose and resolved it through further discussion. It should be noted, however, that there was very little disagreement between the two experts. In order to do the evaluation, we divided all the selected gestures by GPT-4 into the following two categories.
Category 1:
This category consists of gestures that the model generates for specific parts of an utterance, where there is a corresponding gesture on that (or close) part of the utterance in the actual speech (i.e., ground truth is available). Experts have assigned one of the following tags to these gestures:

- **Appropriate**: The gesture does convey the semantic meaning that the speaker is trying to convey. Regarding the definition of appropriateness, we defined this as gestures that are not only appropriate for the utterance (i.e., 'fit' with the intended meaning) but also for the context of the speaker.
- **Inappropriate**: The gesture does not convey the semantic meaning that the speaker is trying to convey.
- **No corresponding gesture**: The speaker produced a gesture but the model did not propose a gesture at the corresponding part of the utterance.

Figure 1a shows the number of gestures that belong to each label across the proposed prompting approaches. The results indicate that providing more information to

the prompts yields more appropriate gestures and less inappropriate gestures. Specifically, approach 3 is better than approaches 1 and 2, and approaches 1,2, and 3 are better than approach 0. Moreover, providing the model with examples (approach 2) is slightly more effective in selecting more appropriate gestures compared to providing the model with solely the explanation of the gestures (approach 1).

Category 2:

This category consists of gestures that the model generates for specific parts of an utterance, yet there is no corresponding gesture on that (or close) part of the utterance in the actual speech (i.e., ground truth is not available). The experts labeled each gesture with either the appropriate or the inappropriate tag.

The results in Figure 1b suggest that approach 1 has the highest number of appropriate gestures and the lowest percentage of inappropriate gestures.

Combining all the gestures in the two categories listed above, we analyzed the overall appropriateness, regardless of whether a corresponding ground truth was present or not. The results, shown in Figure 1c, indicate that there is a slight difference between approaches 1,2 and 3, but they all outperform approach 0. Additionally, the performance of approach 1, where the model was not prompted with any annotations, is promising. This suggests that LLMs can minimize the need for extensive annotations.

Furthermore, we analyzed how often the gestures generated by GPT-4, regardless of their appropriateness, were aligned with the gestures performed by the speaker, meaning how often the speaker and GPT-4 gesture on the same phrase. Figure 2a and 2b depicts the frequency of gestures being aligned, or not aligned. In not-aligned cases, GPT-4 either produced gestures where the speaker did not gesture or the speaker gestured but GPT-4 did not produce a gesture. Approach 2 has a higher alignment of gestures between GPT-4 and the speaker. On the other hand, approach 0 produced the highest misalignment. These results suggest that limiting the gesture classes to the speaker's gesture repertoire helps to constrain gesture selection to produce more accurate gesture timings. In conclusion, the approaches that prompted the model with a defined set of gestural intents generated more appropriate gestures and less inappropriate gestures. Moreover, they helped in selecting gestures that were happening concurrently with the speaker's gestures and thus were more aligned.

Inference time

As mentioned earlier, a potential use case for LLMs in gesture selection is to suggest gestures in real-time. Therefore, we explored the inference times of gesture selection across different prompting approaches. In this work, the inference times are actually recorded as network latencies, to evaluate real-time suitability for human-agent interactions. In these approaches, where GPT-4 was queried to report gesture name, its properties, gesture description, and the associated phrase; the average inference time per utterance across different approaches ranges from 6.51 to 9.29 seconds. Specifically, approach 0 showed the highest inference time, and approach 1 had the lowest inference time. However, for the purpose of animating gestures in the virtual agent, since the animations are already defined, we truncated the query into asking only the gesture type and the associated phrase. With a target inference time of less than one

second for online, real-time, interactions, our next goal is to adapt a smaller, faster model that can achieve this speed.

5.3.3 Approach: When to Gesture

We used GPT-4 driven discourse analysis to identify the rheme and theme of each utterance. We then explored whether the rheme and theme can serve as criteria to decide on when to gesture, or which part of the utterance is most likely to be accompanied by a gesture. Specifically, we prompted GPT-4 to identify the rheme and theme in each utterance. Such a rheme and theme analysis could be used on top of the gesture selection system, to prioritize which gesture/s to use in the animation system; with gestures associated with the rheme part of the utterance being given higher priority. To evaluate, we analyzed whether they correspond to the part of the utterance that the speaker gestured on.
To explore GPTs inference of rheme and theme and its use in selecting when to gesture, we ran our speaker's dialogues through it and asked it to report its rheme and theme. We then evaluated whether the identified rheme corresponded to the speaker's use of gestures.

5.3.4 Results: When to Gesture

To address RQ2, we evaluated how often the speaker's gestures occur within the rhemes identified by GPT-4. To do so, we considered all the gestures made by the speaker, including deictics, beats, metaphorics, and iconics; and counted the number of times these gestures happened within the rheme of the utterances identified by GPT-4.
The results show that out of 49 gestures made by the speaker, 43 of them occur within the identified rheme and 6 of them occur outside the identified rheme (they occurred on the identified theme). Among these six gestures, were one deictic, two beat, and three metaphoric gestures.

5.3.5 Discussion of GPT4 Results

In this section, we evaluated the use of LLMs to automate gesture selection, examining various prompting approaches that yield diverse results suitable for different applications. As to be expected, Approach 0, where the model was not prompted with any prior information, had the most flexibility in the gestures it recommended and could generate creative variations of gestures. However, this approach had the highest number of inappropriate gestures, the lowest appropriate gestures, and the lowest alignment with the speaker's gestural performance and gestural timings. One might imagine it being useful for suggesting gestures to a designer, as opposed to automating gesture selection. In other approaches, adding information to the prompts, including the explanation of gestural intents and examples from the speaker, increased the number of appropriate gestures and decreased the number of inappropriate gestures. Also, gestures tended to be more aligned with original speaker, with slight variations between approaches 1, 2 and 3.

Looking at the results provided by these different approaches, several qualitative distinctions were apparent. In our promptings, we asked the model to also provide a description of the physical properties of the gestures. In general, this was very surprising, as it represented deep relationships between meaning in the utterance and the physical motion, as argued by gesture researchers [2, 33].

In Approach 0, which lacked gestural intents or annotation examples, the model sometimes produced shallow responses, such as self-referential deictics triggered by words like 'I,' 'we, and 'our.' Additionally, this approach frequently generated inappropriate iconic gestures that detracted from speech. Approach 1, which was prompted with gestural intents and descriptions, had fewer gestures than one might expect. Also sometimes the gestural intent was off while the description of the physical motion was good. Approach 2, which was prompted with annotations, but no list or description of gestural intents, would come up with novel gestural intents not in the annotations. Looking from the perspective of building a virtual agent, the animation system needs to transform the gesture selection to a corresponding animation. We will discuss this further in Section6. This can require a mapping between gesture labels (intents) and a defined set of animations. Therefore, limiting the model's choice of gestures as was done is aligned with such an animation system. It is also a means to enforce a designer's intent on the virtual human's performance so that the performance is consistent with the virtual human's role and the interaction context. Consequently, we used approach 3 in our implementation. The generated gestures are in this case constrained to the gestural intents and annotations in our sample of the speakers' style of gesturing, which in some cases is a positive outcome in terms of personality/role consistency in behavior but in others may be viewed as limiting, especially compared to end-to-end machine learning based approaches to gesture animation. Though to be clear, what is limited here is the space of gestural types, not the language that triggers those gestures which is very general due to the use of an LLM. Given the fact that there is a many-to-many mapping between language and gesture, especially when considering an individual's gestural style, this arguably is not as limiting as it might seem. Additionally, our findings on the rheme and theme analysis suggest that this linguistic construct can be used to narrow down the selection of gestures from an LLM, focusing on those most likely to match the original speaker's actions.
We elaborate our further analysis on GPT4:

Rhetorical Structures

Because gestures are often organized into higher-level ideational units that align with rhetorical structures, we evaluate GPT4's ability to infer these structures from discourse as a prerequisite for coherent gesture selection.

In order to evaluate GPT4's ability to infer these rhetorical structures and their associated gestures, we identified three rhetorical structures apparent in the data, then prompted GPT4 with their definitions, and asked the model to identify the rhetorical structures within the utterances. Below is the list of these rhetorical relations:

- Elaboration: Providing additional details about an element previously introduced.
- Enumeration: Listing multiple elements or examples that all support or elaborate on a single concept.

• Contrast: Highlighting differences between two elements.

```
Prompt:
Rhetorical structure theory (RST) is a theory of
text organization that describes relations that
hold between parts of text.
Let's focus on these three rhetorical structures.
- Elaboration: Provides additional details about an
element previously introduced.
- Enumeration: Lists multiple elements or examples
that all support or elaborate on a single concept.
- Contrast: Highlights differences between two
elements.
Now, I will provide you with an utterance from a
social worker activist. For that utterance, please
identify the rhetorical relations. For each
rhetorical relation, report the phrase associated
with it, and suggest an appropriate gesture.

Utterance:
So, sometimes scheduling might not be necessarily
as much of an issue for some of these work-family
types of concepts than other issues, such as paid
leave, paid sick leave, paid family leave.

GPT4 Response:
Rhetorical Relation: Contrast
Phrase: scheduling might not be necessarily as much
of an issue ... than other issues
Explanation: This contrasts scheduling with other
work-family issues, emphasizing that scheduling is
less problematic in comparison.
Gesture: Two-handed lateral spread, alternating
emphasis (e.g., left hand outward for \scheduling,"
right hand outward for \other issues") to visually
represent the contrast.
Rhetorical Relation: Enumeration
Phrase: paid leave, paid sick leave, paid family
leave
Explanation: This is a list of related policies
that are all examples of the \other issues"
mentioned in the contrast.
Gesture: Rhythmic pointing or chopping motion, one
per listed item (e.g., downward palm gesture or
finger point for each: \paid leave" → \paid sick
```

```
leave" → \paid family leave").
```

Spatial Location of Gestures

When there are multiple gestures in an utterance, the spatial location of gestures, and connecting one to the other becomes an important factor in ensuring the consistency of the gestures and making sure the right information is being conveyed. For example, GPT-4 sometimes identified multiple gestures where motion of one gesture is meaningfully tied to the preceding gesture, such as a container gesture along with a gesture suggesting something being removed or placed in the container [49].

This raises the issue as to the capabilities of GPT-4 to specify spatial relations between gestures in ideational units. To automatically animate, it is necessary to identify where gestures begin and end in the gestural space. The following explores this issue through a prompt that specifically requests spatial information.

```
Prompt:
In the following utterance, please identify any
Image Schemas evoked by the utterance, the Phrase
of the utterance that comprise the image schema
and what gesture might be used to illustrate that
Phrase as well as where that gesture Spatially
Begins and  Spatially Ends, center, left or right.
Please provide the answer in the following format
[Type of Image Schema:]
[Phrase : Gesture : Spatially Begins: Spatially Ends]
[Phrase : Gesture : Spatially Begins: Spatially Ends]
[Phrase : Gesture : Spatially Begins: Spatially Ends]
Here is the utterance:
\For example, up in New York with SEIU-1199, you
know, coming together with workers and employers
putting a little bit of contribution from each
into a trust fund to provide for care for --
whether it's, you know, elderly or children or
whatever, to subsidize that."

GPT4 Response:
To analyze the utterance for image schemas and propose
gestures that illustrate the phrases, we can identify
several key elements in the text:
[Type of Image Schema: CONTAINER]
["with SEIU-1199" : Cupped hands gesture :
   Spatially Begins:Center : Spatially Ends:Center]
["into a trust fund" : Gesture of placing objects into a
   container (using both hands) :
   Spatially Begins:Left : Spatially Ends:Center]
```

```
[Type of Image Schema: PATH]
["coming together" : Gesture of moving hands toward
   each other :
   Spatially Begins:Left and Right : Spatially Ends:Center]
["putting a little bit of contribution" : Gesture mimicking
    placing or dropping something along a path :
    Spatially Begins:Left : Spatially Ends:Center]
[Type of Image Schema: SUPPORT]
["provide for care" : Hands held flat, palms up, as if
     supporting something :
     Spatially Begins:Center : Spatially Ends:Center]
These gestures and their spatial dynamics help to visually
represent the conceptual ideas expressed in the text.
```

Although above illustrations on GPT4s ability to infer rhetorical structures and spatially coordinating gestures are promising, more extensive testing is required.

5.4 Analysis on Fast Models

The final goal of this gesture selection system is to be integrated into a virtual human framework, where it is capable of interacting with the users in real-time. Achieving this goal requires an inference time small enough to be suitable for a real-time interaction. In this regard, we looked at three fast models: 1)GPT-OSS-20b, 2)Llama3, and 2)GPT5.2 (the instant version).

Local inference environment

All experiments with locally deployable models(GPT-OSS-20B and Llama3) were conducted on a desktop workstation equipped with a 12th Gen Intel(R) Core(TM) i9-12900KF CPU (3.19 GHz, 64-bit), 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU, running a 64-bit Windows operating system. The gpt-oss-20 and Llama 3 models were served using the Ollama runtime, with GPU acceleration enabled. Inference timings reported in this paper for the local models were obtained on this machine under a batch size of one, matching the real-time use case in our virtual human pipeline.
We had the following inference times for different models:

GPT-OSS-20b: max(s): 9.43 , min(s): 2.39 , avg(s):4.50
Llama3: max(s): 1.23 , min(s): 0.36 , avg(s): 0.61
GPT5.2(instant version): max(s): 2.00 , min(s): 0.61 , avg(s): 1.12

Given our focus on real-time interaction, we exclude GPT-OSS-20B from further evaluation due to its inference latency, which makes it unsuitable for low-latency deployment.

5.4.1 Approach: What Gesture to Use

Data and annotation labels

We evaluated the gesture generation capabilities of LLAMA, and compared it to two large language models GPT-4, GPT-5.2 (the instant version). The gestures selected by these models were evaluated against the data which was used for evaluation in section 5.1. In this data, the gestures used by the actual speaker are called gold standard. In the gold standard data, each utterance might be associated with zero gestures (in this case, we mark the gold standard as null), or one or more gestures (in this case, the gold standard is marked as nonnull). In this evaluations, the experts provided two judgments for each of the selected gestures by the model: (i) *matchness* and (ii) *appropriateness*. Matchness indicates whether the selected gesture is the exact same gesture as in the gold standard; a gesture can be labeled as match or non-match. Appropriateness indicates the appropriateness of the gesture; each gesture can be labeled as appropriate or inappropriate. A gesture is considered appropriate when it is both semantically appropriate and appropriate to the role of the speaker and the context of the speech.

In our annotation scheme, any gesture labeled as `match` is necessarily also `appropriate`, since it was used by the speaker herself. Accordingly, each predicted gesture falls into exactly one of three mutually exclusive categories:

1. **Match** (and therefore appropriate),
2. **Non-match & appropriate** (an appropriate alternative to the gold standard),
3. **Non-match & inappropriate**.

Prompting Technique

For these set of analysis, we picked Approach3 5.3.1 for prompting the language models, as it resulted in the highest appropriateness level. This prompt includes a description of gestural intents, along with examples of those gestural intents being used by the speaker. Then, in each call to the model, the model is provided with the utterance, and it is asked to return gestures along with their associated phrases in that utterance. During all the model calls on utterances, the language model has access to all the previous user and system messages, which includes previous utterances and the model's suggested gestures. The same prompting technique was used across all three models.

Aggregation levels

In the gold standard, an utterance may contain multiple gestures associated to different phrases in that utterance. Also, models may produce different numbers of gestures per utterance since the model is asked to select the gestures as well as the phrases. Given the varying number of gestures per utterance, we report our results at two aggregation levels.

Gesture-level (output-weighted metric).

Each model-selected *gesture* is treated as one observation. Gesture-level metrics summarize the quality of a model's overall outputs by aggregating counts across all selected gestures.

Utterance-level (strict metric).
Each *utterance* is treated as one observation. In utterance-level metric, an utterance is labeled as `appropriate` if all the gestures selected by the model are appropriate; otherwise it is labeled as `inappropriate`. In other words, for a model, an utterance is labeled as appropriate if strictly all the selected gestures by the model for that utterance are appropriate. An utterance is marked as inappropriate if there is at least one inappropriate gesture among the selected gestures. This strict metric reflects robustness of the model at the utterance level, where only one single inappropriate gesture can affect the user perception of the entire utterance.

5.4.2 Results: What Gesture to Use

We report figures derived from these labels and aggregation levels.

Figure 3: Appropriateness (gesture-level).
For each model, we compute the proportion of selected gestures that are `appropriate` and the proportion that are `inappropriate`:

$$\text{AppropriateRate} = \frac{\#(\texttt{match}) + \#(\texttt{non-match \& appropriate})}{\#(\text{all predicted gestures})}, \tag{1}$$

$$\text{InappropriateRate} = \frac{\#(\texttt{non-match \& inappropriate})}{\#(\text{all predicted gestures})}. \tag{2}$$

Figure 4: Appropriateness (utterance-level).
For each model, we compute the proportion of utterances that contain *any* inappropriate gesture:

$$\text{StrictInappropriateRate} = \frac{\#(\text{utterances with } \geq 1 \text{ inappropriate predicted gesture})}{\#(\text{utterances})}, \tag{3}$$

$$\text{StrictAppropriateRate} = 1 - \text{StrictInappropriateRate}. \tag{4}$$

Figure 5: Appropriateness and matchness (gesture-level).
For each model, we compute the distribution of gestures over three mutually exclusive categories: `match`, `non-match & appropriate`, and `non-match & inappropriate`. This figure jointly captures appropriateness and matchness by distinguishing suggested gestures that are the exact match to the actual speaker (and thus appropriate) from the suggested gestures that are not the exact match, but are appropriate alternatives to them.

Figure 6: Overall Matchness rate (gesture-level).
In this figure, we first show how often the models are matching the speaker (or gold standard). We further decompose these matched gestures into gold null and gold non-null categories. Meaning we show how often these matched gestures occur on instances

where the speaker gestures (gold nonnull) and does not gesture(gold null). Let T denote the total number of predicted gestures for a given model, and let M_{null} and M_{nonnull} denote the counts of matched selected gestures that occurred in gold-null and gold-nonnull utterances, respectively. We plot:

$$\frac{M_{\text{null}}}{T} \quad \text{and} \quad \frac{M_{\text{nonnull}}}{T}, \tag{5}$$

so that their sum equals the overall matchness rate:

$$\text{MatchRate} = \frac{M_{\text{null}} + M_{\text{nonnull}}}{T}. \tag{6}$$

Figure 7: False positive and Recall (utterance-level).

In this figure we analyze two different metrics: (i) How often the model gestures when the speaker does not gesture (false positive rate), (ii) How often the model gestures when the speaker also gestures (recall). For each model, we compute:

- **false positive rate:**

$$\text{FP}_{\text{null}} = \frac{\#(\text{gold-null utterances where model predicts } \geq 1 \text{ gesture})}{\#(\text{gold-null utterances})}, \tag{7}$$

- **Recall:**

$$\text{Recall}_{\text{nonnull}} = \frac{\#(\text{gold-nonnull utterances where model predicts } \geq 1 \text{ gesture})}{\#(\text{gold-nonnull utterances})}. \tag{8}$$

5.4.3 Discussion of Fast Models

Figures 3–7 provide complementary views of model performance. Figures 3–4 quantify appropriateness at gesture and utterance levels, Figure 5 separates exact matching from appropriate alternatives, and Figures 6–7 characterize behavior conditioned on whether the gold standard is null or non-null, including the model's ability to decide when to gesture.

Appropriateness.

Figure 3 depicts gesture-level appropriateness across models, i.e., the ratio of all selected gestures annotated as appropriate as well as those annotated as inappropriate. Overall, all three models produce mostly appropriate gestures, with slight differences betwen them: Llama3 achieves the highest gesture-level appropriateness (21/26; 80.8%), followed by GPT-4 (22/29; 75.9%) and GPT-5.2 (26/36; 72.2%). It should be noted that almost all the gestures were semantically appropriate; however, some were inappropriate based on their usage in the utterance according to the role of the speaker and the context of the speech. It is important to note that different models

produce different number of gestures, with GPT-5.2 producing the highest number of gestures, followed by GPT-4 and LLAMA3. Because of this difference, Figure 3 should be interpreted as the *quality of produced gestures*, rather than a direct measure of per-utterance robustness.

Therefore, we looked at the gesture selection of these models at an utterance level. Figure 4 follows the same trend as the gesture-level appropriateness, LLAMA3 achieves the highest utterance-level appropriateness (17/22; 77.3%), followed by GPT-4 (16/22; 72.7%) and GPT-5.2 (13/22 ; 59.1%). The difference is that in the utterance-level analysis the appropriateness of the models dropped, with the GPT-5.2 dropping the most.

These results suggest two trends:
(i) At the gesture level, stronger models such as GPT-5.2 appear to support richer, more enigmatic semantic inference, enabling them to map utterance semantics to more distally related gestural intents. While this capability can produce gestures that are semantically appropriate, it may also lead to *over-interpretation*: the model proposes gestures that may be directly semantically related to the utterance, thus misaligned with the speaker's communicative intent or interaction context. As a result, some of these far inferences are judged inappropriate.
Example:

```
Utterance: "... but our values have not changed".
GPT5.2 output: {"Gesture:Regress", "Phrase:have not changed"}
Analysis: This gesture is marked as inappropriate.
```

(ii) At the utterance level, the same tendency toward richer inference shows itself as suggestion of more gestures per utterance. This higher count of suggested gestures increases the likelihood that an utterance contains at least one inappropriate gesture, which is penalized by the strict utterance-level metric. Consequently, models that generate more gestures per utterance may exhibit lower utterance-level appropriateness. It can make parts of the utterance being gesturely emphasized interfere with the speakers intent, stressing further the need to appropriately address the "when to gesture" issue. In addition, metaphoric gestures, or more broadly image schemas are intended to facilitate understanding by embodying abstract concepts. Nevertheless, an excessive number of metaphoric gestures within a single utterance may be counter-productive, as it can overload the listener cognitively, and hinder rather than support comprehension.
Example:

```
Utterance: "It's demanding our policy makers, pressuring our policy makers
to pass an agenda that's already ready for passage, right?"
GPT5.2 output:
{"Gesture":"Progress","Phrase":"demanding our policy makers"},
{"Gesture":"Progress","Phrase":"pressuring our policy makers"},
{"Gesture":"Progress","Phrase":"pass an agenda"}
Analysis:
```

`The first gesture is evaluated as inappropriate and the other two as appropriate.`

Matchness.

In addition to appropriateness, we analyze *matchness* to quantify how often a model reproduces exactly the gestures that the speaker used. Matchness is a more strict metric for evaluating gestures' correctness. Figure 5 shows that most of the suggested gestures fall into the `non-match & appropriate` category, indicating that models frequently generate gestures that although not exactly matching the speaker, are plausible alternatives to the them. This analysis implies that low matchness does not necessarily mean poor gestural behavior, but rather suggests variability in gesture choices and potential differences between the model's inferred intent and the speaker's intent. At the same time, matchness remains a valuable metric for reproducibility and controllability in gesture selection systems, specifically in applications that require consistent gesture realization. As the Figure 5 indicates, LLAMA3 has the highest ratio of match gestures, followed by GPT5.2 and GPT4. Also, GPT4 has the highest ratio of non-match & appropriate, followed by LLAMA3 and GPT5.2. Additionally, we were interested in how often the matchness happens on utterances where the gold standard is null and where the gold standard is nonnull. Figure 6, shows that in the gesture-level analysis, LLAMA3, has the highest number of matchness (34%), with (24%) matchness when the gold standard is null. Meaning that LLAMA3 has higher rate of not gesturing when the speaker does not gesture.

Overall, this analysis reveals an interesting paradox in selecting a language model for gesture selection: despite having weaker general reasoning capabilities, a smaller language model may outperform larger, stronger models in practical gesture selection, by producing gestures that are more appropriate when the speaker gestures and by withholding from gesturing when the speaker does not gesture. This findings suggest that a high inference capability might not be always preferable, and that a mechanism for restraining the gesture production is necessary.

Furthermore, these results indicate the issues associated with selection of too many gestures. While more nonverbal output may appear beneficial for expressiveness, excessive gesture generation can increase the likelihood of inappropriate gestures and may overwhelm, even transform, the communicative intent of an utterance. In the context of virtual agents, generating a large number of gestures does not always enhance realism; rather, it can reduce clarity and naturalness, ultimately lowering user perception of the agent.

6 Implementation

This section details the implementation of our proposed LLM-based nonverbal behavior generation system and explains its integration within SIMA, the Socially Intelligent Multimodal Agent. Figure 8 illustrates the architecture for selecting gestures, assuming the virtual human uses text-to-speech. Upon initiation, the LLM processes a System Prompt that includes conversational context and, potentially, examples of utterance-gesture pairings relevant to that context. The prompt may be tailored for various roles, such as a labor union representative at a panel or a presidential candidate at a

rally. Furthermore, this prompting stage explores different prompting approaches to assess their impact on the creativity and accuracy of gestures proposed by the LLM. Within our architecture, the text-to-speech and behavior scheduling processes are standard components commonly found in virtual human architectures, as described in [50]. The text-to-speech system generates a schedule for the behavior scheduler, which aligns nonverbal behaviors (gestures, visemes) with the corresponding audio. This behavior schedule is then sent to the animation engine in Behavior Markup Language (BML) [51], which manages co-articulation and blending of gestures. Here, we focus on how the LLM can provide specific gesture information to guide the behavior scheduler and animation engine, including labels or physical properties for real-time animation. The full implementation, and information on all prompting approaches is available on GitHub[2].

6.1 SIMA Gesture Generation

The gesture generation component in SIMA takes speech in textual format and generates a BML file specifying the gestures that the virtual human should perform. The model used in SIMA is GPT-4, which is prompted with prompting approach 3, as specified earlier.

6.1.1 Input Processing

The input dialogue text is tokenized and marked with <mark name=""> tags in XML format. These <mark> tags allow the text-to-speech engine to replace markers with precise BML timings, enabling synchronization between spoken words and gestures. In the current implementation, GPT-4 directly performs the structural and semantic analyses of the dialogue text relevant to its proposed gestures.

6.1.2 Gesture Prompting

Using predefined prompts, GPT-4 identifies potential gestures for corresponding parts of the utterance, specifying gestural intents (e.g., container) and their associated phrases (part of the speech where the gesture occurs). Due to how the animation system handles gestures, this step does not include the physical properties of gestures. The GPT4's output is in JSON format, which is then converted into BML. The gestural intent identified by the model is set as the gesture lexeme in the BML, and the first word in the associated phrase (identified by its word number within the complete sentence) in the output of the LLM is set as the stroke-start point in the BML. All gestures generated by GPT-4 are embedded in a single BML file, which is then passed as input to the animation engine. The animation engine then executes the provided gestures according to the provided timings (the stroke-start points). The formatting of the behavior BML block is as follows:

```
<gesture stroke-start="T3" lexeme="Container"
type="METAPHORIC" emotion="neutral" />
```

[2] https://github.com/pariesque/SIMA

6.2 Animation Realization

SIMA uses SmartBody [52, 53] as its animation engine, an open-source framework for real-time animation of conversational agents. SmartBody converts the BML generated by the SIMA gesture generation into character animation aligned with the speech audio. SmartBody's behavior processing consists of a behavior & and schedule manager and a motion controller engine. The behavior & schedule manager parses the BML, extracting behavior requests and their synchronization points. These behaviors encompass gestures, speech visemes, and other nonverbal cues. This system retrieves the timing of the speech markers to synchronize the speech with the behaviors. Then, the scheduler assigns the absolute timing to the behaviors' time markers. Meanwhile, the motion controller manipulates skeletal joints, coordinating movements like gaze, gestures, and head nods to ensure fluid, synchronized animation.

The overall pipeline is automated, from the utterance to the crafting of speech audio, gesture selection by LLM, mapping to the BML specification of the animation, and realization in the SmartBody animation system.

7 Final Discussion

We conducted two complementary analyses to understand how large language models (LLMs) can be used for gesture selection in virtual humans, and what design choices are necessary to make these systems both accurate and deployable. First, we examined multiple prompting strategies for guiding an LLM toward gestures that are meaningful, aligned with the speaker's style, and compatible with a gesture-animation pipeline. Second, we compared three models (Llama3, GPT-4, GPT-5.2) using quantitative metrics that jointly capture gesture *appropriateness*, *matchness* to a gold standard.

Prompting approaches comparison.

Across prompting approaches, we observe a clear trade-off between flexibility and controllability. As expected, the unconstrained baseline (Approach 0) produced the most creative and varied gesture suggestions, but also yielded the highest rate of inappropriate gestures and the weakest alignment with the original speaker's gesture timing and style. In contrast, enriching the prompt with information such as gestural intents, intent descriptions, and speaker-specific examples (Approaches 1–3) systematically improved appropriateness while reducing inappropriate outputs. Qualitatively, model responses were often surprising in their ability to connect utterance meaning to physical motion descriptions, reflecting deep relationships between semantics and gesture form [2, 33]. From a deployment perspective, our animation system (Figure 8) requires a mapping from gesture labels (intents) to a fixed set of realizable animations; therefore, constraining the model's output space (as in Approach 3) aligns naturally with practical virtual human pipelines. Finally, our findings on rheme–theme analysis suggest that discourse structure can further narrow gesture selection to the most salient segments of an utterance, improving alignment with the original speaker's behavior while reducing over-generation.

Model choice comparison

We compared LLAMA3, GPT-4, and GPT-5.2 and observed a consistent pattern: better and stronger language models do not necessarily have better gesture selection capabilities. Although larger models are capable of doing far inferences and suggesting semantically rich and expressive gestures, they also exhibit a stronger tendency to over-gesture, generating more gestures per utterance, including gestures that do not reflect the speaker's intent.

This over-generation, mostly observed in GPT5.2, increases the likelihood of producing inappropriate gestures; and thus when deployed in a virtual human, can reduce perceived realism by introducing unnecessary or distracting motion. On the contrary, a smaller model like LLAMA3, exhibits more restrained behavior. It produces fewer gestures and tends to refrain more from gesturing on instances where the actual speaker does not gesture, which leads to more practical performance in terms of both appropriateness and alignment with the speaker's behavior. Overall, these results highlight that the key challenge for LLM-based gesture selection is not only selecting meaningful gestures, but also controlling *how many* gestures are produced and *when* they occur.

In addition, comparing these models from the perspective of inference timing, we observed that LLAMA3 with average inference time of below one second per utterance, proves suitable for a real-time interaction, which given its higher appropriateness score proves the best choice for gesture selection.

8 Conclusions and Future work

The assumption underlying this work is that different designers may have different intentions so their approach to automating gesture generation may differ. In this work, we are exploring ways to automate gesture selection assuming that a designer wants to retain a degree of control over the form and the usage of the gestures. We specifically demonstrated the use of LLMs to automate the selection of semantically rich gestures for virtual humans. As part of the effort, the approach considered ways of tailoring gesture selection to a specific individual's or role's use of gestures. We first evaluated alternative prompting approaches that varied to the degree that prompting constrained gesture use. The approaches were evaluated by their semantic appropriateness (RQ1). In general, the semantic appropriateness was impressive regardless of whether the LLM was prompted to restrict its gesture use to those gestures common to a specific speaker. We evaluated as well the approaches in terms of the speed of inference (RQ1), in order to determine the suitability for incremental real-time gesture selection as opposed to offline gesture design, as well the capability of the LLM to assess when to gesture (RQ2). We additionally presented initial explorations in the ability to realize spatially consistent gestures in ideational units. Those results were very promising but need further evaluation of methods for the LLM to control the use of gestural space across gestures.

Although the gesture-generation capability of GPT-4 proves promising, it lacks a key requirement for interactive virtual humans: *fast inference*. Real-time human-agent interaction demands low-latency behavior generation, which motivates investigating models that are optimized for efficient inference and preferably locally deployable.

To address this challenge (RQ3), we evaluated a locally deployable language model (Llama3) and the instant version of the newest OpenAI model (GPT-5.2) alongside GPT-4. Our results reveal an important trade-off: smaller models can generate more desirable behavior for gesture selection in practice. In particular, they tend to suggest more appropriate gestures in cases where the actual speaker gestures, and they are more likely to suggest no gesture when the speaker does not gesture.

These findings highlight a broader issue observed in LLM-based behavior generation, namely, a tendency to *over-gesture*. This over-gesturing is also observed in many end-to-end machine learning models that generate gestures for virtual humans. Over-gesturing can reduce the perceived realism of a virtual human and reduce the quality of the interaction by introducing unnecessary or distracting motions. taken together, this analyses suggest that a practical LLM-based gesture selection requires not only high semantic knowledge; but also more control in the generated behavior. These observations emphasized the importance of using additional constraints and selection mechanisms, such as prosodic cues or discourse-based analysis (e.g., rheme–theme analysis), to reduce the gesture space and better align predicted gestures with the speaker's intent.

Finally, we demonstrated how this overall approach was implemented within a virtual human architecture.

Our future steps would be to explore fine-tuning and the use of Retrieval Augmented Generation (RAG) on Llama3 and other small language models.

9 Acknowledgements

This material is based upon work supported by NSF Grant #2128743. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Kendon, A.: Gesture. Annual review of anthropology **26**(1), 109–128 (1997)

[2] Calbris, G.: Elements of Meaning in Gesture vol. 5. John Benjamins Publishing, Amsterdam (2011)

[3] Goldin-Meadow, S., Alibali, M.W.: Gesture's role in speaking, learning, and creating language. Annual review of psychology **64**, 257–283 (2013)

[4] McNeill, D.: Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992)

[5] Tversky, B., Hard, B.M.: Embodied and disembodied cognition: Spatial perspective-taking. Cognition **110**(1), 124–129 (2009)

[6] Jamalian, A., Tversky, B.: Gestures alter thinking about time. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 34, pp. 503–508 (2012)

[7] Bavelas, J.B.: Gestures as part of speech: Methodological implications. Research on language and social interaction **27**(3), 201–221 (1994)

[8] Hostetter, A.B.: When do gestures communicate? A meta-analysis. Psychological Bulletin **137**(2), 297 (2011) https://doi.org/10.1037/a0022128 . Publisher: US: American Psychological Association. Accessed 2022-09-23

[9] Foley, G.N., Gentile, J.P.: Nonverbal Communication in Psychotherapy. Psychiatry (Edgmont) **7**(6), 38–44 (2010). Accessed 2023-04-25

[10] Hall, K., Goldstein, D.M., Ingram, M.B.: The hands of Donald Trump: Entertainment, gesture, spectacle. HAU: Journal of Ethnographic Theory **6**(2), 71–100 (2016) https://doi.org/10.14318/hau6.2.009 . Publisher: The University of Chicago Press. Accessed 2023-04-25

[11] Seizer, S.: On the Uses of Obscenity in Live Stand-Up Comedy. Anthropological Quarterly **84**(1), 209–234 (2011). Publisher: The George Washington University Institute for Ethnographic Research. Accessed 2023-04-25

[12] Özer, D., Göksun, T.: Gesture use and processing: A review on individual differences in cognitive resources. Frontiers in Psychology **11**, 573555 (2020)

[13] Kita, S.: Cross-cultural variation of speech-accompanying gesture: A review. Speech Accompanying-Gesture, 145–167 (2020)

[14] Saund, C., Marsella, S.: Gesture generation. In: The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition, pp. 213–258 (2021)

[15] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G.E., Neff, M.: A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. arXiv:2301.05339 [cs] (2023). https://doi.org/10.1111/cgf.14776 . http://arxiv.org/abs/2301.05339 Accessed 2023-04-26

[16] Hensel, L.B., Yongsatianchot, N., Torshizi, P., Minucci, E., Marsella, S.: Large language models in textual analysis for gesture selection. In: Proceedings of the 25th International Conference on Multimodal Interaction, pp. 378–387 (2023)

[17] Fries, U.: Theme and rheme revisited. In: Modes of Interpretation: Essays Presented to Ernst Leisi on the Occasion of His 65th Birthday, pp. 177–192 (1984)

[18] Grady, J.: Foundations of meaning: Primary metaphors and primary scenes (1997)

[19] Cienki, A.J., Koenig, J.-P.: Metaphoric gestures and some of their relations to

verbal metaphoric expressions. Discourse and cognition: Bridging the gap, 189–204 (1998)

[20] Cienki, A.: Image schemas and gesture. From perception to meaning: Image schemas in cognitive linguistics **29**, 421–442 (2005)

[21] Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Description and construction of text structures. In: Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics, pp. 85–95. Springer, New York (1987)

[22] Quintilian: Institutio Oratoria 11.3. https://penelope.uchicago.edu/Thayer/e/roman/texts/quintilian/institutio_oratoria/11c*.html

[23] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, pp. 413–420 (1994)

[24] Hensel, L.B., Cheng, S., Marsella, S.: Gesres: A richly annotated dataset of spontaneous. interaction **26**(29), 15

[25] Grady, J.E.: In: Hampe, B. (ed.) Image schemas and perception: Refining a definition, pp. 35–56. De Gruyter Mouton, Berlin, New York (2005). https://doi.org/10.1515/9783110197532.1.35 . https://doi.org/10.1515/9783110197532.1.35

[26] Clough, S., Duff, M.C.: The Role of Gesture in Communication and Cognition: Implications for Understanding and Treating Neurogenic Communication Disorders. Frontiers in Human Neuroscience **14** (2020). Accessed 2023-05-08

[27] Guellaï, B., Langus, A., Nespor, M.: Prosody in the hands of the speaker. Frontiers in Psychology **5** (2014). Accessed 2023-05-04

[28] Fares, M., Grimaldi, M., Pelachaud, C., Obin, N.: Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding (2023). https://hal.science/hal-03972415 Accessed 2023-04-27

[29] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., Shapiro, A.: Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '13, pp. 25–35. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2485895.2485900 . http://doi.acm.org/10.1145/2485895.2485900

[30] McNeill, D., Duncan, S.: Growth points in thinking-for-speaking. Language and gesture (1987), 141–161 (2000)

[31] Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477–486 (2001)

[32] Cassell, J., Stone, M., Yan, H.: Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of the First International Conference on Natural Language Generation - Volume 14. INLG '00, pp. 171–178. Association for Computational Linguistics, USA (2000). https://doi.org/10.3115/1118253.1118277 . https://dl.acm.org/doi/10.3115/1118253.1118277 Accessed 2023-04-25

[33] Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge (2004)

[34] Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago press, Chicago (2008)

[35] Neff, M.: Hand gesture synthesis for conversational characters. Handbook of Human Motion, 1–12 (2016)

[36] Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: International Conference on Intelligent Virtual Agents, pp. 243–255 (2006). Springer

[37] Lhommet, M., Xu, Y., Marsella, S.: Cerebella: automatic generation of nonverbal behavior for virtual humans. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 4303–4304 (2015)

[38] Ravenet, B., Pelachaud, C., Clavel, C., Marsella, S.: Automating the production of communicative gestures in embodied characters. Frontiers in psychology **9** (2018)

[39] Zhang, F., Ji, N., Gao, F., Li, Y.: Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In: International Conference on Multimedia Modeling, pp. 231–242 (2023). Springer

[40] Zhang, F., Ji, N., Gao, F., Zhao, B., Wu, J., Jiang, Y., Du, H., Ye, Z., Zhu, J., Zhong, W., et al.: Dim-gesture: Co-speech gesture generation with adaptive layer normalization mamba-2 framework. arXiv preprint arXiv:2408.00370 (2024)

[41] Mughal, M.H., Dabral, R., Habibie, I., Donatelli, L., Habermann, M., Theobalt, C.: Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1388–1398 (2024)

[42] Deichler, A., Mehta, S., Alexanderson, S., Beskow, J.: Diffusion-based co-speech gesture generation using joint text and audio representation. In: Proceedings of the 25th International Conference on Multimodal Interaction, pp. 755–762 (2023)

[43] Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. arXiv preprint arXiv:2305.04919 (2023)

[44] Sun, Y., Xu, Z., Zhou, H., Guan, J., Yang, Q., Wang, K., Liang, B., Li, Y., Feng, H., Wang, J., et al.: Cosh-dit: Co-speech gesture video synthesis via hybrid audio-visual diffusion transformers. arXiv preprint arXiv:2503.09942 (2025)

[45] He, X., Huang, Q., Zhang, Z., Lin, Z., Wu, Z., Yang, S., Li, M., Chen, Z., Xu, S., Wu, X.: Co-speech gesture video generation via motion-decoupled diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2263–2273 (2024)

[46] Cheng, Q., Li, X., Fu, X.: Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. In: SIGGRAPH Asia 2024 Conference Papers, pp. 1–11 (2024)

[47] Chen, B., Li, Y., Zheng, Y., Ding, Y.-X., Zhou, K.: Motion-example-controlled co-speech gesture generation leveraging large language models. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers, pp. 1–12 (2025)

[48] Pang, H., Ding, T., He, L., Tao, M., Zhang, L., Gan, Q.: Llm gesticulator: leveraging large language models for scalable and controllable co-speech gesture synthesis. In: Eighth International Conference on Computer Graphics and Virtuality (ICCGV 2025), vol. 13557, p. 1355702 (2025). SPIE

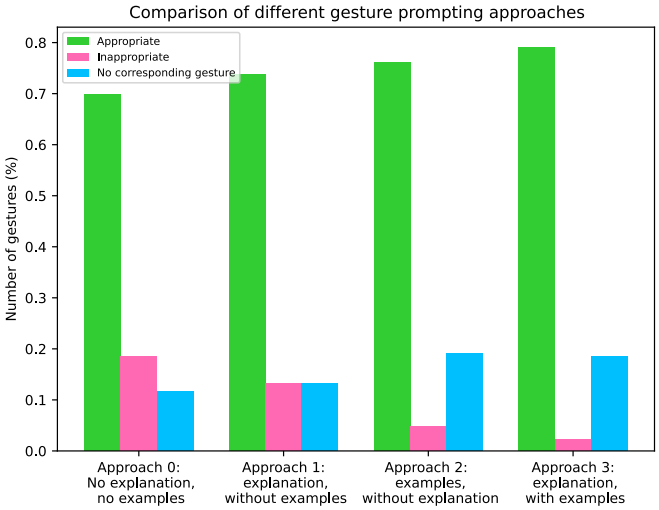
[49] Lhommet, M., Marsella, S.: Metaphoric gestures: towards grounded mental spaces. In: Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14, pp. 264–274 (2014). Springer

[50] Hartholt, A., Fast, E., Li, Z., Kim, K., Leeds, A., Mozgai, S.: Re-architecting the virtual human toolkit: towards an interoperable platform for embodied conversational agent research and development. In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, pp. 1–8 (2022)

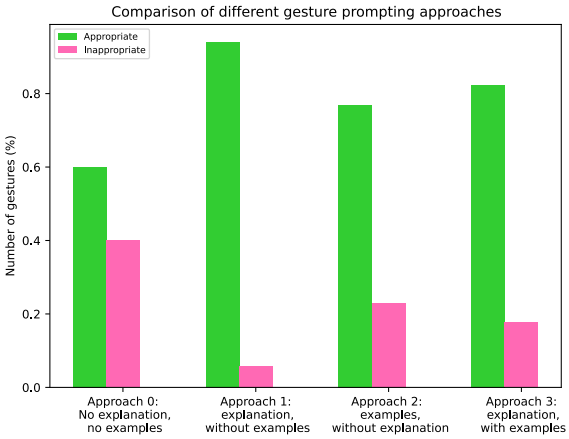
[51] Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) Intelligent Virtual Agents, pp. 205–217. Springer, Berlin, Heidelberg (2006)

[52] Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M.: Smartbody: Behavior realization for embodied conversational agents. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent systems-Volume 1, pp. 151–158 (2008)

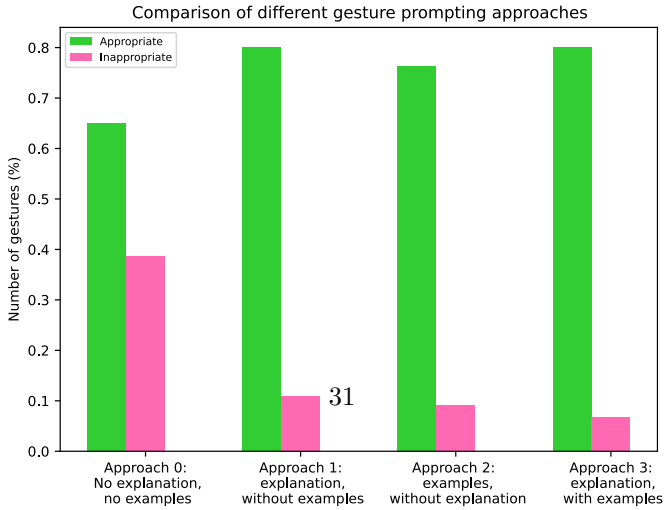
[53] Shapiro, A.: Building a character animation system. In: Motion in Games: 4th International Conference, MIG 2011, Edinburgh, UK, November 13-15, 2011. Proceedings 4, pp. 98–109 (2011). Springer



(a) Selected gestures whose utterances have ground truth

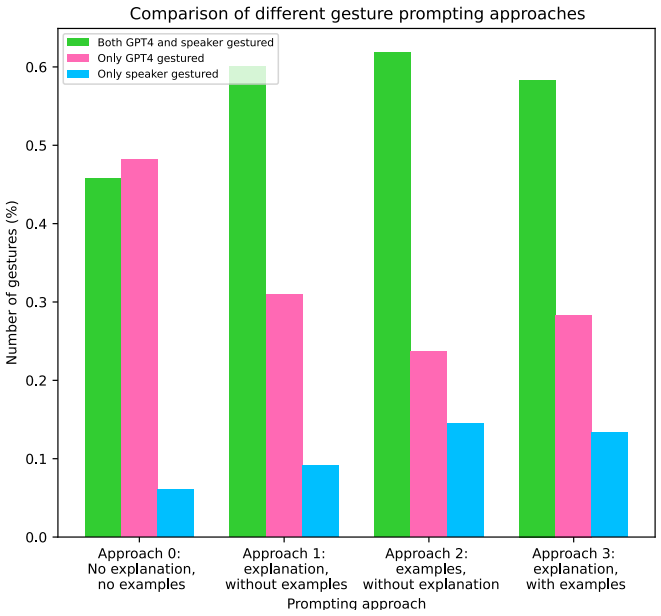


(b) Selected gestures whose utterances do not have ground truth

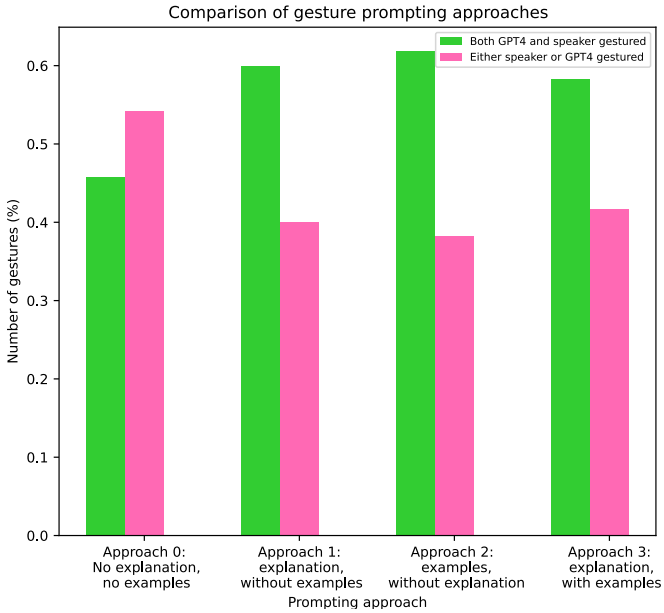


(c) all gestures

Fig. 1: comparison of different prompting approaches in terms of their appropriateness



(a) Both speaker and GPT-4 gestured, GPT-4 gestured, or speaker gestured



(b) Both speaker and GPT-4 gestured, either GPT-4 gestured or speaker gestured

Fig. 2: comparison of different prompting approaches in terms of their alignment with the speaker

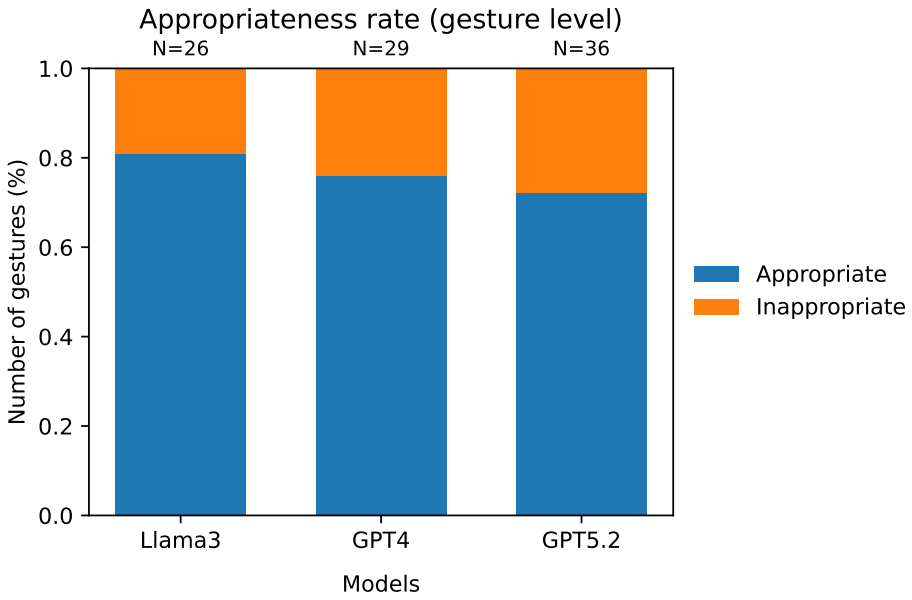


Fig. 3: **Gesture-level appropriateness.** Proportion of predicted gestures labeled appropriate vs inappropriate for each model.

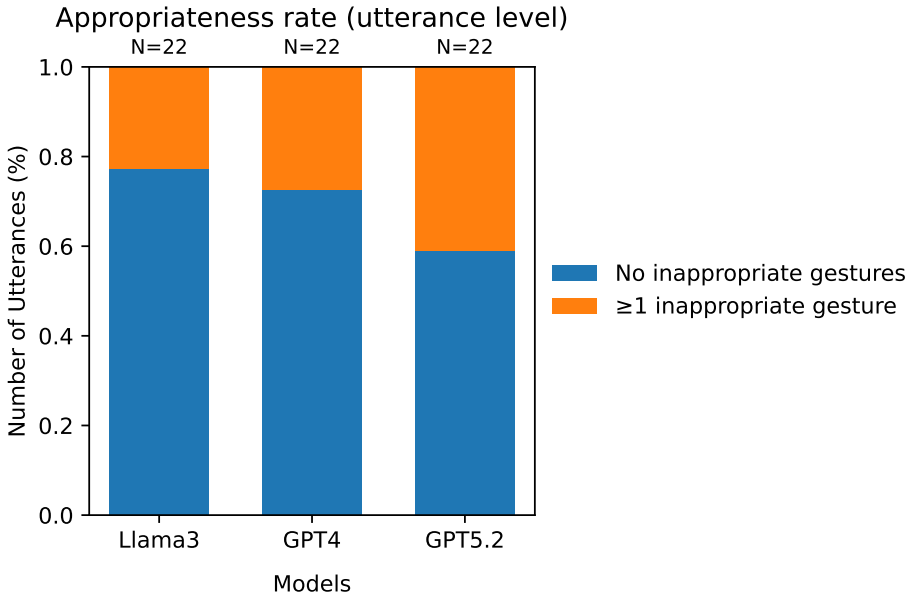


Fig. 4: **Utterance-level appropriateness (strict).** Proportion of utterances containing at least one inappropriate gesture vs none, per model.

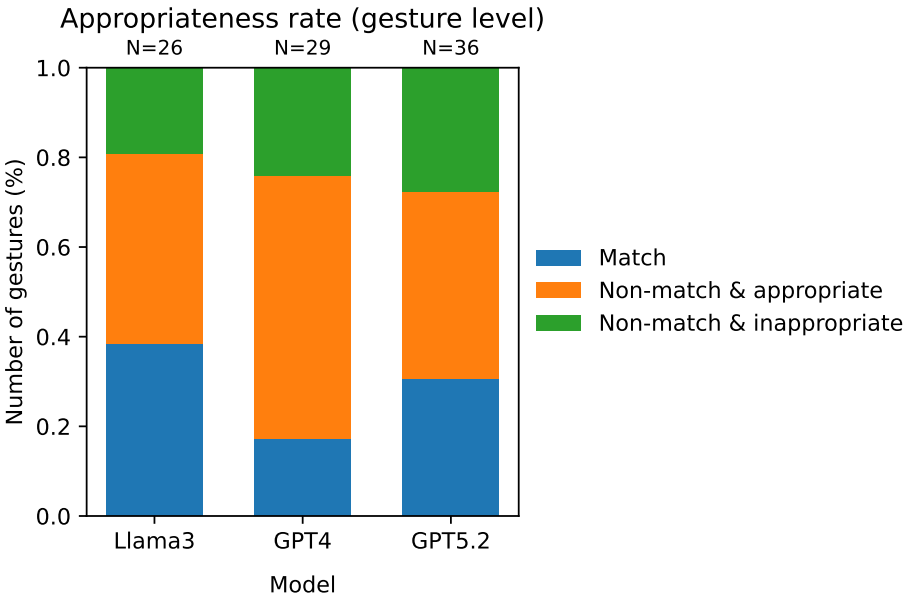


Fig. 5: **Gesture level three categorization** Distribution over Match, Non-match & Appropriate, and Non-match & Inappropriate outcomes for each model.

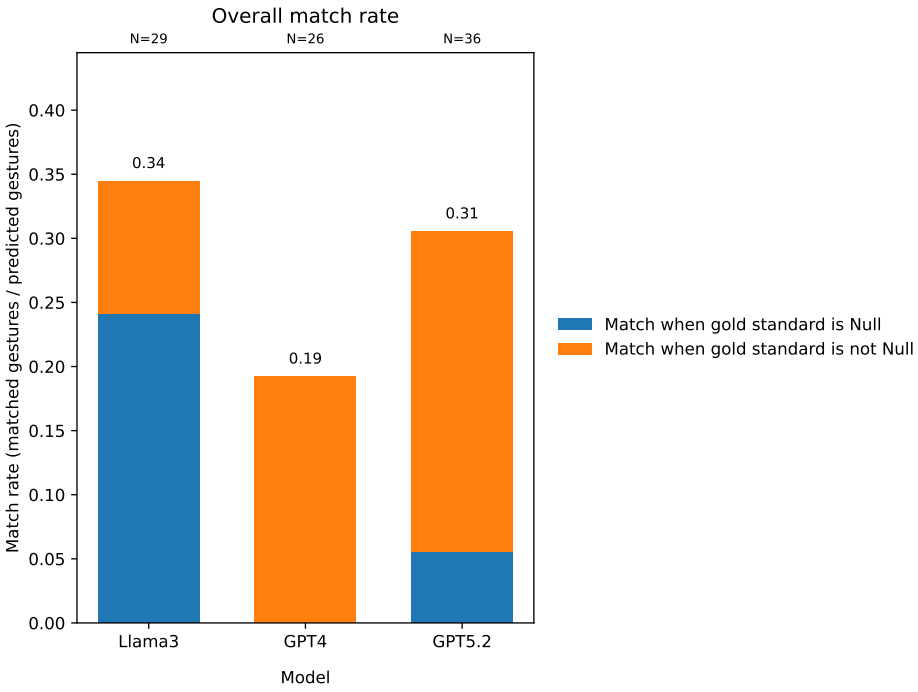


Fig. 6: **Overall match rate (gesture-level).** Overall match rate decomposed into contributions from gold-null and gold-nonnull utterances (stack segments sum to total match rate).

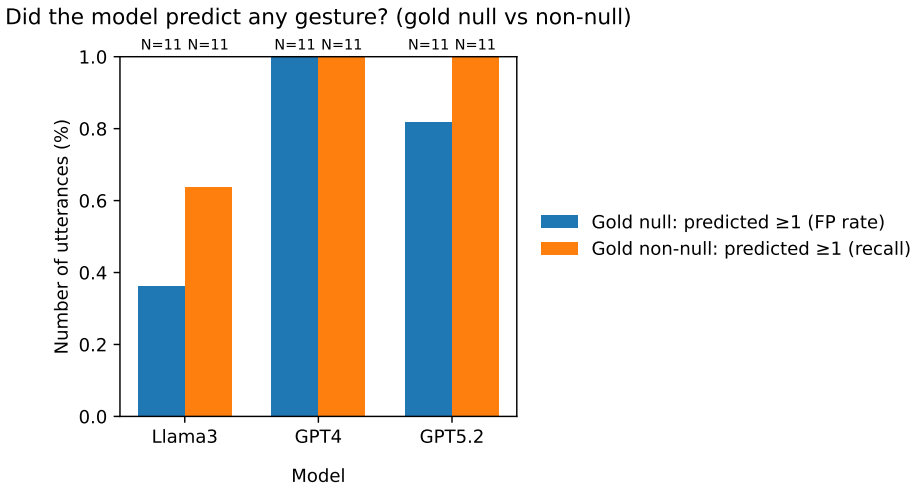


Fig. 7: **FP and Recall (utterance-level).** On gold-null utterances, the bar shows the rate of predicting at least one gesture (over-gesturing / false positives). On gold-nonnull utterances, the bar shows the rate of predicting at least one gesture (recall for gesturing).

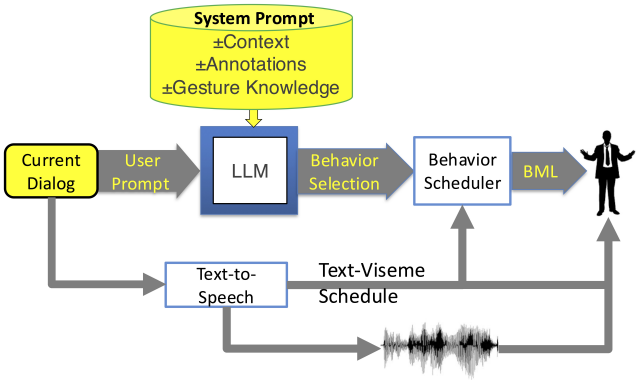


Fig. 8: LLM Approach to selecting gestures.
Based on the type of approach, the input of the LLM can contain the context, annotation (examples), or gesture knowledge (gesture description)