



# Large language models in textual analysis for gesture selection

Laura B. Hensel\*  
University of Glasgow  
laura.hensel@glasgow.ac.uk

Nutchanon Yongsatianchot  
Northeastern University  
nutjung.nutlc@gmail.com

Parisa Torshizi  
Northeastern University  
ghanadparisa@gmail.com

Elena Minucci  
University of Glasgow  
2648114m@student.gla.ac.uk

Stacy Marsella  
University of Glasgow and  
Northeastern University  
stacymarsella@gmail.com

## ABSTRACT

Gestures perform a variety of communicative functions that powerfully influence human face-to-face interaction. How this communicative function is achieved varies greatly between individuals and depends on the role of the speaker and the context of the interaction. Approaches to automatic gesture generation vary not only in the degree to which they rely on data-driven techniques but also the degree to which they can produce context and speaker specific gestures. However, these approaches face two major challenges: The first is obtaining sufficient training data that is appropriate for the context and the goal of the application. The second is related to designer control to realize their specific *intent* for the application. Here, we approach these challenges by using large language models (LLMs) to show that these powerful models of large amounts of data can be adapted for gesture analysis and generation. Specifically, we used ChatGPT as a tool for suggesting context-specific gestures that can realize designer intent based on minimal prompts. We also find that ChatGPT can suggest novel yet appropriate gestures not present in the minimal training data. The use of LLMs is a promising avenue for gesture generation that reduce the need for laborious annotations and has the potential to flexibly and quickly adapt to different designer intents.

## CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models.**

## KEYWORDS

gesture analysis; gesture selection; large language models

### ACM Reference Format:

Laura B. Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. 2023. Large language models in textual analysis for gesture selection. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23)*, October 09–13, 2023, Paris, France. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3577190.3614158>

\*Corresponding Author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*ICMI '23, October 09–13, 2023, Paris, France*

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0055-2/23/10.

<https://doi.org/10.1145/3577190.3614158>

## 1 INTRODUCTION

Human co-speech gestures have been extensively studied. Research has demonstrated that gestures convey a range of meaning and can have a powerful impact on face to face interaction [12, 27, 37, 48]. Gestures positively impact the speaker's persuasiveness as well as an addressee's comprehension, recall, engagement, and trust in the speaker [7, 35, 68]. However these impacts are dependent on the particular gestures being used and the context in which they occur [e.g., 33]. Broadly speaking, people differ in the degree they use gestures, what types of gestures they use and the impact those gestures have [17, 38, 49, 57]. This is especially clear when one considers the situation and role of the gesturer. A clinician may use gestures to establish a safe, trusted place for the interaction with a client as they discuss difficult issues [24]. A politician may seek to be dismissive of opponents [32], or arouse fear and anger [13]. A comedian may use entirely different gestures to generate laughter at their own foibles or ridicule others [65]. A lack of gestures or use of inappropriate gestures can appear uncanny [55, 66].

This potential power of gestures can be exploited in the creation of virtual agents. Selecting gestures for virtual agents is, however, a complex problem due to the multidimensionality of the gesture space. To solve this problem, researchers have focused on two main approaches. The first approach is to predict appropriate gestures based on analysis of the virtual human's utterances [e.g., 16, 44]. Because of a lack of data sets that include both language and co-speech gestures, work in this area has sometimes exploited existing language tools [16, 60, 64] to analyze utterance content and suggest suitable gestures. In particular, machine learning derived tools have been used to analyze prosody and parse rhetorical structure of utterances [36] while ontology tools like Wordnet and Trains [4, 50] have been used to analyze lexical and semantic content. In essence, these approaches aim to generalize example gestures to semantically related utterances. However, the emphasis is increasingly shifting to use of state of the art deep learning approaches to generate gestures directly from the utterance, in part driven by the fact that larger language+gesture corpora have become available (see [54] for a comprehensive review).

With any approach, the data that informs gesture selection is critical to the appropriateness of the output gestures. For example, it may be undesirable for a virtual human clinician to gesture like a politician, a comedian, or someone who rarely gestures. Similarly, it may be counterproductive to base the virtual human's gestures on an average over these different people and roles. Tailoring a virtual human's gestures to the role the virtual human inhabits and the designer's intent for the gesture's impact has potential to

greatly improve the virtual human’s impact [61, 62]. Given this context-dependency of gestures, the central question in gesture generation becomes what gestural data can realize designer intent.

Here, we explore this question from a large language model (LLM) perspective. LLMs have the key advantage that they comprise enough data to perform textual analysis. However, the issue remains how to map that analysis to gestures, in a way consistent with design intentions. In this paper, we demonstrate that LLMs, specifically OpenAI’s ChatGPT and GPT-4, are a powerful tool for selecting appropriate gestures [3, 10]. We address the critical issue of design intention through the use of speaker annotations (see [41] for related work) consistent with design intentions to achieve mappings from the utterance to gesture.

We illustrate the use of LLMs for context-relevant gesture selection based on several examples, starting with the popular but less recent model BERT (Bidirectional Encoder Representations from Transformers [20]) before moving on to explore GPT in more detail. These examples demonstrate the potential power of the approach to not only select individual gestures, but also suggest novel gestures outside the data the system was provided. With this, we aim to demonstrate the transformative potential of LLM’s for gesture generation research. However, we also note that the evaluation of these gestures is still a fundamental challenge in the field [63, 69] and lay out a road map to formally developing the use of LLM’s for gesture generation.

## 2 RELATED WORK

Generating co-speech gestures faces two distinct challenges: The specification problem and the animation problem [52]. While the specification problem deals with *what* gesture should be generated, the animation problem is concerned with *how* to achieve the animation of that gesture. In this paper, we focus on the specification problem. For a review of the work on the animation problem see [54]. The specification problem can be approached with rule-based or data-driven approaches which the following sections discuss.

### 2.1 Rule-based gesture generation

As the name suggests, rule-based approaches rely on a knowledge base along with inference procedures to determine gesture selection, where gestures are often chosen from a gesture dictionary, i.e., a corpus of gestures [14]. This approach frequently relies on manual annotations of gestures and co-speech. These annotations are then used to curate a gesture lexicon as well as synthesize rules for gesture selection. For example, the Behavior Expression Animation Toolkit relies on heuristics derived from extensive behavioral research to select pre-recorded gestures based on linguistic and contextual information [15]. The system uses a bespoke language tagging module and WordNet [50] to automatically extract this information from text. Similarly, Marsella et al. [46] determined the communicative function of utterances using a collection of machine learning based prosodic and syntactic analyses as well as semantic database tools. The communicative functions were then mapped to gestures based on rules codified in an indexed gesture dictionary.

In contrast, Ravenet et al. [60] based their heuristics on image schemas – recurring cognitive processes that aid in establishing meaning [39, 47]. Example image schemas include ‘big’, ‘full’, or

‘object’. The system uses WordNet to identify synonyms for in-text words that correspond to image schemas which they then map to gestures via a gesture dictionary compiled from the literature [60].

While such rule-based gesture generation systems offer a high degree of flexibility when it comes to adjusting rules for gesture-selection, and thereby changing designer intent, they have several key limitations. Firstly, their reliance on gesture dictionaries, i.e., pre-determined sets of gestures, limits the range of gestures that can be produced. As a result, the gestures lack diversity and, by extension, these systems typically only produce a limited number of different gestures for the same speech input. Secondly, generating such extensive heuristics and corresponding gesture dictionaries is time consuming, relies on expert-knowledge and is therefore not scaleable. Because of these limitations, a growing body of work is focusing instead on data-driven gesture generation.

### 2.2 Data-driven gesture generation

Early approaches to data-driven gesture generation often relied on probabilistic mapping from semantic content to gesture type [34, 53]. For example, several approaches used annotated gesture data to produce gesture profiles to which they could apply statistical analyses to ultimately predict plausible gestures for new, annotated input-text [40, 53]. Similarly, Bergmann and Kopp [8] built a Bayesian network to determine whether and what type of gesture should be generated, based on variables such as the communicative action the gesture aims to achieve. While these approaches are more flexible than rule-based approaches, they still often rely on laboriously annotated data and a relatively small set of variables to apply the statistical analyses to.

Because of this, data-driven gesture generation increasingly uses machine learning techniques, and specifically deep learning [1, 54]. Many of these approaches use encode-decoder recurrent neural networks that learn the mapping from utterance text to gestures [6, 72, 73]. However, these approaches often still rely on gesture lexica [6] and in all cases, these data-driven approaches are dependent on the data they are trained on, with limited data sets available. Additionally, much of the semantic richness is often lost [54].

### 2.3 Large Language Models

In contrast, the extensive pre-training of LLMs lends itself as a natural next step in automatic gesture generation. Currently, there are multiple available LLMs, such as OpenAI’s GPTs, Meta’s LLaMA, Google’s Bard, and Anthropic’s Claude, to name a few (for a recent review, see [71]) [3, 5, 10, 28, 67]. One LLM that has found much traction in the research community is BERT [20]. BERT consists of transformer encoder layers and is designed for language modeling and next-sentence prediction. It has been shown to powerfully predict semantic similarities between word- and sentence-pairs [51, 58] and this makes it a potentially valuable tool for gesture generation as mappings from words to gestures can be generalized over synonyms [23, 43]. However, BERT does not have in-built explainable AI, limiting its application to realizing designer intent. For this reason, we explore BERT only briefly before we primarily focus on models from OpenAI.

We chose OpenAI’s models for two key reasons: First, they provide easy access API and second, at the time of writing, GPT-4 is

the most recent and powerful model [59]. Specifically, we use two of OpenAI's Generative Pre-trained Transformer models or GPT: ChatGPT (gpt-3.5-turbo) and GPT-4 [3, 10]. GPTs are an autoregressive decoder-only transformer-based large language models. They are pretrained on large data sets to predict next words and fine-tuned using Reinforcement Learning with Human Feedback to make the models align to human preferences and common tasks [56]. One of the key capabilities of these LLMs in relation to this work is in-context learning – their ability to learn a new task given only a few examples without having to fine-tune the model [10, 70].

Recent work has begun to demonstrate the ability of GPTs' to approximate human cognitive processes. For example, Binz and Schulz [9] studied GPT-3's decision-making, information search, and causal reasoning using common cognitive psychological tests such as the conjunction and base-rate fallacies. They found that GPT-3 can solve these problems at a similar or better than human level but fall for some common human biases. Similarly, Kosinski [42] examined Theory of Mind in LLMs using standard false-belief tasks and found that ChatGPT and GPT-4 can solve most of these tasks. Similarly, Bubeck et al. [11] tested GPT-4 in various domains including mathematics, coding, medicine, law, and psychology, including Theory of Mind and emotion understanding. They showed that GPT-4 can perform these tasks well and outperforms ChatGPT. Taken together, this body of works shows GPTs' promise for applications that normally require human cognitive input.

One such application is gesture generation. GPTs' extensive textual analysis training has the potential to yield gestures that are semantically rich and flexible while not relying on manual annotations. The first foray into this approach used ChatGPT to identify the intent of an utterance, based on this their system then selected a gesture based on a pre-compiled gesture dictionary [25]. The work demonstrated that ChatGPT can successfully aid in gesture generation but did not yet fully realize its potential which lies in ChatGPT's ability to a) suggest new gestures with little input and b) adapt gestures to designer intent.

## 2.4 Gesture generation driven by designer intent

Work that aimed to realize designer intent has often focused on producing gestures for a specific novel target, i.e., a single gesturer. For example, Ahuja et al. [2] were able to generate gestures for a novel target by adapting a pre-trained source model based on just two minutes of target data. Their model did this by learning the difference in the crossmodal grounding, i.e., binding between language and physical gestures, between the source model and the target data. In contrast, other work has focused on learning different speaker styles that can then be used in a zero-shot style transfer to generate gestures for targets not present in training data [22, 26]. These approaches are often very efficient but are also tailored to generate gestures for specific targets.

## 3 EXAMPLES

One approach to applying modern LLMs to gesture generation is to use them to replace the generalization from examples. Earlier rule-based work discussed in previous sections relied on analyses using a range of ontology based tools [4, 29, 50]. Such tools can be

used to analyze text to find similarities with, and thereby generalize from, a set of annotated text-gesture pairings. For example, consider the phrase "important concept". A rule based ontological approach may codify the metaphoric mapping from "important" to *big* and the semantic mapping from "concept" to *idea* which in turn can be metaphorically mapped to a *container* gesture. As the following example shows, language models have the potential to replace this codification.

### 3.1 BERT

The pre-trained BERT model can realize a mapping from text to gesture directly through its cosine similarity measures, even without re-training. Given some phrase-gesture example mappings, derived for example from some annotations of a speaker, and a novel utterance, BERT can measure the cosine similarity between each example utterance and a novel utterance. Additionally it provides a metric for this similarity which can be used to choose potential alternative gestures. Consider the following example where BERT gives similarity measures between *important concept* and five phrases of varying relatedness.

Seed Phrase: *important concept*

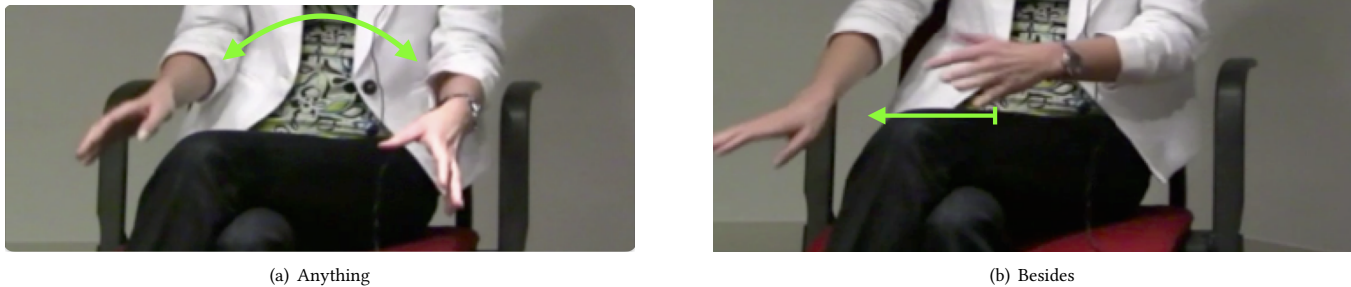
- big idea correlation: 0.861
- great thought correlation: 0.81
- red napkin correlation: 0.239
- weak idea correlation: 0.469
- silly idea correlation: 0.674

We see a strong relation between the seed *important concept* and *big idea* as well as *great thought*. As expected the relation to *red napkin* is the smallest as they are semantically unrelated. The seed's relation with *weak idea* as one might expect is somewhat stronger but still weaker than *big idea* and *great thought*. In contrast, the seed's relation to *silly idea* is slightly stronger than *weak idea*, arguably because silliness has both positive and negative attributes. Critically, however, the exact reason for this lower correlation is hidden from the user at this level of probing the model.

In short, BERT's pre-trained model could substitute for the lexical, semantic, metaphoric analyses discussed above. One would expect BERT to realize a much broader generalization than an ontological approach, though this depends on the boundary condition set on the cosine similarity. A key benefit of BERT is that it provides a clear, overall metric on this mapping as the above examples show. However, the derivation of this metric is opaque and outside the designer's control.

In contrast, the ontology based approach is more transparent as to how this mapping of a novel utterance to gesture suggestion is realized. The semantic, lexical and metaphoric elements are well-defined, which potentially supports finer grain designer manipulations. For example, specific analysis elements could be eliminated or discounted, such as mapping "important" to *size*, or specific analyses could be generalized, such as generalizing abstract concepts to any concept so any object, abstract or physical could be conveyed by a container.

Whereas BERT is arguably similar to prior approaches, ChatGPT, specifically its prompting ability, provides another, novel approach to achieving generalization from example annotations while allowing for designer control through the prompting mechanism.



**Figure 1: Ideational Unit showing a container gesture depicted a protected space followed by a sweep gesture suggesting something discarded from the container**

### 3.2 ChatGPT and GPT-4

To illustrate the power of LLMs for gesture selection, we prompted ChatGPT and GPT-4 with text-gesture pairings which in practice would be drawn from annotations of an actual speaker or speakers. In the following example, we intentionally used opaque terms for the gesture that ChatGPT could not interpret, such as "lcg" meaning "large container gesture" and "ng" for "negation gesture". Based on just three example prompts when given the phrase "it is an excellent idea", ChatGPT responded with "lcg", shown in bold italics.

*ChatGPT.*

- Example prompt: *It is an important idea. lcg*
- Example prompt: *It is a silly idea. ng*
- Example prompt: *It is a big idea. lcg*
- Prompt: *It is an excellent idea.*
- Output: ***lcg***

This method is very similar to the BERT approach – it identifies the similarity between "important", "big", and "excellent" and produces the corresponding gesture. However, if we supply gesture descriptions that are interpretable to ChatGPT, then it produces novel gestural suggestions based on its interpretation of the gestural prompts. The following example demonstrate this.

*ChatGPT.*

- Example prompt: *It is a important idea. large container gesture*
- Example prompt: *It is a bad idea. no gesture*
- Example prompt: *It is a big idea. large container gesture*
- Prompt: *It is an excellent thought.*
- Output: ***thumbs up gesture***

ChatGPT suggests a gesture that is appropriate for illustrating the utterance but was never prompted: "*thumbs-up gesture*". Such novel suggestions raise the question as to the degree to which the prompting is reflecting an underlying relation between utterance and gesture. To illustrate the depth of GPT's ability in this regard, consider an utterance sequence adapted from an interaction between a clinician and woman suffering from PTSD. This utterance and gestural sequence is adapted<sup>1</sup> from an 90 minute corpus in which a real world clinician is going through a PTSD protocol with a woman actor playing the role of a client suffering from PTSD. She

<sup>1</sup>We did not use the exact utterance on one hand to make the context explicit, i.e., a clinician talking to a client, and additionally to remove dysfluencies [45].

is talking to the clinician because of her husband who wants her to change behavior. A key moment in the interaction is as follows:

The clinician is asking a client whether (1) there is anything the client wants to work on, adding the proviso of (2) besides what her husband wants. She uses the following gestures: As she says (1) "anything", she performs a container gesture suggesting a set of all the possible topics they could discuss. Specifically, this container gesture depicts a domed region of space, perhaps suggesting a protected space (see Figure 1). Next, she clarifies what topic she does not want to discuss, (2; "besides what he wants") using a gesture that sweeps to the right, suggesting removing that topic from the container formed by the previous gesture. The "he" is a reference to the client's husband. See Figure 1. These gestures constitute what Calbris [12] calls an ideational unit, a sequence of related gestures. Such gestures tend to set up gestural space in a consistent fashion, such as depicting sets of abstract things and then removing or putting items into the set.

To explore GPT's ability to form ideational units, we first prompted ChatGPT to provide an explanation for the use of the container gesture in the clinician example. This yielded the following explanation that is consistent with a prior analysis of the clinician's gesture done by gesture researchers [19]:

*ChatGPT.*

- Prompt: *The clinician said we can talk about anything. She used a container gesture when she said anything.*
- ChatGPT: *The clinician's use of a container gesture when saying "anything" may suggest that she wants you to feel safe and comfortable to discuss any topic that comes to mind, but also to keep those topics contained within the therapeutic space. The container gesture is a nonverbal cue that can signal a sense of containment and boundaries, creating a safe space for exploration and self-discovery.*

Thus, ChatGPT captured the relation of the container to the possible topic. The response also unexpectedly captured the notion of a protected space which aligns with the specialized dome-shaped container gesture used by the clinician which we did not include in the prompt. GPT-4 yielded similar results but did not capture the subtlety of a "safe space" as explicitly. It instead suggested a container gesture to show inclusiveness and make the patient comfortable.

Importantly, the following example shows that GPT-4 can provide a good prediction of what gesture the clinician could perform next consistent with the gesture sequence the clinician actually used (ChatGPT yielded similar results). This suggests an ability to construct an ideational unit.

GPT-4.

- Prompt: *The clinician said "We can talk about anything" and then she added "besides what your husband wants us to talk about." She used a container gesture to illustrate "talk about anything." To illustrate "except what your husband wants us to talk about", what gesture might she use after the container gesture.*
- GPT-4: *After using the container gesture to illustrate "talk about anything", the clinician could use a swiping or brushing away gesture with one hand to represent "except what your husband wants us to talk about". This gesture would visually convey the idea of removing or excluding a specific topic from the otherwise open and inclusive container of conversation topics.*

Finally, we prompted GPT-4 to visualize the entire action:

- Prompt: *Can you visualize it in some way?*
- GPT-4: *Sure! Here's a step-by-step visualization of the gestures:*
  - (1) The clinician starts by using a container gesture to illustrate "talk about anything." She holds her hands in front of her, palms facing each other, and forms an imaginary circle or container in the air.
  - (2) To illustrate "besides what your husband wants us to talk about," she then moves one hand, palm facing down, from one side of the imaginary container to the other, as if swiping or brushing away a specific topic.
  - (3) The clinician's facial expression could also change to emphasize the exclusion, such as raising an eyebrow or giving a slight smile to show that she's aware of the exception and is still open to discussing other topics.
- This combination of gestures and facial expressions would help to visually convey the message that all topics are open for discussion, except for the specific topic the husband wants them to talk about.

Note GPT-4 has gone beyond the hand gestures to suggest facial actions. These examples illustrate that there is important, and often surprising, potential for gesture selection and analysis in LLMs. The next sections explore this potential more systematically.

## 4 APPROACH

In this section, we detail two approaches to investigating the use of LLMs to a) predict gestures from annotated examples and b) suggest gestures without example prompts. Note that these do not form an exhaustive exploration of using LLMs for gesture selection but are aimed at giving initial insight into the power of these tools. In the discussion, we set out a road map for developing these approaches and the use of LLMs in gesture generation. Note that all analyses in this paper are based on GPT versions from March 2023.

### 4.1 Data collection

To demonstrate GPT's potential for gesture selection, two researchers jointly annotated the first four minutes of Barack Obama's 2020 speech to the Democratic National Convention. Specifically, we focused on three gesture categories that were easily identifiable and semantically meaningful: sweep, span, and container. In total, we identified thirty-seven such gestures. We further distinguished between six unique physical gesture types: span, container, and four types of sweeps with different orientations of the palm (up, down, in, and forward). We also classified these gestures in terms of the semantic meaning conveyed, such as temporal span, negative sweep, and enumeration sweep. The final dataset contained seventeen unique semantic gestures (semantic gesture description + gesture category) and fifteen unique semantic gesture descriptions. Code and data are available at: [https://osf.io/c82tq/?view\\_only=4e0c9c1923cc432abd2c82d0ed8f05c8](https://osf.io/c82tq/?view_only=4e0c9c1923cc432abd2c82d0ed8f05c8)

### 4.2 GPT gesture prompting

For this work, we used ChatGPT and GPT-4. In all of our experiments, we set the temperature parameter to 0, making the outputs mostly deterministic to maximize the replicability of our results. All other parameters were set to default.

To elicit gesture suggestions, we prompted ChatGPT and GPT-4 in the following ways. First, we provided each model with a number of examples (2, 4, 6, or 33, i.e., leave-one-out) for each gesture category. For example, we provided two container gestures, two sweep gestures, and two span gestures. Each prompt was structured in three parts:

- (1) The following contextual statement:  
Barrack Obama is giving a speech at the Democratic National Convention.
- (2) Examples of gestures Obama used. Each example consisted of three elements: The segment of the speech the utterance occurred in, the utterance associated with the gesture, and the gesture. For example:  
He said "I want to talk as plainly as I can about the stakes in this election because what we do these next 76 days will echo through generations to come." When he said "these next 76 days", he used the following gesture: "span".
- (3) The target utterance excluding the gesture. The models filled in the gesture at this point.

We investigate four gesture specifications: Gesture category (span, sweep, container), physical gesture descriptions (e.g., sweep with palm facing up/down), semantic gesture descriptions (e.g., negative sweep), and semantic description only (e.g., negative).

The semantic specification in particular yielded a variety of gestures. We summarise an illustrative selection of these in Table 1 for the reader's convenience.

Finally, to explore GPT's ability to suggest appropriate gestures without any additional user input, we input each utterance prompt without gesture examples. An example includes:

Barrack Obama is giving a speech at the Democratic National Convention. He said "I want to talk as plainly as I can about the stakes in this election because what

Original gesture	ChatGPT	GPT-4
<b>inclusive span</b>	inclusive span/sweep	inclusive span/sweep, compare and contrast
<b>temporal span</b>	temporal span, spatial span, forward span, positive span	temporal span, negation span, counting on fingers or tapping on wrist, forward sweep, guiding motion, continuation, positive sweep, process sweep
<b>positive sweep</b>	upward sweep, emphatic sweep, temporal span	upward sweep
<b>negative sweep</b>	negative/negation span/sweep, inclusive span, negative head shake	negative sweep, downward sweep, negation span, negation or dismissive wave, shrug or open palm, uncertain span

**Table 1: ChatGPT and GPT-4 gesture suggestions for semantic specifications**

we do these next 76 days will echo through generations to come.” When he said “these next 76 days”, what gesture did he use?

## 5 ANALYSIS

To explore LLMs’ performance at predicting context-specific gestures, that is, gestures produced by Barack Obama during his speech, we first calculated prediction accuracies for the four gesture specifications (gesture type, physical gesture description, semantic gesture, semantic only) and for each level of example numbers (2, 4, 6, and leave-one-out). Note, that for semantic specifications, we based this analysis on the first part of the semantic descriptor (e.g., ‘negat’, ‘inclu’) to avoid bias from slight deviations in spelling. Next, we measured pairwise cosine similarity between each original gesture and GPT-4’s suggestions based on word embeddings derived using BERT (ChatGPT excluded for brevity, results are similar). This analysis accounted for cases in which GPT’s description of the gesture differed from the annotation, therefore appearing inaccurate, but nevertheless showed high semantic similarity. Figure 2 shows the results.

Specifically, in the top row, each plot shows the prediction accuracy (y-axis) for each level of examples (x-axis), where ‘LOO’ denotes leave-one-out, for ChatGPT (purple) and GPT-4 (yellow), separately. The dotted line indicates chance performance (1/number of unique gesture types of that category in the dataset). In the bottom row, each plot shows the distribution of pairwise cosine similarities, ranging from 0 (very different) to 1 (very similar), for each level of examples (x-axis) for GPT-4.

Results show that both models predicted the correct gesture category and the correct physical gesture description with above chance accuracy when provided with four or more example prompts. Moreover, both models correctly predicted gesture types with semantic description and on only semantic descriptions with above chance accuracy from as little as two example prompts. Across all four gesture specifications, increasing the number of example prompts improved prediction performance and we observed the highest performance during gesture category and semantic description prediction. However, two examples appear to be too few to predict specific physical gestures. Instead, the models frequently generated gesture descriptions that did not map directly onto the gesture types we examined, thus causing the low performance. Notably, GPT-4 outperformed ChatGPT when predicting gesture categories and physical descriptions but not when predicting specifications involving semantic information such as ‘negative’ or ‘inclusive’.

Figure 3 shows that misclassifications were primarily driven by confusions between span and sweep gestures. GPT-4 performed particularly poorly in this regard (shown in Figure 3). This indicates that the models used these gestures interchangeably.

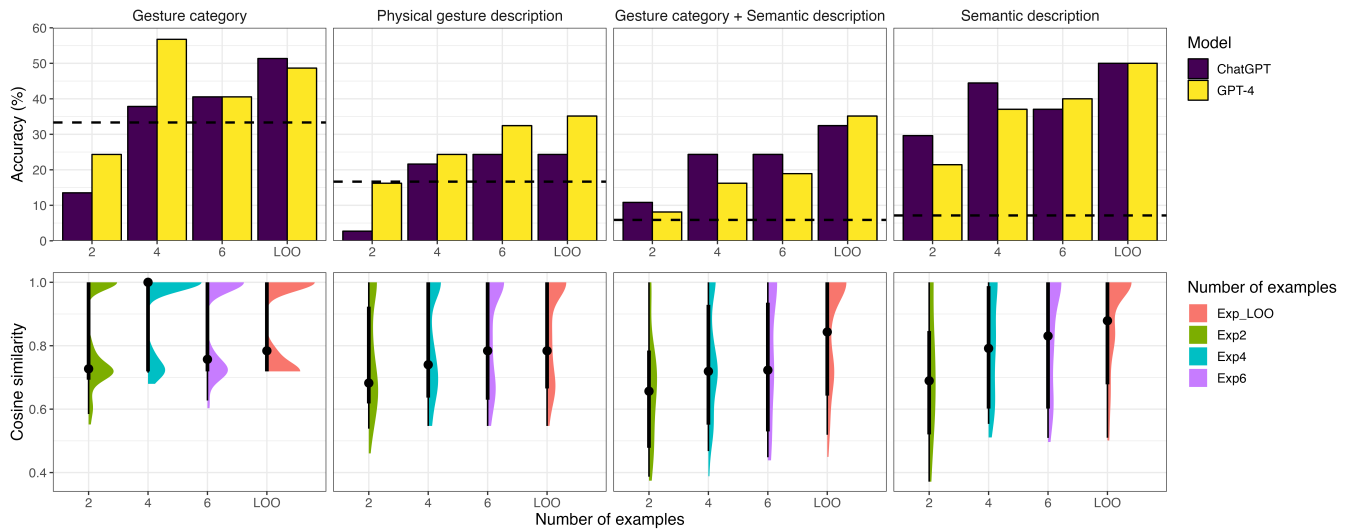
Finally, to explore GPT-4’s ability to generate appropriate co-speech gestures without any example input, we categorized the models’ gesture output when no examples were provided based on its appropriateness. Specifically, we applied one of the following labels to each output gesture: Similar (to the original gesture), Different appropriate (the gesture was different but still appropriate in the context of the speech), Different inappropriate (the gesture was different and not appropriate in the context of the speech), No gesture (the model did not supply a gesture). Two researchers categorized all 37 output gestures of GPT-4, for brevity, and discussed any ambiguous cases until an agreement was reached. Figure 4 shows the resulting appropriateness levels. Red denotes inappropriate suggestions.

Results showed that in 43.2% of cases GPT-4 suggested a gesture similar to the true gesture. Importantly, when the produced gesture was not similar to the true gesture, in most cases (43.2%), the novel gesture was nevertheless appropriate in the context of Obama’s speech. We found GPT-4 only produced inappropriate gestures in 13.5% of cases.

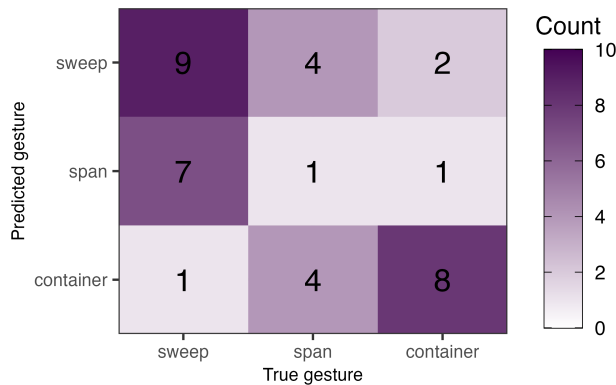
Notably, the model not only made appropriate suggestions for gestures but frequently supplied relevant and appropriate justifications for these suggestions. For example, GPT-4 described the gesture for ‘unleashed’ as “releasing something or setting free”. This description very closely mirrored the actual gesture observed which was a quick motion of the right hand from the palm facing toward the body to the palm facing upward and the hand moving forward.

## 6 DISCUSSION

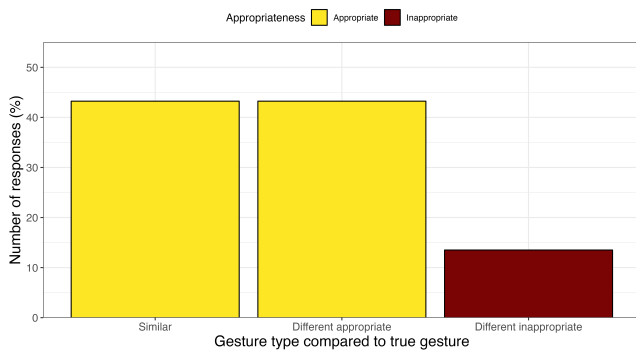
Here, we illustrated several key benefits of using LLMs for gesture selection. We showed that LLMs (1) can predict gestures used in a real-world context with above chance accuracy from as few as two example prompts and (2) can generalize over annotations to suggest novel, previously unseen gestures that are context-appropriate. Combined, these results demonstrate a key benefit of using LLMs for gesture selection – designer control, which is enabled by the selection of the speaker and context. For example, if the design goal for a virtual human is a clinical application where nonverbal behavior can play a critical role [31], the annotation of a trained



**Figure 2: Prediction performance for four gesture specifications. Top: Prediction accuracy (percent correct). Bottom: Cosine similarity between true and predicted gesture (GPT-4).**



**Figure 3: Confusions of gesture category prediction by GPT-4.**



**Figure 4: Appropriateness of gestures suggested by GPT-4 without example prompts.**

clinician with highly effective gesturing skills interacting in a clinical session provides an ideal seed. Indeed, the clinician example highlighted the LLM’s powerful ability to adapt to specific contexts as it appeared sensitive to the clinical context when providing an interpretation of the clinician’s gestures.

Note that we provided the models with very few example prompts and required outputs that matched relatively strict criteria (e.g., ‘sweep’, ‘palm down sweep’, ‘negat sweep’, ‘negat’; depending on gesture specification). Additionally, the differentiation between spans, sweeps, and containers is nuanced because each suggests a region in space. This further adds to the difficulty of the task. Finally, the models were not fine-tuned to achieve gesture selection. Taken these challenges into account, the difficulty of the task should not be underestimated and the observed prediction accuracies of well above chance represent an impressive performance. Moreover, both models not only predicted gesture types and physical gesture descriptions but also semantic gesture specifications. This is especially impressive as we did not provide the models with information about different gesture types or how they relate semantically. Nevertheless, the models were capable of suggesting plausible gestures.

However, a key challenge to applying LLMs to gesture generation is the annotation process itself. For one, manual annotation is laborious. Additionally, designing a gesture naming scheme can be difficult due to the wide variety of real-world gestures and the lack of an agreed upon coding system for gestures, unlike the facial action coding system used in facial expression research[21]. Any coding scheme used needs to reliably distinguish the speaker’s gestures and there can be significant individual and situational differences [33].

Additionally, coding schemes need to distinguish physical properties across semantically related sequences of gestures, i.e., ideational units[12]. As discussed earlier, in an ideational unit, the physical properties of a gesture can be carried over to subsequent gestures.

For example, Obama uses a beat gesture with a precision grip to emphasize a specific point and then holds that precision grip as he does a sweep gesture to illustrate a related issue. The precision grip is not intrinsic to this sweeping gesture but rather is carried over from the previous beat gesture. During annotation, the annotator must ideally distinguish such carryovers of form.

These shortcomings can be circumvented by not using annotations but instead prompting LLMs with just the context and utterance. However, the unconstrained gestural suggestions provided by the models can be hard to parse by machine making automated gesture generation on this basis difficult. Another approach is automatic annotation which would minimize laborious user input. However, this would also cede some degree of designer control. On the upside, we showed that leveraging an LLM's powerful generalization capability does not require a large amount of annotations. The important requirement that remains is to have sufficient annotations to span the speaker's gesture repertoire.

Another important issue is timing, specifically what parts of an utterance are accompanied by a co-speech gesture. Gestures are often associated with what the speaker seeks to emphasize [18]. As a result, altering which part of an utterance is emphasized by a gesture can transform the meaning conveyed. It is therefore critical for co-speech gesture generation to derive emphasis information about the utterance. A common approach to this is to use prosodic cues such as pitch and pauses [22, 30]. However, this presumes that the spoken utterance that is driving gesture generation includes prosodic cues appropriate for the context in which the utterance occurs. A more flexible method is discourse analysis. For example, discourse analysis can identify the *rheme* of the utterance. This is the part of the utterance commenting on and providing new information about the topic or background of the conversation, what is referred to as the *theme*. The rheme tends to be more associated with gesturing [16, 48]. It is likely that LLMs can also be leveraged to identify rhemes.

However, in addition to temporal disconnects between gesture and speech, co-speech gestures often convey additional information that is not transmitted by the utterance at all. For example, in his speech, Obama says "I have sat in the Oval Office with both of the men who are running for president", accompanying the 'oval office' with a sweep of his right hand. This may simply function as an emphasis but, equally, alludes to the length of time that he has spent in that office – 8 years – and the experience he has therefore accumulated. It is a temporal sweep not conveyed by the semantic content of the speech and thus difficult to identify not only for humans but also for machine approaches.

Finally, the models used in this paper are proprietary and not open source. As a result, any update to the models may change the results without recourse for the researcher. This poses an issue for both reproducibility and reliability of any applications based on these models.

It is clear then from the work in this paper that LLMs hold great potential for gesture selection and, with that, generation. Yet, a number of challenges remain to be addressed. In the following we set out our current research efforts to formally and systematically explore the use of LLMs for gesture selection and implementation for gesture generation.

- (1) Expansion of analysis to several types of speakers, including, for example, politicians, psychotherapists, and orators.
- (2) Annotation of these videos, going beyond the restricted set of gestures used in this paper and instead annotating all gestures.
- (3) Expansion of the types of prompts used to explore a) which prompts most effectively aid GPT in suggesting gestures and b) the minimal approach necessary to elicit appropriate and interpretable gestures.
- (4) Extension of analysis of the gestures produced to include rheme, theme, and ideational units.
- (5) Evaluation of the appropriateness of the suggested gestures using human observer studies in addition to expert opinion.
- (6) Implementation of gesture generation by extracting gestures selected by GPT, matching to a gesture dictionary, and generating gestures for virtual human avatars.

## 7 CONCLUSION

In this paper we showcased the use of LLMs to realize a generative mapping between verbal utterances and accompanying co-speech gestures. We explored approaches to using LLMs for gesture selection based on minimal to no example prompts. LLMs consistently suggested context-appropriate gestures. While this approach largely relies on laborious manual annotation, GPT required very few annotated examples to suggest suitable gestures. Nevertheless, it's important to note that the effectiveness of these gesture suggestions hinges on annotations derived from performances by a skilled speaker in a context and role that mirror the designer's intent for a virtual agent. Throughout our exploration of LLMs in this context, we not only highlighted their potent ability to select gestures but also underscored the challenges that still need to be surmounted for the complete implementation of LLMs for gesture generation. We outlined our ongoing research endeavors aimed at overcoming these obstacles.

In conclusion, we view this novel approach as a valuable and versatile tool for gesture researchers to use. However, we recognize that its suitability and alignment with goals may vary, as it either necessitates annotation (when example prompts are used) or interpretation (when example prompts are absent). We nevertheless believe that it will be a fruitful avenue for many gesture researchers.

## ACKNOWLEDGMENTS

The work in this article has been supported by EPSRC Grant Number EP/T021136/1 (LBH, SM) and UKRI CDT in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1 (EM).

## REFERENCES

- [1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1884–1895. <https://doi.org/10.18653/v1/2020.findings-emnlp.170>
- [2] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. 2022. Low-Resource Adaptation for Personalized Co-Speech Gesture Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 20534–20544. <https://doi.org/10.1109/CVPR52688.2022.01991>
- [3] Open AI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

- [4] James Allen, Hannah An, Ritwik Bose, Will de Beaumont, and Choh Man Teng. 2020. A broad-coverage deep semantic lexicon for verbs. *arXiv preprint arXiv:2007.02670* (2020).
- [5] Anthropic. 2023. Claude. <https://www.anthropic.com/product>
- [6] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. *ACM Transactions on Graphics* 41, 6 (Nov. 2022), 209:1–209:19. <https://doi.org/10.1145/3550454.3555435>
- [7] Janet Beavin Bavelas. 1994. Gestures as part of speech: Methodological implications. *Research on language and social interaction* 27, 3 (1994), 201–221.
- [8] Kirsten Bergmann and Stefan Kopp. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 361–368.
- [9] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrk, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [12] Geneviève Calbris. 2011. *Elements of meaning in gesture*. Vol. 5. John Benjamins Publishing.
- [13] Daniel Casasanto and Kyle Jasmin. 2010. Good and Bad in the Hands of Politicians: Spontaneous Gestures during Positive and Negative Speech. *PLOS ONE* 5, 7 (July 2010), e11805. <https://doi.org/10.1371/journal.pone.0011805> Publisher: Public Library of Science.
- [14] Justine Cassell, Matthew Stone, and Hao Yan. 2000. 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)*. Association for Computational Linguistics, USA, 171–178. <https://doi.org/10.3115/1118253.1118277>
- [15] J Cassell, H. Vilhjálmsso, and T. Bickmore. 2001. BEAT: The behavior expression animation toolkit. In *Proceedings of ACM SIGGRAPH*.
- [16] Justine Cassell, Hannes Högni Vilhjálmsso, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.
- [17] Mingyuan Chu, Antje Meyer, Lucy Foulkes, and Sotaro Kita. 2014. Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General* 143, 2 (2014), 694.
- [18] Sharice Clough and Melissa C. Duff. 2020. The Role of Gesture in Communication and Cognition: Implications for Understanding and Treating Neurogenic Communication Disorders. *Frontiers in Human Neuroscience* 14 (2020). <https://www.frontiersin.org/articles/10.3389/fnhum.2020.00323>
- [19] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. (2014), 1061–1068.
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [21] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).
- [22] Mireille Fares, Michele Grimaldi, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. <https://hal.science/hal-03972415>
- [23] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2022. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. <http://arxiv.org/abs/2110.04527> arXiv:2110.04527 [eess].
- [24] Gretchen N. Foley and Julie P. Gentile. 2010. Nonverbal Communication in Psychotherapy. *Psychiatry (Edgmont)* 7, 6 (June 2010), 38–44. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898840/>
- [25] Nan Gao, Zeyu Zhao, Zhi Zeng, Shuwu Zhang, and Dongdong Weng. 2023. GesGPT: Speech Gesture Synthesis With Text Parsing from GPT. <http://arxiv.org/abs/2303.13013> arXiv:2303.13013 [cs].
- [26] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. <https://doi.org/10.1111/cgf.14734> \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14734>.
- [27] Susan Goldin-Meadow and Martha Wagner Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology* 64 (2013), 257–283.
- [28] Google. 2023. Claude. <https://bard.google.com>
- [29] Joseph Grady. 1997. Foundations of meaning: Primary metaphors and primary scenes. (1997).
- [30] Bahia Guellai, Alan Langus, and Marina Nespor. 2014. Prosody in the hands of the speaker. *Frontiers in Psychology* 5 (2014). <https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00700>
- [31] Judith A. Hall, Jinni A. Harrigan, and Robert Rosenthal. 1995. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology* 4, 1 (1995), 21–37. [https://doi.org/10.1016/S0962-1849\(05\)80049-6](https://doi.org/10.1016/S0962-1849(05)80049-6)
- [32] Kira Hall, Donna M. Goldstein, and Matthew Bruce Ingram. 2016. The hands of Donald Trump: Entertainment, gesture, spectacle. *HAU: Journal of Ethnographic Theory* 6, 2 (Sept. 2016), 71–100. <https://doi.org/10.14318/hau6.2.009> Publisher: The University of Chicago Press.
- [33] Autumn B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297. <https://doi.org/10.1037/a0022128> Publisher: US: American Psychological Association.
- [34] Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A Speech-Driven Hand Gesture Generation Method and Evaluation in Android Robots. *IEEE Robotics and Automation Letters* 3, 4 (Oct. 2018), 3757–3764. <https://doi.org/10.1109/LRA.2018.2856281> Conference Name: IEEE Robotics and Automation Letters.
- [35] Azadeh Jamalian and Barbara Tversky. 2012. Gestures alter thinking about time. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34. 503–508.
- [36] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41, 3 (2015), 385–435.
- [37] Adam Kendon. 1997. Gesture. *Annual review of anthropology* 26, 1 (1997), 109–128.
- [38] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
- [39] Adam Kendon et al. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25, 1980 (1980), 207–227.
- [40] Michael Kipp. 2003. Gesture generation by imitation : from human behavior to computer character animation. Universität des Saarlandes. <https://doi.org/10.22028/D291-25852>
- [41] Michael Kipp and Jean-Claude Martin. 2009. Gesture and emotion: Can basic gestural form features discriminate emotions?. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 1–8.
- [42] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023).
- [43] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. ACM, Virtual Event Netherlands, 242–250. <https://doi.org/10.1145/3382507.3418815>
- [44] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 243–255.
- [45] Margot Lhommet and Stacy Marsella. 2014. Metaphoric gestures: towards grounded mental spaces. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*. Springer, 264–274.
- [46] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual Character Performance from Speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation (Anaheim, California) (SCA '13)*. ACM, New York, NY, USA, 25–35. <https://doi.org/10.1145/2485895.2485900>
- [47] David McNeill. 1985. So you think gestures are nonverbal? *Psychological review* 92, 3 (1985), 350.
- [48] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- [49] David McNeill. 2005. Gesture, gaze, and ground. In *International workshop on machine learning for multimodal interaction*. Springer, 1–14.
- [50] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
- [51] Faith Wavinya Mutinda, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. *Methods of Information in Medicine* 60, S 01 (June 2021), e56–e64. <https://doi.org/10.1055/s-0041-1731390>

- [52] Michael Neff. 2016. Hand Gesture Synthesis for Conversational Characters. [https://doi.org/10.1007/978-3-319-30808-1\\_5-1](https://doi.org/10.1007/978-3-319-30808-1_5-1)
- [53] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.
- [54] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. <https://doi.org/10.1111/cgf.14776> arXiv:2301.05339 [cs].
- [55] Terry H. Ostermeier. 1997. *Gender, Nonverbal Cues, and Intercultural Listening: Conversational Space and Hand Gestures*. Technical Report. <https://eric.ed.gov/?id=ED416520> ERIC Number: ED416520.
- [56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [57] Demet Özer and Tilbe Gökşun. 2020. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology* 11 (2020), 573555.
- [58] Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7047–7055. <https://doi.org/10.18653/v1/2020.acl-main.630>
- [59] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277* (2023).
- [60] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018).
- [61] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.
- [62] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217.
- [63] Carolyn Saund and Stacy Marsella. 2021. The Importance of Qualitative Elements in Subjective Evaluation of Semantic Gestures. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.
- [64] Carolyn Saund, Marion Roth, Mathieu Chollet, and Stacy Marsella. 2019. Multiple metaphors in metaphoric gesturing. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 524–530.
- [65] Susan Seizer. 2011. On the Uses of Obscenity in Live Stand-Up Comedy. *Anthropological Quarterly* 84, 1 (2011), 209–234. <https://www.jstor.org/stable/41237487> Publisher: The George Washington University Institute for Ethnographic Research.
- [66] Chidchanok Thepsoonthorn, Ken-ichiro Ogawa, and Yoshihiro Miyake. 2021. The Exploration of the Uncanny Valley from the Viewpoint of the Robot’s Nonverbal Behaviour. *International Journal of Social Robotics* 13, 6 (Sept. 2021), 1443–1455. <https://doi.org/10.1007/s12369-020-00726-w>
- [67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [68] Barbara Tversky and Bridgette Martin Hard. 2009. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition* 110, 1 (2009), 124–129.
- [69] Pieter Wolfert, Jeffrey M Girard, Taras Kucherenko, and Tony Belpaeme. 2021. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 494–502.
- [70] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080* (2021).
- [71] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *arXiv preprint arXiv:2304.13712* (2023).
- [72] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 1–16. <https://doi.org/10.1145/3414685.3417838>
- [73] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.