# Estudo Experimental sobre o Uso de Modelos de Linguagem de Grande Escala na Predição de Intenções de Usuários em Reuniões Híbridas

**Renata Rodrigues dos Santos Briet[1], Elizabete Munzlinger[2], Fabricio B. Narcizo[3,4], Mario T. Shimanuki[5]**

[1] Cursando Tecnologia em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Câmpus Caraguatatuba, renata.briet@aluno.ifsp.edu.
[2] Industrial Ph.D. student, Computer Science Department, IT Univerity of Copenhagen, København S, Denmark, munzlinger@itu.dk.
[3] AI Research Scientist, GN Advanced Science, GN One, fbnarcizo@jabra.com.
[4] Part-time Lecturerомputer, Computer Science Department, IT University of Copenhagen, København S, Denmark, narcizo@itu.dk.
[5] Professor no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, IFSP, Campus Caraguatatuba, mario@ifspcaragua.net

Área de conhecimento (Tabela CNPq): 1.03.02.00-1 – Sistemas de Computação

**RESUMO:** O avanço do trabalho remoto aumentou o uso de Plataformas de Comunicação Unifica (PCUs), como Microsoft Teams, Google Teams e Zoom Meetings. Contudo, a execução de comandos básicos, como controle de volume e câmera, ainda é limitada por interfaces convencionais (mouse e teclado), o que compromete a naturalidade e eficiência da comunicação. Diante dessa lacuna de usabilidade, o presente estudo investiga a aplicação de Modelos de Linguagem de Grande Escala (LLMs) para o reconhecimento de intenção dos usuários por meio da fala. O trabalho propõe, ainda, a integração desse reconhecimento à detecção de gestos de mãos, visando tornar a tecnologia mais robusta e minimizar erros de intepretação na comunicação.

**PALAVRAS-CHAVE**: Reconhecimento de Intenção; Plataformas de Comunicação Unificada; Gestos de Mãos.

## Experimental Study on the Use of Large Language Models for Predicting User Intentions in Hybrid Meetings

**ABSTRACT:** The rise of remote work has boosted the use of Unified Communications Platforms (UCPs), such as Microsoft Teams, Google Meet, and Zoom Meetings. However, the execution of basic commands (e.g., volume and camera control) is often still limited by conventional interfaces (mouse and keyboard), which compromises the fluidity and efficiency of the interaction. Given this usability gap, the present study investigates the application of Large Language Models (LLMs) for recognizing user intent through speech. Furthermore, the work proposes the integration of this recognition with hand gesture detection, aiming to make the technology more robust and minimize interpretation errors in communication.

**KEYWORDS**: Intent Recognition; Unified Communications Platforms; Hand Gestures.

## INTRODUÇÃO

Com o avanço tecnológico a demanda por trabalhos home office cresceu, somente no Brasil, em 2022, o contingente de teletrabalhadores chegou a 7,4 milhões de pessoas (IBGE, 2023), tornando o computador uma ferramenta essencial de trabalho, e as reuniões, antes presenciais, são majoritariamente realizadas em Plataformas de Comunicação Unificada (PCUs), como o Microsoft Teams, Google Teams e Zoom Meetings. A adoção dessas plataformas traz praticidade e flexibilidade, permitindo que empresas superem barreiras geográficas e formem equipes remotas sem restrição territorial.

Apesar do papel fundamental dessas plataformas no emprego de milhões de pessoas, a interação e usabilidade para execução de comandos ainda apresentam limitações práticas. Atualmente, a execução de comandos básicos como controle de volume ou silenciamento, depende do uso de dispositivos convencionais (mouse e teclado) ou, em raras implementações, da entrada por voz. Contudo, no contexto dinâmico das reuniões, a dependência desses dispositivos convencionais interrompe o fluxo da comunicação e exige que o usuário, muitas vezes distante do computador, pause sua atividade para interagir com a tela (MUNZLINGER et al. 2025). Essa limitação demonstra uma lacuna na usabilidade, indicando a necessidade de métodos de interação mais intuitivos e acessíveis, como o reconhecimento de gestos de mãos.

A busca por métodos de interação que ultrapassem o paradigma do mouse e teclado é um tema relevante na área de Interação Humano-Computador (IHC). Estudos demonstram que a incorporação da execução natural de comandos otimiza a experiência do usuário, pois reduz a carga cognitiva estranha e melhora a eficiência na execução de tarefas (NIELSEN, 1993). Além disso, a comunicação humana é essencialmente multimodal, onde os gestos desempenham um papel fundamental na expressão e no entendimento (MCNEILL, 1992).

Embora o reconhecimento de gestos de mãos para a execução de comandos proporcione benefícios como interações mais naturais, intuitivas e reuniões mais eficientes, ele gera a necessidade de que o sistema seja robusto. É crucial minimizar os erros de interpretação que podem ocorrer quando o usuário realiza um gesto de forma espontânea, sem a intenção de executar um comando (BRIET et al. 2025).

Com base nesse cenário, este estudo visa avaliar a capacidade dos Modelos de Linguagem de Grande Escala (LLM) em identificar intenções de comandos a partir da fala dos usuários, explorando a viabilidade da análise das transcrições das falas dos participantes durante as reuniões, utilizando a capacidade dos LLMs de processar e identificar "gatilhos" ou padrões de linguagem que sinalizam uma intenção de comando. Essa abordagem inicial, focada no texto, estabelece a base para futuras implementações que visam a integração de dados multimodais, como gestos, para uma interpretação contextual mais robusta.

## METODOLOGIA

A pesquisa adota uma abordagem qualitativa voltada à análise e associação entre transcrições verbais e comandos de controle da plataforma em contextos de reuniões virtuais, sendo continuidade de um estudo realizada em 2024, que definiu gestos intuitivos para comandos comuns em Plataformas de Comunicações Unificadas (PCUs). O atual objetivo do processo é entender a viabilidade de utilizar LLMs como ferramenta auxiliar para prever as intenções dos usuários a partir da análise das transcrições das falas. O processo metodológico envolveu a coleta e o pré-processamento de um conjunto de dados misto, a codificação da *pipeline* e a realização de testes de validação.

O atual estudo segue como base os oito comandos abordados em (BRIET et al. 2024), e comuns em ambientes virtuais: Aumentar Volume, Diminuir Volume, Mutar Microfone, Desmutar Microfone, Desligar Câmera, Ligar Câmera, Pedir para Falar e Encerrar Chamada. Este estudo ainda amplia o escopo para dez comandos, incluindo "Compartilhar Apresentação" e "Passar Slide", por sua relevância prática.

O conjunto de dados utilizado para o Modelo de Linguagem de Grande Escala (LLM) foi composto por dados de duas fontes:

a) Vídeos do YouTube: Foram coletados sete vídeos de reuniões virtuais no YouTube, com duração de 4 minutos até 180 minutos. Embora a busca inicial tenha sido de conteúdos em português, o estudo foi redirecionado para vídeos em inglês, devido à limitação de material que apresentasse variedade de falas espontâneas, relacionadas à necessidade de ajustes da plataforma, conforme lista de comandos previamente citada. O objetivo foi coletar vídeos que apresentassem em seu conteúdo falas dos participantes relacionadas às intenções de ajustes do ambiente virtual, com alguma associação com quaisquer dos oito comandos previamente abordados no estudo e supracitados. Foram realizadas anotações das falas e dos gestos espontâneos (como erguer as mãos para falar) que indicassem a mesma intenção do comando.

b) Simulação de reuniões: Também foram gravadas simulações de reuniões virtuais com voluntários na plataforma Zoom Meetings. Para manter a conformidade dos dados, as conversas foram realizadas em inglês, seguindo cinco *scripts* de conversas pré-definidos para simular diferentes assuntos e incluir os comandos estudados. Os voluntários foram instruídos a realizar gestos de sua preferência com as mãos ao expressarem verbalmente a intenção de executar um comando. O objetivo foi obter transcrições de vídeos que cobrissem os dez comandos a serem analisados no estudo.

O áudio das simulações (salvo automaticamente pelo Zoom) e o *script* das falas dos vídeos do YouTube foram transcritos utilizando a implementação *Faster Whisper* da *Systran*[1]. A versão utilizada nesta pesquisa foi adaptada com um *script* complementar para automatizar o *download* do áudio dos vídeos em formato MP3 (via biblioteca yt_dlp) e a geração das transcrições em arquivos de texto (.txt), contendo o tempo do vídeo e as falas correspondentes. O código dessa adaptação está disponível no repositório: https://github.com/RenataRSBriet/transcription-faster-whisper. Todas as transcrições coletadas compõem o conjunto de dados que está sendo utilizado para o experimento com a LLM.

A FIGURA 1 apresenta um exemplo de trecho de transcrição de vídeo utilizado na coleta de dados, ilustrando a formatação com o tempo de ocorrência e as falas correspondentes. O trecho destaca frases que expressam a intenção de comandos específicos, como ligar a câmera, passar slide e pedir a palavra, de acordo com o que foi dito pelo participante.

```
224   [914.00s → 915.00s]  I'm back.
225   [915.00s → 918.00s]  I'll continue explaining.
226   [918.00s → 921.00s]  So I'm turning my camera back on.
227   [921.00s → 923.00s]  Could you move to the next slide please?
228   [923.00s → 924.00s]  Got it.
229   [924.00s → 927.00s]  Move to the next slide now.
230   [927.00s → 928.00s]  Perfect.
231   [928.00s → 931.00s]  Also, I need to request to speak for a quick update.
```

Figura 1 – Fragmento de marcação temporal e intenções de comando.

Para o reconhecimento da intenção dos usuários a partir da fala, o ChatGPT Plus foi selecionado como o Modelo de Linguagem de Grande Escala (LLM) a ser avaliado. O LLM é alimentado com o conjunto de dados de fala transcrita, utilizando tanto os *scripts* de vídeos do YouTube quanto os vídeos capturados pelos autores. Esse processo visa dotar o modelo da capacidade de entender ambiguidades e variações naturais, o que é essencial para o sistema, já que o modelo deve antecipar ou reforçar comandos gestuais, resultando em maior precisão.

Para o experimento com o modelo foi criada uma versão customizada do ChatGPT na plataforma da OpenAI com engenharia de prompt em *zero-shot*, ou seja, sem apresentação de nenhum exemplo de entrada-saída, conforme visto na FIGURA 2.

```
ROLE AND CONTEXT
You are the definitive Intent Validation Engine for a hybrid meeting tool. Your function is to confirm
user intent from speech, reducing command errors.

ACTION:
For every user input, analyze the speech and classify the intent based ONLY on the VALID INTENTIONS
list. If the input is general conversation, use the 'NO_COMMAND' intention. You MUST provide the output
as a JSON object, and ONLY the JSON object.

VALID INTENTIONS (Do not deviate from these):
1. INCREASE_VOLUME
2. DECREASE_VOLUME
3. CAMERA_ON
4. CAMERA_OFF
5. MUTE_MIC
6. UNMUTE_MIC
7. ASK_TO_SPEAK
8. END_CALL
```

[1] Disponível em: https://github.com/SYSTRAN/faster-whisper. Acesso em: 8 out. 2025.

```
9. SHARE_SLIDES
10. NEXT_SLIDE

OUTPUT FORMAT (CRITICAL):
You MUST provide the output as a LIST of JSON objects (an array), enclosed EXCLUSIVELY within a Markdown
code block (```json ... ```). DO NOT add any conversational text, explanations, or markdown before the
code block.
```json
[
  {
    "Timestamp": "[The timestamp read at the start of the line, e.g.,]",
    "Intention":
"[INCREASE_VOLUME|DECREASE_VOLUME|CAMERA_ON|CAMERA_OFF|MUTE_MIC|UNMUTE_MIC|ASK_TO_SPEAK|SHARE_SLIDES|N
EXT_SLIDE|END_CALL]",
    "Confidence": "[High|Medium|Low]",
    "Original_Speech": "[The analyzed speech input, including speaker and line]"
  }
]

Rule: If the user input is ambiguous or has multiple possible meanings, set "Confidence" to "Medium".
If the intent is perfectly clear, set "Confidence" to "High".
```

Figura 2 – Prompt fornecido ao Modelo de Linguagem de Grande Escala (LLM)

A FIGURA 2 retrata o *prompt* fornecido ao Modelo de Linguagem de Grande Escala (LLM), este *prompt* tem função de transformar o LLM em um "Motor de Validação de Intenção". O objetivo principal é que o modelo analise a fala dos usuários a partir das transcrições e classifique a intenção comunicada, atuando como um sistema auxiliar para prever comandos de controle.

Para garantir a relevância e a precisão da classificação, o *prompt* utiliza dez Intenções Válidas, todas diretamente relacionadas ao controle de um ambiente de reunião híbrida, como INCREASE_VOLUME, MUTE_MIC e NEXT_SLIDE. O formato de saída é uma exigência crítica: o LLM deve gerar a resposta como uma lista de objetos JSON, exclusivamente em um bloco de código *Markdown*, sem quaisquer explicações ou texto conversacional adicional. Essa padronização é essencial para que os resultados do experimento sejam processados e avaliados de maneira automatizada e eficiente. O formato JSON inclui campos como *Timestamp*, a *Intention classificada* e o *Original Speech* analisado.

O campo *"Confidence"* (Confiança) representa uma métrica importante para a avaliação do LLM. O modelo é instruído a classificar a clareza da intenção detectada como *"High*" (Alta), *"Medium"* (Média) ou *"Low"* (Baixa). Essa auto validação permite que o estudo analise não apenas o que o modelo previu, mas também o grau de certeza em diferentes cenários de fala ambígua ou complexa, fornecendo *insights* sobre as limitações do uso de dados puramente textuais.

## RESULTADOS E DISCUSSÃO

O presente estudo demonstra como a detecção de intenção pela fala atua como um fator preditivo para comandos gestuais, aprimorando a usabilidade e a confiabilidade do sistema. É importante ressaltar que o módulo LLM ainda se encontra em fase de treinamento e validação. Os resultados parciais, no entanto, validam a estratégia de coleta de dados e o treinamento do modelo, comprovando a capacidade inicial de extrair intenções de comandos sequenciais em contextos de diálogo natural.

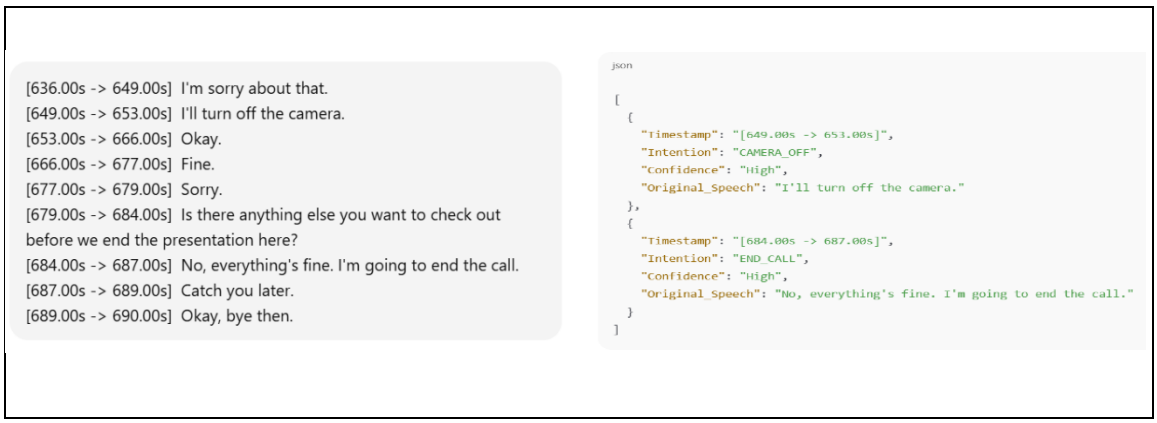


Figura 3 – Demonstração do Fluxo de Trabalho do Modelo de Linguagem de Grande Escala (LLM).
```

A FIGURA 3a (Entrada) apresenta o *script* do diálogo submetido, enquanto a Figura 3b (Saída) ilustra o retorno do LLM no formato de lista JSON: CAMERA_OFF ([649.00s -> 653.00s]) e END_CALL ([684.00s -> 687.00s]). O sucesso na extração sequencial e o retorno do *timestamp* para cada comando validam o LLM como um módulo preditivo capaz de fornecer a referência temporal necessária para a sincronização com um sistema de detecção gestual. Ambas as classificações foram atribuídas com confiança "High", reforçando a robustez do treinamento em classificar comandos diretos.

A análise do trecho de encerramento também revelou a capacidade do LLM de priorizar a intenção acionável em relação às interações sociais. Embora o diálogo final contivesse frases como *'Catch you later'* e *'Okay, bye then'*, que dessem a entender o encerramento da chamada, o modelo classificou a intenção de comando uma única vez na frase *'I'm going to end the call'*. Essa precisão garante que o comando seja acionado no ponto exato da decisão, ignorando saudações posteriores, o que é fundamental para a eficiência do sistema.

No entanto, este achado levanta uma questão crítica sobre a ambiguidade de intenção em chamadas coletivas. O LLM não diferencia semanticamente entre a intenção de encerrar a reunião para todos e sair individualmente. Este risco de erro de execução em comando único exige o refinamento do vocabulário do LLM e a inclusão de comandos que permitam essa diferenciação, como LEAVE_MEETING, para diferenciar as intenções de encerrar ou sair da chamada.

Esta limitação do LLM em diferenciar saídas individuais demonstra a incapacidade do reconhecimento unimodal (somente fala) de resolver ambiguidades, e reforça a necessidade da validação gestual. Se o LLM reconhece o END_CALL de forma ambígua, a sincronia com o módulo gestual se torna a camada de segurança para a execução correta: o comando só deve ser executado se houver a concordância do gesto. Em última análise, a capacidade de isolar, classificar e referenciar no tempo as intenções demonstra a robustez do LLM como módulo preditivo, e sua combinação com gestos é o caminho para garantir a confiabilidade e a segurança do sistema em contextos de reuniões dinâmicas.

## CONSIDERAÇÕES FINAIS

O estudo demonstrou a viabilidade do uso de LLMs para interpretar falas transcritas de usuários de modo a localizar e entender conteúdos relacionados à intenção de execução de comandos em reuniões virtuais, como os previamente listados.

Os resultados preliminares demonstraram que o treinamento do LLM tem alta eficácia na extração de comandos de intenção em contextos de diálogo natural e sequencial, classificando ações com alta confiança. Este sucesso na extração sequencial e no referenciamento temporal (*timestamp*) valida o LLM como um módulo preditivo essencial para a sincronização com o sistema de detecção gestual.

Contudo, a análise dos achados levantou uma questão importante sobre as limitações do reconhecimento unimodal. A incapacidade do LLM de diferenciar semanticamente comandos ambíguos, como o encerramento da chamada para todos e a saída individual da reunião, enfatiza a necessidade da validação gestual. Essa ambiguidade prova que a integração de gestos não é apenas uma melhoria de usabilidade e inclusão, mas sim uma camada de segurança crucial para a execução correta e confiável dos comandos.

## TRABALHOS FUTUROS

Como próximos passos, o projeto seguirá para a codificação final e testes para consolidar este caminho para interações virtuais mais intuitivas e eficazes, melhorando a experiência dos usuários. Na fase final, o foco central está em realizar ajustes no prompt de comando da LLM, avaliando resultados para *zero shot* e *few shot*, bem como avaliações de *benchmark*. Especificamente, os esforços envolverão a codificação da lógica de associação de eventos e a execução de testes finais de validação, visando confirmar a eficácia plena e a robustez do sistema multimodal em operação.

## AGRADECIMENTOS

Os autores agradecem ao Instituto Federal de Educação, Ciência e Tecnologia de São (IFSP), por ter celebrado o *Memorandum of Understanding (MOU)* entre IFSP e GN Audio A/S, e ao Programa Institucional de Bolsas de Iniciação Científica (PIBIFSP 2025), pelo apoio e pela concessão da bolsa de Iniciação Científica, que viabilizou a realização deste projeto.

**REFERÊNCIAS**

MUNZLINGER *et al*. Eliciting User-Defined Mid-Air Hand Gestures for Hybrid Meeting Platform Control: Results, Insights, and Design Implications. In: INTERACT 2025, 20., 2025, Belo Horizonte. Proceedings. **Lecture Notes in Computer Science**. Springer, 2025. Disponível em: < https://link.springer.com/chapter/10.1007/978-3-032-05005-2_22>. Acesso em: 25 set. 2025.

BRIET et al. Reconhecimento de Gestos como Alternativa Eficiente para Execução de Comandos em Reuniões Virtuais. In: **CONICT** –16., 2025, Campinas. (No prelo).

BRIET *et al*. Interação por Gestos de Mãos em Plataformas de Comunicações Unificadas. In: **SICLN** –14., 2024, Caraguatatuba. Disponível em: < https://ocs.ifspcaraguatatuba.edu.br/sicln/XIVSICLN/paper/view/515>. Acesso em: 25 set. 2025.

**IBGE (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA).** Pesquisa inédita do IBGE mostra que 7,4 milhões de pessoas exerciam teletrabalho em 2022. *Agência de Notícias IBGE*, Rio de Janeiro, 25 out. 2023. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/38159-pesquisa-inedita-do-ibge-mostra-que-7-4-milhoes-de-pessoas-exerciam-teletrabalho-em-2022. Acesso em: 25 set. 2025.

MCNEILL, David. **Hand and mind: What gestures reveal about thought**. University of Chicago press, 1992. Disponível em: https://www.jstor.org/stable/1576015?origin=crossref. Acesso em: 22 set.2025.

NIELSEN, Jakob; LANDAUER, Thomas K. A mathematical model of the finding of usability problems. In: **Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems**. 1993. p. 206-213. Disponível em: https://doi.org/10.1145/169059.169166. Acesso em: 24 set. 2025.

PAAS, Fred; VAN GOG, Tamara; SWELLER, John. Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. **Educational psychology review**, v. 22, n. 2, p. 115-121, 2010. DOI: 10.1007/s10648-010-9133-8. Acesso em: 25 set. 2025.

SYSTRAN. **faster-whisper: Faster Whisper transcription with CTranslate2.** GitHub. Disponível em: https://github.com/SYSTRAN/faster-whisper. **Acesso em: 8 out. 2025.**

YASEEN et al. Next-gen dynamic hand gesture recognition: Mediapipe, inception-v3 and lstm-based enhanced deep learning model. **Electronics**, v. 13, n. 16, p. 3233, 2024. Disponível em: https://www.mdpi.com/2079-9292/13/16/3233. Acesso em: 03 set. 2025.