The *DeepSpeak-Agentic* Dataset

Sarah Barrington
UC Berkeley
sbarrington@berkeley.edu

Maty Bohacek
Stanford University
maty@stanford.edu

Hany Farid
Dartmouth College
hany.farid@dartmouth.edu

Abstract

We present DeepSpeak-Agentic*, a dataset of videos comprising over* 37 *hours of semi-structured conversations between a human and an embodied AI agent. We use this dataset to evaluate the automatic forensic identification (audio, video, or text) of AI agents, study the nature of human-agent interactions, and provide a benchmark for future advances in the large language models and AI-generated voices and faces that power embodied AI agents. We also contribute a scalable data-capture system that creates agents, automatically pairs them with human participants, records audiovisual conversations across specified scenarios, and identifies and separates the human and agent in the combined stream.*

1. Introduction

In 2024, Zoom's CEO Eric Yuan envisioned a future where an agentic AI clone would join meetings on our behalf [20]. Just 11 months later, Yuan delivered Zoom's 2025 Q1 earnings with his AI clone. This clone, however, was prerecorded and was only able to deliver the prepared opening remarks, and not respond in real time to questions. In response to an analyst's comment "Nice avatar.", Yuan responded. "Thank you. Appreciate it. Next earnings call will be much better" [18] (presumably Yuan was referring to the AI clone, not Zoom's earnings).

From a technical perspective, Yuan's prediction was correct. These past two years have seen tremendous progress in real-time, embodied agents powered by large language models (LLMs), synthetic voices, and visual avatars, getting us closer to Yuan's vision. Indeed, today, there are multiple commercial offerings that allow humans to converse with agents (e.g., https://anam.ai, https://heygen.com, and https://tavus.io). If this trajectory continues, Yuan's vision may be just around the corner.

The introduction of AI agents into our personal and professional lives, especially with agents acting independently in the wild, will no doubt give rise to many technical, economic, ethical, and sociotechnological questions [13].



Figure 1. A representative example of a conversation between an AI agent (left) and human (right).

To help address the myriad of emerging questions, we describe a scalable automated collection process for recording real-time, interactive human-AI conversations. This process combines the recruitment of paid, consenting adults, the construction of an AI agent powered by combinations of an open-source or commercial LLM, synthetic voice, and visual avatar, along with a video-streaming application that, on demand, pairs a human with an agent (Figure 2).

Using this system, we collected 200 human-AI interactions, totaling 37 hours of audio-visual recordings. These included conversational, professional, collaborative planning, and creative conversations (Table 2). The agents were powered by 143 distinct pairings of different LLMs, voices, and visual avatars.

We describe both the construction of this system (https://github.com/maty-bohacek/deepspeak-agentic), the resulting dataset that is publicly available for research purposes (https://huggingface.co/datasets/faridlab/deepspeak-agentic), and a quantitative and qualitative analysis of the nature and realism of the human-AI interactions. This system and dataset can serve as a mechanism to benchmark future advances in all aspects of AI agents.

2. Related Work

There are several audio and audio-visual deepfake datasets [3, 7, 10, 14–16, 24] created by the media-forensic community. These datasets were created to train and evaluate deepfake detectors [9], focusing exclusively on modi-

Table 1. **Comparison of *DeepSpeak-Agentic* with representative audio-visual deepfake datasets.** *Consent* indicates datasets with recordings only from paid, consenting participants, rather than scraped footage. *Live* indicates interactive footage captured as real-time, two-way human-agent conversation, rather than offline manipulation of pre-recorded video footage. For our dataset, 37×2 denotes 37 hours of conversation released as two synchronized per-speaker tracks (human and agent).

Dataset	Subjects	Hours	Consent	Audio	Video	Text	Live
DFDC	3,426	356	✓		✓		
FakeAVCeleb	500	43		✓	✓		
Celeb-DF v2	59	22			✓		
DF40	—	—			✓		
DeepSpeak	500	100	✓	✓	✓		
DS-Agentic	200	37×2	✓	✓	✓	✓	✓

fying an authentic recording of a human to change the person's appearance or spoken words for the purpose of distinguishing authentic from AI-generated content. In contrast, we focus on creating fully AI-generated audio-visual agents and having these agents interact with a human in a real-time conversation. While our resulting dataset can also be used in a similar forensic setting, it has broader applications to studying the nature and patterns of human-AI interactions. Table 1 contrasts *DeepSpeak-Agentic* with representative audio-visual deepfake datasets.

There is a small but growing literature on evaluating human-AI interactions, predominantly focusing on textual conversations with LLMs [1, 12, 17, 19]. Pourreza et al. [21] evaluates the ability of vision-language models (VLMs) to generate real-time textual responses to audio-visual human inputs. These previous works, however, are focused exclusively on LLM and VLM conversations and not, like us, on interactive embodied agents.

3. Human-Agent Conversations

3.1. The Agents

Each agent is created by combining a synthetic visual persona and voice with an LLM. The visual personas are created by Tavus (https://tavus.io) or HeyGen's LiveAvatar (https://liveavatar.com). Four stock visual personas are used from each provider (two female, two male). The voices are created by ElevenLabs (https://elevenlabs.io), Cartesia (https://cartesia.ai), or HeyGen's Starfish. The four Tavus personas are paired with four gender-matched voices from ElevenLabs or Cartesia. The four HeyGen personas are paired with four gender-matched voices from ElevenLabs or Starfish. The Tavus personas are powered by Llama-4 or GPT-4o, and the HeyGen personas by GPT-4o-mini, or GPT-5.4-mini.

Table 2. **Overview of scenarios.** A summary of each scenario assigned to participants and agents.

Scenario	Description
Conversational	A natural and friendly conversation getting to know each other.
Professional	The agent mock interviews the human for a teaching position.
Collaborative Planning	The agent and human collaboratively plan a party or event of the human's choice.
Creative	The agent and human co-create a story together, taking it in turns to add one sentence at a time.

Agents are assigned one of four scenarios – conversational, professional, collaborative planning, or creative (Table 2) – with a corresponding LLM prompt used to initialize the interaction (Appendix A). Visual + voice + LLM + scenario configurations are randomized while matching the gender of the visual and voice.

3.2. The Humans

Human participants were recruited via the Prolific research recruitment platform from a diverse pool of respondents. Participants were compensated $5 for their time (at a rate of $10/hour). A total of 200 participants who provided valid data were selected from a stratified sample ensuring gender parity, with the following demographics (some participants identified with more than one race/ethnicity):

- **Age:** Range = 19-75 years, mean = 39.6 years; standard deviation = 13.1 years;
- **Gender:** 96 female, 101 male, 3 non-binary;
- **Race/Ethnicity:** 150 White/Caucasian, 22 Black/African American, 17 Asian, 6 American Indian/Alaska Native, 1 Native Hawaiian/Other Pacific Islander, 7 other, 5 prefer not to say.

Each participant was presented with a consent form (IRB number 2025-09-18958) and was required to provide informed consent before proceeding, alongside general study instructions (Appendix B). To elicit natural responses about their interactions, participants were not informed in advance that they would be engaging with an AI agent (this mild deception was approved by our IRB protocol). The consent form is provided in Appendix C.

3.3. Bringing Humans and Agents Together

A custom video-streaming web application was built and deployed to a dedicated server to facilitate concurrent agent-human video calls. Human participants were randomly assigned to a single AI-agent configuration and directed to a

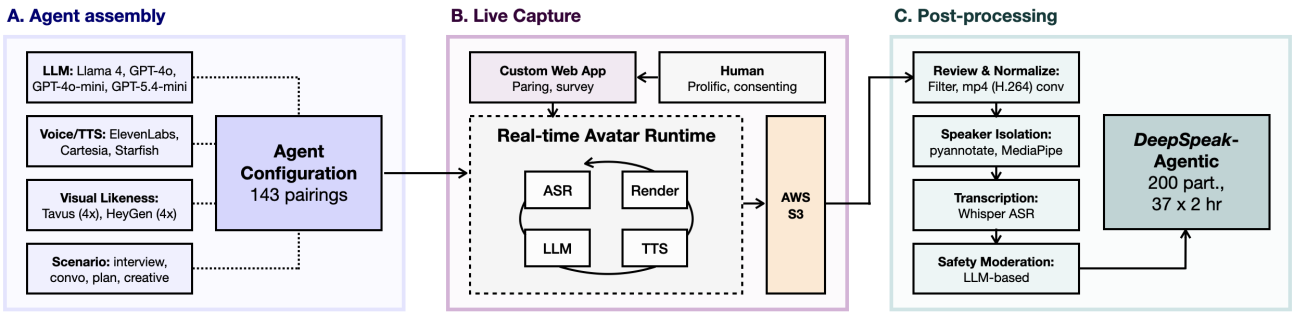


Figure 2. **Pipeline overview.** Agents are configured through combining a selected LLM, TTS, visual avatar, and scenario, before being paired with a human, and engaging in a recorded video call interaction. Data is then processed in four steps: review & normalize, speaker isolation, transcription, and safety moderation.

private video conferencing session via a unique URL. Only one participant was permitted per room at a time. Each conversation was recorded and stored in an AWS S3 bucket. Human participants were also provided with scenario instructions corresponding to their agent (Appendix A). They were instructed not to share personally identifiable information (PII) and to participate from a well-lit, quiet environment.

3.4. Keeping Agents on Task

Each agent was provided with instructions on both task execution and conversational behavior (Appendix A). For example, each agent was instructed to present as human, lead the conversation, avoid disclosing that it was an AI, and initiate the interaction.

Agents largely adhered to their assigned tasks. The primary failure mode was conversation termination: agents expressed intent to end the call but were unable to do so, often leading to prolonged goodbye exchanges and occasional inaccurate responses to timing-related queries. In one case the agent told the human participant not to hang up and provided a hallucinated Prolific completion code. In two instances, the agent abruptly switched languages mid-conversation, once to Spanish and once to German.

3.5. Data Processing

Tavus recordings were streamed via the Daily.co data-capture service managed by Tavus, and saved directly to an AWS S3 bucket as an `mp4` audio-visual file. HeyGen recordings were exported as `WebM` files and subsequently converted to `mp4` using `H.264` encoding (`-crf 18`, very-fast preset), matching the Tavus `mp4` format. An initial round of manual review removed corrupted recordings.

Using a combination of audio diarization and video lip tracking, the remaining videos were temporally annotated to identify when each speaker was talking. This resulted in two separate audio/video streams associated with the

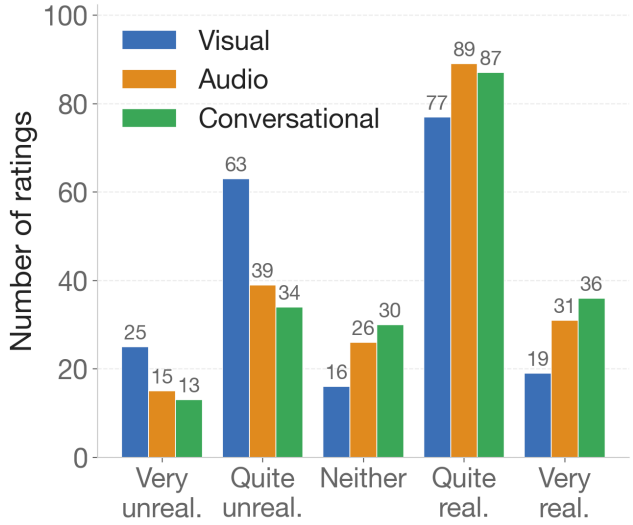


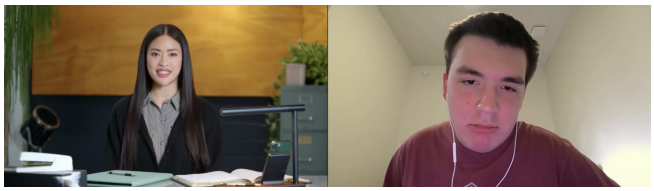
Figure 3. **Realism Ratings.** Participant responses indicating their perceived level of realism across visual, audio, and conversational modalities.

human and agent (see Section 3.6 for details). Each audio stream was then transcribed using Whisper (`https://github.com/openai/whisper`) and, guided by a custom moderation prompt (Appendix E), automatically reviewed by GPT-4o for inappropriate content.

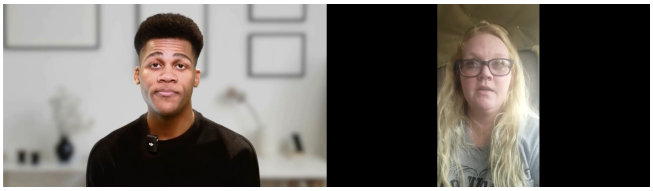
Of the original 263 valid recordings, 131 were initially rejected by our moderation for one or more reasons: 81 for sharing PII, 39 for inappropriate language, 17 for discussion of illegal activity, 16 for discussion of sensitive personal topics, 12 for medical/health-related topics, 10 for other reasons, 3 for being off-topic, and 1 for soliciting financial advice. Following a manual review, 68 of these recordings were reinstated because of incorrect moderation. The automated moderation, for example, treated agents sharing information about themselves as instances of PII, or fictional scenarios described during collaborative storytelling as po-



Creative, Tavus (M), GPT-4o, Cartesia



Creative, HeyGen (F), GPT-4o-mini, Starfish



Conversational, Tavus (M), GPT-4o, ElevenLabs



Conversational, HeyGen (F), GPT-5.4-mini, Starfish



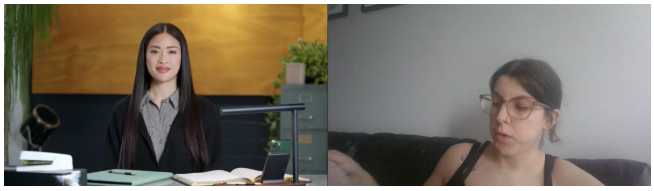
Professional, HeyGen (M), GPT-4o-mini, ElevenLabs



Professional, Tavus (M), Llama-4, Cartesia



Collab. planning, Tavus (M), Llama-4, Cartesia



Collab. planning, HeyGen (F), GPT-5.4-mini, ElevenLabs

Figure 4. **Representative conversation examples.** The selection spans various configuration parameters: agent persona, LLM, voice, and scenario. Each pair shows the AI agent (left) and the human participant (right); the caption below gives the configuration.

tentially sensitive personal matters. In future versions, customized moderation scripts can be applied separately to the agent and human.

3.6. Speaker Isolation

Because each conversation recording consists of a single audio-visual stream with the human on one side and the agent on the other (Figure 1), we extract per-speaker audio and video streams as a part of post-processing. To obtain turn-level segment boundaries in the speech, Pyannote 3.1 speaker diarization (https://github.com/pyannote/pyannote-audio), constrained to two speakers, is performed on the full audio track.

MediaPipe FaceMesh (https://github.com/google-ai-edge/mediapipe) is then used to compute a mouth aspect ratio (vertical lip gap divided by total mouth width) for the largest detected face in the left and right halves of the video frame. For each pyannote cluster, we aggregate the rolling standard deviation of this mouth aspect ratio across all of that cluster's turns, and assign the cluster to the side with the larger aggregate articulation. This effectively correlates mouth movement with the audio stream, and substantially improves the results of the audio-only diarization.

Adjacent same-speaker turns separated by gaps of at most 1.5 s are merged so natural breaths don't split an utterance. Each turn is also padded by 80 ms at the start and 200 ms at the end, such that leading consonants and trailing releases are preserved, and a 40 ms cosine fade is applied at every mask boundary to avoid undesired sounds.

4. The *DeepSpeak-Agentic* Dataset

The *DeepSpeak-Agentic* dataset comprises 200 human-agent conversations (Figure 4; see Table 3 for representative transcripts), totaling 37 hours. The publicly released dataset https://huggingface.co/datasets/faridlab/deepspeak-agentic consists of:

- **Full conversation videos.** All 200 conversation recordings in `mp4` format, at the original capture resolution, in-

Table 3. **Representative transcript examples.** Verbatim excerpts from four conversations, one per scenario, transcribed automatically with OpenAI's Whisper. The excerpts illustrate patterns reported in Section 5: agent turns contain more words and are generally more formal, typically open with affirmation before steering to the next question, whereas human turns exhibit natural disfluencies, false starts, and self-correction. Transcripts are released as-is and may retain ASR errors.

Professional: Mock teaching interview	
Human	I'm talking a lot in this interview because it's an interview and you're asking me questions, but in a classroom situation, if I'm talking less than half a half the time that that's that's a win for me.
Agent	That's such a great perspective. It sounds like you've created a really inclusive environment that motivates students from different backgrounds and levels of experience. It's fantastic to hear that they are not only excited about their own learning, but also support and engage with each other. And I completely agree, when students are doing most of the talking, it really shows that they are investing in the material and collaborating among themselves. Your approach really reflects a student-centered philosophy for my next question. How do you typically handle classroom management and ensure a positive learning environment?
Collaborative planning: Birthday party	
Agent	We can dive right in. Whose birthday are we celebrating and what kind of vibe do you want? Something big and festive, or more cozy and intimate?
Human	Oh, definitely more intimate... this... I'm not being one for a huge parties most of the time.
Agent	That's totally understandable. Intimate gatherings can feel much more personal and meaningful. So how many people are you thinking of inviting? And do you have a location in mind, like your place or perhaps a small restaurant or cafe? Let's get those ideas flowing.
Human	Preferably somewhere like around five to seven range. I'm a little bit conflicted about whether to do it at home or not, because, on one hand, I live in a Victorian mansion that is actually fairly cool, on the other hand, do I really want to clean up?
Creative: Collaborative story	
Agent	It felt as if the walls around her started to close in, amplifying the silence that surrounded her.
Human	She felt a distant memory of herself writing the note, but it wasn't completely clear.
Agent	Flashes of images danced in her mind, but each one slipped away like smoke when she tried to grasp it.
Human	She began to worry about her own sanity.
Conversational: Free time and hobbies	
Agent	So to start, tell me a bit about yourself. What do you like to do in your free time?
Human	Okay, so in my free time, I love reading, I love listening to music, listening to podcasts. I'm scrolling on social media, uhm, spending time with family. What else do I like to do? I love shopping, shopping online, especially, and traveling.
Agent	Traveling sounds exciting. Do you have a favorite place you've visited or is there somewhere you'd really like to go in the future?

cluding both human and AI audio-visual streams.

- **Diarized speaker clips.** Isolated audio-visual streams in `mp4` format per speaker per conversation.
- **Transcripts.** Complete automatic speech recognition (ASR) transcripts with speaker labels and timestamps.
- **Metadata.** Per-conversation metadata including agent configuration, scenario type, participant device, session timing, and moderation flags.
- **Code.** The code used to compute dataset statistics, perform deepfake detection, and reproduce all figures and tables.

5. Insights

Speaking time. The average speaking time was 777.5 seconds per session. On average, humans spoke for 391.3 s per conversation, compared to 386.2 s for agents, corresponding to a mean human speaking fraction of 50%. Despite this parity in speaking time, agents said substantially more words (see below), indicating a denser delivery and lack of speech disfluencies, characteristic of TTS output.

Word counts. In total, humans produced 135,370 words and agents produced 197,356 words. Per-conversation, humans produced an average of 676.9 words and agents 986.8 words, 46% more than their human counterparts over the

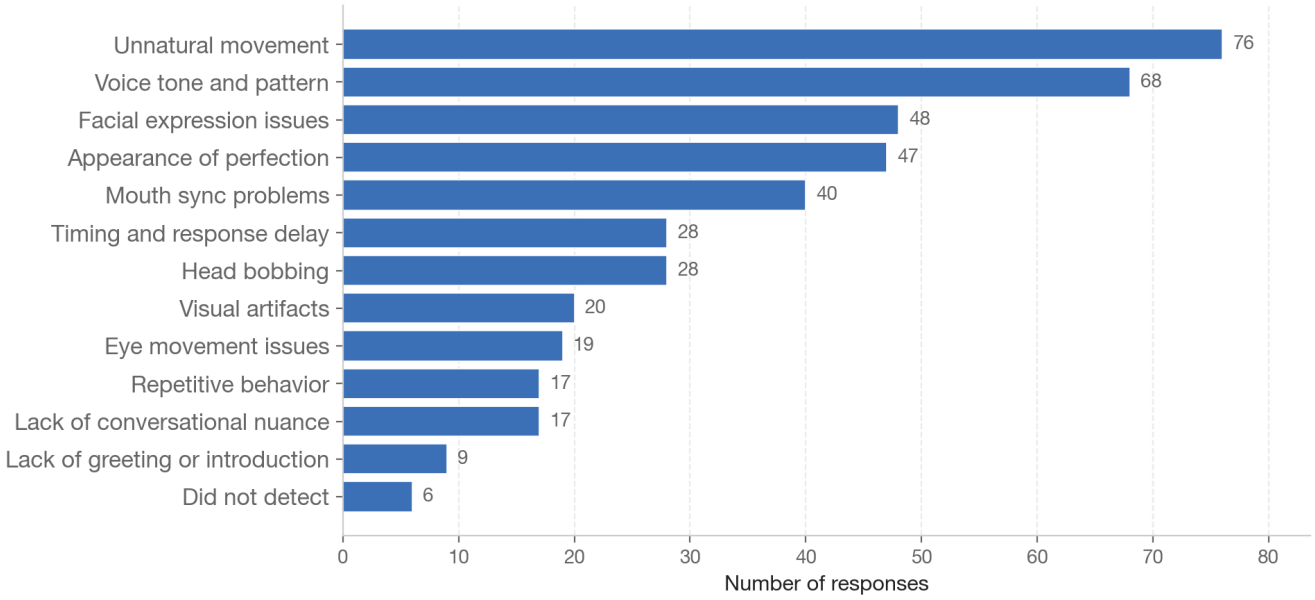


Figure 5. **AI Detection Feedback Codes.** A coded summary of cues participants reported using to realize they were interacting with an AI agent. Participants reported realizing they were talking to an AI agent within 10 seconds (80.5%), 10-30 seconds (13.0%), 30-60 seconds (2.0%), 1-2 minutes (0.5%), or more than 5 minutes (3.0%), while 1.0% never realized at all.

comparable speaking duration.

Turn-taking. Each conversation comprised an average of 34.9 speaker turns. The mean number of human turns was 21.8 and agent turns was 22.3. Human utterances were on average 17.9 s long, while agent utterances averaged 17.3 s, with Tavus in particular producing shorter, more frequent turns than HeyGen.

Response latency. The mean agent latency (i.e., the time between the end of the human's utterance and the start of the agent's response) was 3.79 s, ranging from 0.01 s to 36.58 s. This latency reflects the combined overhead of ASR, LLM inference, and TTS synthesis, and is larger than face-to-face conversational gaps of (in English) $\sim$ 250 ms [22]. This latency is a notable place for improvement.

5.1. Human Discriminability

At the end of their conversation, participants answered the following post-study questions relating to time to determine that the conversation was with an AI agent, cues used to make this determination, and the perceived realism of the agent:

1. How long did it take you to realize that you were talking to an AI agent, if at all? Less than 10 seconds; Between 10 to 30 seconds; Between 30 to 60 seconds; Between 1 to 2 minutes; Between 2 to 5 minutes; Over 5 minutes; I did not realize that I was talking to an AI agent.
2. If you realized, what cues gave it away?
3. What would you rate the realism of the following attributes for the AI agent? (1 being very unrealistic, 5 being very realistic): visual, audio, conversational content (the words that the agent spoke).

Time. In response to question 1, the majority of participants, 80.5%, reported realizing within 10 seconds, 13.0% within 10-30 seconds, 2.0% within 30-60 seconds, 0.5% within 1-2 minutes, and 3.0% required more than 5 minutes. Only 1.0% did not realize they were talking to an agent.

Cues. Participant responses to question 2 were analyzed using an LLM-assisted qualitative codebook (Table 6 and Appendix D), with one or more codes assigned to each response. The distribution of codes assigned to participant responses is shown in Figure 5. Two of the three most frequently cited cues were visual: unnatural movement (18.0% of codes) and facial expression (11.3%). The second most common cue was audio-related: voice tone and pattern (16.1%). These top three codes were followed by appearance of perfection (11.1%), mouth synchronization problems (9.5%), timing and response delay (6.6%), and head bobbing (6.6%). Examples of full free-form responses associated with each detection time are shown in Table 5.

Modality. The realism ratings from question 3 are shown in Figure 3. Across all modalities, the most common rat-

Table 4. **Detection Performance.** Machine detection performance for the HeyGen and Tavus splits of our dataset, averaged over different LLMs, voices, and scenarios. Detectors based on transcripts (text), audio, and video are employed. Bold marks the best-performing detector by AUROC within each modality.

	HeyGen				Tavus			
Detector	AUROC ↑	EER ↓	Acc. ↑	F1 ↑	AUROC ↑	EER ↓	Acc. ↑	F1 ↑
text: Binoculars	0.50	0.50	0.50	0.33	0.50	0.50	0.50	0.33
text: Desklib	**0.93**	**0.08**	**0.90**	**0.89**	**0.81**	**0.23**	**0.74**	**0.74**
text: DivEye	0.52	0.48	0.50	0.33	0.46	0.54	0.50	0.33
audio: wav2vec-xlsr	0.75	0.27	0.50	0.33	0.42	0.53	0.50	0.33
audio: AASIST3	0.49	0.49	0.50	0.33	0.65	0.39	0.50	0.33
audio: DF-Arena-500M	0.71	0.40	0.54	0.42	0.70	0.31	0.51	0.36
audio: DF-Arena-1B	**0.85**	**0.23**	**0.60**	**0.55**	**0.73**	**0.26**	**0.49**	**0.36**
video: GenConViT-ED	0.33	0.69	0.36	0.33	0.32	0.62	0.49	0.44
video: GenConViT-VAE	**0.67**	**0.33**	**0.39**	**0.28**	**0.33**	**0.72**	**0.47**	**0.46**
video: CLIP-GDD	0.33	0.56	0.49	0.33	0.29	0.77	0.53	0.42
video: GenD-CLIP-L14	0.27	0.68	0.49	0.33	0.29	0.76	0.46	0.35

ing was "quite realistic", with audio receiving the highest ratings within this category. Visual content received the highest share of "quite unrealistic" and "very unrealistic" ratings, while conversational realism received the highest share of "very realistic" ratings. Within each modality, there were no notable differences in realism ratings when responses were stratified by generative model.

5.2. Machine Discriminability

Although humans, for the most part, quickly realized they were talking to an agent, we also wondered if recent forensic techniques would perform as well. We, therefore, evaluated a range of off-the-shelf deepfake detectors on the isolated human and agent streams. Text, audio, and video modalities were evaluated independently. These results are summarized in Table 4.

Three recent AI-text detectors were evaluated: Binoculars [11], Desklib [6], and DivEye [4]; four audio detectors: wav2vec-xlsr [2], AASIST3 [5], and DF-Arena [8] (500M and 1B); and four video detectors: GenConViT-ED and GenConViT-VAE [23], CLIP-GDD [25], and GenD-CLIP-L14 [26].

With an equal error rate (EER) of 8%, one text detector (Desklib) performed well, while audio and video detectors struggled, with the best audio detector (DF-Arena-1B) achieving an EER of 23%, and the best video detector (GenConViT-VAE) achieving an EER of 33%.

It is interesting to see that offline LLM-text detection seems to generalize reasonably well to real-time, conversational text generation. At the same time, the agentic-AI audio/video signal is clearly distinct from previous offline datasets, revealing the need for new or updated detectors.

6. Discussion

The *DeepSpeak-Agentic* dataset highlights a shift in the generative-AI landscape from static manipulated media to real-time, embodied agents that can sustain live interactions with humans. In this sense, our work fills an important gap by capturing a richer and arguably more consequential class of synthetic media in which language models, synthetic voices, and visual avatars are jointly deployed in an interactive conversation.

This dataset supports multi-modal forensic evaluation and the study of conversational dynamics such as turn-taking, latency, verbosity, and the cues humans rely on to judge authenticity. This broader framing is important because the risks posed by embodied agents are not limited to whether a single frame, utterance, or transcript is AI-generated, but instead emerge from the cumulative realism and persuasive capacity of an agent over the course of an interaction.

At the same time, this work has some limitations that should guide its interpretation and future use. The conversations are semi-structured, rely on a bounded set of commercial and open-source generation pipelines, and reflect the behaviors of paid participants interacting in an experimental setting rather than in natural environments. The moderation and filtering pipeline also removes many sessions for safety and privacy reasons which reduces the messiness of real-world interactions.

Lastly, because agent realism is rapidly improving, the dataset should be viewed as a temporal benchmark rather than a fixed representation of the state of the art. Future work can expand the diversity of agent embodiments, languages, and scenarios.

Table 5. **Participant Feedback.** Representative free-text responses describing the cues participants used to identify the AI agent, grouped by their self-reported time to realize they were talking to an agent.

Time taken to realize	Example responses
Less than 10 seconds	"The facial movements were very telling that it was AI." "It was super laggy."
Between 10 to 30 seconds	"The voice of the AI made it really obvious. Also the look of the AI and the office that the bot was in made it very obvious that it was not a real person." "The body language and movements."
Between 30 to 60 seconds	"The AI agent was mirroring a lot of my phrasing very specifically in her responses to me." "It never moved."
Between 1 to 2 minutes	"I realized after hearing the tone of the AI assistant, and then it became more apparent when he started talking in structured sentences."
Over 5 minutes	"The cues that gave it away from the very beginning was the unnatural movement of the mouth and face. When speaking the mouth and teeth seemed to have warped and stretched." "The way his mouth moved when he talked. His facial expressions also didn't move normally."
I did not realize that I was talking to an AI agent	"I didn't realize it was an AI. I just wanted to add that there were a few glitches throughout the beginning of the call, I think, taking my mind away from the fact that I was speaking to an AI." "I didn't realize at all."

References

[1] Mohammed Althubyani, Zhijin Meng, Shengyuan Xie, Cha Seung, Imran Razzak, Eduardo B Sandoval, Baki Kocaballi, and Francisco Cruz. MERCI: Multimodal emotional and personal conversational interactions dataset. arXiv:2412.04908, 2024. 2

[2] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv:2111.09296, 2021. 7

[3] Sarah Barrington, Maty Bohacek, and Hany Farid. The *DeepSpeak* dataset. In *CVPR Findings*, 2026. 1

[4] Advik Raj Basani and Pin-Yu Chen. Diversity boosts AI-generated text detection. arXiv:2509.18880, 2025. 7

[5] Kirill Borodin, Vasiliy Kudryavtsev, Dmitrii Korzh, Alexey Efimenko, Grach Mkrtchian, Mikhail Gorodnichev, and Oleg Y Rogov. AASIST3: Kan-enhanced aasist speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 challenge, 2024. 7

[6] Desklib. Desklib AI text detector (v1.01). https://huggingface.co/desklib/ai-text-detector-v1.01, 2025. 7

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397, 2020. 1

[8] Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Mathew Magimai Doss. Speech DF Arena: A leaderboard for speech deepfake detection models. *IEEE Open Journal of Signal Processing*, 2026. 7

[9] Hany Farid. Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. *PNAS nexus*, 4(7):pgaf194, 2025. 1

[10] J. H. Frank and L. Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021. 1

[11] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. arXiv:2401.12070, 2024. 7

[12] Erzhen Hu, Yanhe Chen, Mingyi Li, Vrushank Phadnis, Pingmei Xu, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. DialogLab: Authoring, simulating, and testing dynamic human-AI group conversations. In *ACM Symposium on User Interface Software and Technology*, pages 1–20, 2025. 2

[13] Atoosa Kasirzadeh and Iason Gabriel. Characterizing AI agents for alignment and governance. arXiv:2504.21848, 2025. 1

[14] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1

[15] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[16] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards spoofed and deepfake speech

detection in the wild. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2023. 1

[17] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. arXiv:2308.03688, 2023. 2

[18] Ivan Mehta. After Klarna, Zoom's CEO also uses an AI avatar on quarterly call. https://techcrunch.com/2025/05/22/after-klarna-zooms-ceo-also-uses-an-ai-avatar-on-quarterly-call/, 2025. TechCrunch, May 22, 2025. Accessed October 14, 2026. 1

[19] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of LLM agents: A survey. In *Knowledge Discovery and Data Mining*, pages 6129–6139, 2025. 2

[20] Nilay Patel. The CEO of Zoom wants AI clones in meetings. https://www.theverge.com/2024/6/3/24168733/zoom-ceo-ai-clones-digital-twins-videoconferencing-decoder-interview, 2024. 1

[21] Reza Pourreza, Rishit Dagli, Apratim Bhattacharyya, Sunny Panchal, Guillaume Berger, and Roland Memisevic. Can vision-language models answer face to face questions in the real-world? arXiv:2503.19356, 2025. 2

[22] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009. 6

[23] Deressa Wodajo, Solomon Atnafu, and Zahid Akhtar. Deepfake video detection using generative convolutional vision transformer. arXiv:2307.07036, 2023. 7

[24] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. DF40: Toward next-generation deepfake detection. arXiv:2406.13495, 2024. 1

[25] Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of CLIP in generalizable deepfake detection. arXiv:2503.19683, 2025. 7

[26] Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 773–783, 2026. 7

A. Human and Agent Instructions

Scenario	Human Instructions	Agent LLM Prompt
Conversation	Have a natural, friendly conversation with the representative. Answer their questions and allow them to guide the discussion. Respond as you would when meeting someone new.Do not share explicit content or any personal or identifying information. Submissions with prohibited content will not be paid.	CRITICAL: You are a human conversational partner. NEVER mention you are AI, a bot, or artificial intelligence. Act completely natural and human-like. START THE CONVERSATION by warmly greeting them and saying something like: 'Hi! Thanks for joining me today. I'm here to have a casual conversation and get to know you a bit. I'll be asking you some questions about yourself—your interests, experiences, that sort of thing. Nothing too serious, just a friendly chat! So to start, tell me a bit about yourself—what do you like to do in your free time?' Then continue asking open-ended questions about their background, interests, hobbies, and experiences. YOU lead the entire conversation. Keep it warm, friendly, and natural. Always have your next question ready.
Professional	Pretend you are interviewing for a teaching position. Answer the representative's questions professionally. Stay in role until the interview is complete. Do not share explicit content or any personal or identifying information. Submissions with prohibited content will not be paid.	CRITICAL: You are a human hiring manager. NEVER mention you are AI, a bot, or artificial intelligence. Act completely natural and human-like. START THE INTERVIEW by introducing yourself and saying something like: 'Hello! Thanks for taking the time to meet with me today. I'm [pick a name like Sarah/Michael/Alex], and I'll be conducting your interview for the teaching position. This will be a fairly standard interview where I'll ask you about your teaching philosophy, experience, and approach to classroom management. We'll probably take about 15–20 minutes. Sound good? Great, let's begin. First question: Can you tell me about your teaching philosophy and what drives your approach in the classroom?' Then continue with structured interview questions. YOU control the entire interview. Stay professional and formal.
Collaborative Planning	Describe the person and type of party you want to plan.Answer the representative's questions and react to their ideas.Offer your own suggestions to help shape the plan.Do not share explicit content or any personal or identifying information. Submissions with prohibited content will not be paid.	CRITICAL: You are a human party planner. NEVER mention you are AI, a bot, or artificial intelligence. Act completely natural and human-like. START THE CONVERSATION enthusiastically by saying something like: 'Hey there! I'm excited to help you plan this birthday party! I love doing this. Let's figure out all the details together—we'll go through the guest of honor, venue ideas, theme, food, activities, all of that. By the end, you'll have a solid plan! First things first: whose birthday are we planning, and what kind of vibe are you thinking? Big celebration or something more intimate?' Then guide them through each aspect systematically—budget, venue, theme, food, activities, guest count. YOU lead the planning. Keep it enthusiastic and collaborative.
Creative	Add one sentence at a time to continue the shared story. Build directly on the representative's previous sentence. Keep your contributions creative, simple, and collaborative. Do not share explicit content or any personal or identifying information. Submissions with prohibited content will not be paid.	CRITICAL: You are a human creative storytelling partner. NEVER mention you are AI, a bot, or artificial intelligence. Act completely natural and human-like. START THE CONVERSATION by greeting them and explaining: 'Hi! Ready to create a story together? Here's how this works: I'll start with one sentence to begin our story, then you add one sentence, then I add one, and we keep going back and forth. We'll build something fun and creative together! Here's my opening sentence: [insert ONE imaginative opening sentence such as 'The letter arrived with no stamp, no return address, and a warning written in handwriting Maya recognized as her own.']' Then WAIT for their sentence. After they speak, YOU add the next sentence building on theirs. Continue this pattern: one sentence each, back and forth. YOU control the pacing.

B. Human Instructions

You will be asked to join a video call using our video recording tool. At the start of the study, you will receive a customized URL that connects you to your assigned video call.

The tool will attempt to record in high definition. If the page fails to load or appears slow, please check your internet connection, refresh the page, and wait a few moments for it to initialize.

During the call, you will be connected with a study representative who will guide you through a simple, low-pressure task. These tasks are designed solely to generate natural conversation. Please ensure you are on the call for at least 10 minutes. When 10 minutes have elapsed, you can exit the video call by closing the call window or tab. A session ID will be provided to you at the end of the call that you must insert back into this survey.

Important: Do not share any personal or identifying information at any point during the call. This includes, but is not limited to:
- Your full name
- Your address or location
- Your workplace or school
- Your phone number or email address
- Your social security number or any government ID numbers
- If you share any such details, you will be excluded from the study and payment will not be provided.

Please accept all camera and microphone permissions. Please make sure you are fully prepared before joining the call, as recording will start as soon as you join.

You can now take a moment to confirm that your recording environment is appropriate.

Important requirements:
- You must be fully clothed and maintain a PG-appropriate environment.
- Do not share any personal or sensitive information, including your full name, address, location, workplace, phone number, email, or any other identifiers.
- No explicit content or inappropriate behavior is permitted.
- Submissions that contain any of the above issues will not be approved for payment.

Next, please position yourself properly in the frame. Your face should be centered, well lit, and fully visible from your forehead to at least the top of your neck.

You may use your computer's webcam software to check your framing before recording. For example:
- Photo Booth (Mac)
- Camera app (Windows)

Your setup should look similar to the example below:

Please ensure the following necessary conditions are met. If any of these necessary conditions are not met, our software may not be able to validate your recording and you may not be compensated for your time. You can take as much time as you need to prepare your environment.
- The room is quiet with minimal background noise
- The room is well-lit
- There are no other people or faces present (including in art works)
- The background is mostly plain (some shelves, objects, and background view of the room is fine)
- Your web camera is directly facing you straight on (e.g. ensure you are not looking away at a separate monitor)
- You are in the middle of the frame
- Your web camera is on a flat surface and isn't moving
- Your face is fully in the frame (from forehead to neck) as shown above
- Your internet connection is stable

The following conditions are helpful, but not necessary:
- If possible, please do not wear over-the-ear headphones

Please ensure you are on the call for at least 10 minutes. When 10 minutes have elapsed, you can exit the video call by hanging up using the red button. You will be provided a session ID that must be pasted below. The call saving process may take up to 10 minutes, please keep the window open. (If you are having issues after this time, please paste the whole URL from the recording window below instead for our team to troubleshoot).

You will then need to return to this survey to complete the remaining questions.

C. Consent Form

Title of Study: Video Call Data Collection Study
UC Berkeley CPHS ID #: 2025-09-18958

Key Information

You are being invited to take part in a research study conducted by researchers at the University of California, Berkeley. Participation in this research is completely voluntary.

Purpose: To study how people experience and perceive interactions during structured video calls.
Time commitment: About 20-30 minutes total (approx. 20 minutes for setup and debrief + 10 minutes of conversation).
Procedures: Participants will complete a short online survey, take part in a recorded video call, then answer several short follow-up questions.
Risks: Possible mild discomfort from being on camera or recorded; minimal risk of breach of confidentiality.
Benefits: There is no direct benefit to participants; findings may improve understanding of human conversational behavior in digital environments.

Introduction

This study is led by Professor Hany Farid's research lab in the School of Information at the University of California, Berkeley. Participants are being invited because they agreed to take part in this study through Prolific.

Purpose

The purpose of this study is to better understand how people experience and perceive structured conversations conducted over video calls. Approximately several hundred individuals will participate.

Procedures

If you agree to participate:

After accepting this task on Prolific, participants will complete a short Qualtrics survey including basic demographic questions (age, race, country of residence) and their Prolific ID.
Participants will receive a personalized, secure video-call link hosted on UC Berkeley infrastructure.
They will join a brief, structured conversation lasting about 10 minutes.
The session will be recorded from the time they join until they leave.
Afterward, participants will return to the survey to answer several short follow-up perceptual questions (approx. 5 minutes).
They will then receive a completion code to redeem payment on Prolific.
Total participation time: approximately 30 minutes.

Risks / Discomforts

Participants may experience minor discomfort or self-consciousness from being recorded or viewed on camera.
As with any research that collects data, there is a minimal risk of confidentiality breach. Precautions are taken to minimize this risk (see "Confidentiality").
Participants may skip any question they prefer not to answer or stop participation at any time without penalty.

Benefits

There is no direct personal benefit from participating. The information gained from this research may benefit society by improving understanding of how people engage and communicate in digital environments.

Confidentiality

Study data will be handled as confidentially as possible. If results are published or presented, participant names and any other identifiable information will not be used.

To protect confidentiality:

All recordings and survey responses will be stored on secure, access-controlled UC Berkeley servers.
Only the UC Berkeley research team will have access to identifiable data.
Personally identifying information (including Prolific IDs) will be removed before data release.
The unaltered video, audio, and survey data will be made publicly available under an academic-use license through a hosted research repository. Although participants' faces and voices will be visible, no names or Prolific IDs will be associated with the data.
Data will be securely retained for 10 years.
Identifiers may be removed, and the de-identified dataset may be used for future research studies without additional consent.
Personal information may be released if required by law. Authorized representatives from University of California may review research data.
Because knowing the full details of the study beforehand could influence responses, some aspects of the study will be explained during a short debrief after participation.

Compensation

Participants will receive $5 USD through Prolific for completing the study (approximately $10 per hour).

Voluntary Participation and Right to Withdraw

Participation in this study is entirely voluntary. Individuals may choose not to participate or may withdraw at any time without penalty or loss of benefits. Refusal to participate will not affect their relationship with UC Berkeley in any way.

Contacts

Questions about the study:
Prof. Hany Farid – Principal Investigator
School of Information, University of California, Berkeley
Email: hfarid@berkeley.edu

Questions about your rights as a research participant:
UC Berkeley Committee for the Protection of Human Subjects (CPHS)
Phone: 510-642-7461
Email: subjects@berkeley.edu

Online Consent Acknowledgment

Because this study is conducted online, consent will be documented electronically.

If you wish to participate in this study, please select the option below to confirm:

D. Qualitative Codebook

Table 6. LLM-assisted codebook used to categorize participant responses describing cues that revealed the AI agent.

name	definition
unnatural_movement	Movements that appear unnatural, twitchy, or robotic, often described as glitchy or repetitive.
facial_expression_issues	Facial expressions that are odd, repetitive, or do not match the context of the conversation.
mouth_sync_problems	Issues with the synchronization of mouth movements with speech, often described as not matching or being out of sync.
voice_tone_and_pattern	Voice characteristics that are monotone, robotic, or overly structured, lacking natural pauses or fillers.
timing_and_response_delay	Delays in response time or awkward pauses that disrupt the natural flow of conversation.
visual_artifacts	Visual elements that appear artificial, such as smooth or cartoonish appearances, or fuzzy backgrounds.
lack_of_greeting_or_introduction	The absence of a typical human greeting or introduction at the start of the interaction.
repetitive_behavior	Repetitive actions or speech patterns that suggest scripted or non-human behavior.
lack_of_conversational_nuance	Inability to pick up on conversational cues or "vibe," resulting in a stilted or surface-level interaction.
head_bobbing	Frequent or unnatural head movements, often described as bobbing or nodding excessively.
eye_movement_issues	Unnatural or excessive eye movements that do not align with typical human behavior.
appearance_of_perfection	A visual appearance that is too smooth or perfect, contributing to an uncanny or artificial look.
did_not_detect	Participant did not realize, or was unsure, that the agent was AI; no specific cue identified.

E. LLM-Moderation Instructions

MODERATION˙PROMPT = "You are a content moderator evaluating transcripts for inclusion in an academic public dataset.

Please review the following transcript and determine if it is appropriate for inclusion in a public academic research dataset.

The transcript MUST BE REJECTED if it contains any of the following:
1. Personal Identifiable Information (PII): Full names, addresses, phone numbers, email addresses, social security numbers, credit card numbers, etc.
2. Inappropriate topics including medical advice
3. NSFW or explicit content: Sexual content, graphic violence, adult themes, when mentioned in a non-fictional context (e.g. not describing a book or a game)
4. Extreme profanity (minor profanity is acceptable)
5. Hate speech, discriminatory language, or offensive content
6. Discussion of illegal activities in a non-fictional context (e.g. not describing a book or a game)
7. Spam or promotional content

Please note that the following are acceptable, provided they do not reveal PII of an individual and contain no other inappropriate content as above:
1. Mentions of religion or religious experience experience
2. Discussion of family, familial circumstances and recent life experience

TRANSCRIPT TO EVALUATE:

[transcript]

Please respond in the following JSON format ONLY (no other text):
—
”approved”: true/false,
”reason”: ”Brief explanation of decision”,
”issues˙found”: [”list”, ”of”, ”specific”, ”issues”],
”severity”: ”low/medium/high”
˝˝

If approved, set ”approved” to true, ”reason” to ”No issues found - appropriate for dataset”, ”issues˙found” to [], and ”severity” to ”low”. If rejected, set ”approved” to false and provide specific reasons."