# Situated Real-time Interaction with a Virtually Embodied Avatar


Sunny Panchal
Qualcomm AI Research†
sunnpanc@qti.qualcomm.com

Guillaume Berger
Qualcomm AI Research†
guilberg@qti.qualcomm.com

Antoine Mercier
Qualcomm AI Research†
amercier@qti.qualcomm.com

Cornelius Böhm
Aignostics GmbH‡
cornelius.vonrekowski@gmail.com

Florian Dietrichkeit
LifeBonus‡
florian.dietrichkeit@lifebonus.health

Xuanlin Li
Qualcomm AI Research†
xuanlinl@qti.qualcomm.com

Reza Pourreza
Qualcomm AI Research†
pourreza@qti.qualcomm.com

Pulkit Madan
Qualcomm AI Research†
pmadan@qti.qualcomm.com

Apratim Bhattacharyya
Qualcomm AI Research†
aprabhat@qti.qualcomm.com

Mingu Lee
Qualcomm AI Research†
mingul@qti.qualcomm.com

Mark Todorovich
Qualcomm AI Research†
mtodorov@qti.qualcomm.com

Ingo Bax
Qualcomm AI Research†
ibax@qti.qualcomm.com

Roland Memisevic
Qualcomm AI Research†
rmemisev@qti.qualcomm.com


## 1. Introduction

A well-known shortcoming of generative language models is that they can generate language which, despite being syntactically and semantically sound, is not grounded in facts [2, 17]. A growing body of recent work has shown how combining language models (LM) with external information sources makes it possible to reduce such hallucinations by letting the model directly attend to external information [15, 18, 24]; an approach commonly referred to as *grounding*. A common approach to grounding is to monitor a model's output for the occurrence of certain syntactic patterns (such as the presence of agreed-upon tags) and to let an external source fill in information in the LM's stead, after which the LM continues its generation [5, 15, 17, 19]. A similar approach may be used to let an LM-based agent interact with an external environment by monitoring for, and executing, generated actions in the environment [4, 20]. These types of interactions between the LM and external systems are often enabled by LM augmentation with external plug-and-play modules, and orchestrators that coordinate LM prompts and generation [17].

This kind of external environment grounding is applicable in typical turn-based dialogue scenarios that iterate between user-posed questions and LM responses. In this scenario, individual turns are typically *blocking*, in that the interaction loop is frozen until the current participant finishes their turn. This blocking aspect continues to be a limiting factor of recent *agentic* instantiations of LMs where given a user instruction, the LM begins to execute a series of, often rigidly implemented, iterative planning and sub-goal execution steps [22].

Our motivation for studying LMs in the context of situated real-world interactions is two-fold: First, situating language in real-world interactions is a critical open problem and a long-standing goal of human-computer interac-

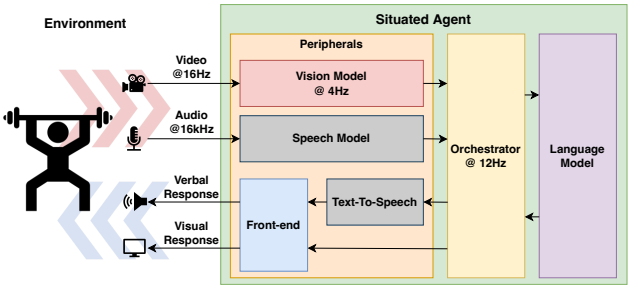


Figure 1. Overview of the state-augmented architecture

†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

‡Work performed at TwentyBN GmbH

tion. Second, the lack of grounding in real-time, sensory information constitutes a fundamental semantic gap between concept formation in language models versus humans, and is a likely culprit for LM hallucinations [13]. Bridging this gap is quite likely an essential step on the path towards general-purpose, human-like AI.

### 1.1. Situated communication testbed and baseline

We consider a human-AI interaction scenario in which real-time visual percepts of human activity, from a raw RGB camera input, are streamed continuously to an orchestrator-managed LM for online lingual response generation.

Despite the rapid recent progress, this is still a highly challenging scenario for existing models; multi-modal vision-based applications remain limited to turn-based VQA-style querying of LMs [1, 9–11]. In contrast, we explore vision-LM interaction as an online and dynamic situated dialogue. This interaction style necessitates a rich input stream with high information density that needs to be rapidly winnowed down to a manageable task-specific scope.

## 2. Multi-modal orchestrator

Reasoning over a real-time visual input stream of discrete image token embeddings, with the full stream preserved, would quickly bloat any LM's context window. Instead, we propose to use a stateful orchestrator that prompts the language model, when appropriate, with distilled task-pertinent visual information. An overview of the architecture, with an optional speech input, is shown in Figure 1.

Driven by activated triggers, the orchestrator calls the language model using contextualized prompts, that include (i) a task description, (ii) pertinent state information, (iii) interaction exemplars, (iv) a history of responses, and (v) a query representative of the activated triggers. Generated responses are then returned to the orchestrator as a candidate reaction to be considered for execution. Internal state management by the orchestrator lifts the burden of long-range consistency from the LM, allowing it to focus on the fusion and consolidation of information within its prompt. Additionally, we leverage the in-context learning abilities of LMs and demonstrate the usage of additional state context and multi-turn dialogue not present during training using a few-shot exemplar prefix. A virtually embodied avatar controlled by Unity will vocalize selected responses with automated lip-sync; animation and UI display commands may also be sent by the orchestrator.

A key feature of our approach is the asynchronous processing of vision, language, and front-end modules from the core orchestrator logic. In contrast to the 4Hz vision prediction stream, the orchestrator operates at 12Hz to continue processing model-free control flow triggers and interaction management processes. This allows for a responsive core process, unencumbered by slower processing times of NN-based modules, and allows the language model to operate at a flexible rate capped at 12 Hz (depending on the availability of inputs from the orchestrator).

## 3. Exercise coaching as a case study on situated communication

We use fitness coaching as a test bed to evaluate our method using vision and language models fine-tuned specifically on expert-curated, fine-grained fitness activity recognition and feedback data. This testbed provides a high-paced dynamic interaction where LM responses must be generated and delivered rapidly, in real-time, to fast-changing environment states. Despite this constituting a highly narrow (albeit real-world) task domain, it is sufficiently open-ended to allow for dialogue.

### 3.1. Exercise video data and vision network

We crowdsourced the recording of short video clips of length 3 to 10 seconds, each showing a single person performing a repetition of a given exercise totalling approximately $300,000$ clips, or approximately $5,000$ clips per exercise. Fine-grained variations of each exercise were collected similar to [14, 21].

Using this, we trained a 3d convolutional network based on the Efficientnet-lite-v4 architecture [23] with multiple distinct recognition heads to map video clips to qualitative, quantitative, and repetition labels; pretraining of the network was done on ImageNet [7].

### 3.2. Language data and model

We collected a comprehensive collection of text responses encompassing a complete interaction spanning introductions, workout generation, exercise explanation, workout navigation, live exercise form feedback, and outgoing feedback summarization and remarks. In total we collected just over 8.5k exercise-specific lines across 136 exercises and 2k general workout lines.

We fine-tuned a handful of publicly available decoder-only language models, namely Pythia (2.8B and 6.9B) [3], and Pythia-chat-base (7B) [6] on the fitness dataset; natural language pretraining data and the OIG dataset [16] were used for regularization.

**Quantitative evaluations:** We assess the diversity of generated responses through Self-BLEU [26], and the quality of generated responses through perplexity and BERTScore [8, 12, 25]. Results are presented in Table 1.

**Human evaluations:** We evaluate model generations in terms of their contextual relevance (labeled "Relevance") and to what degree they reflect multi-hop reasoning (labeled "Multi-hopness"). Results are shown in Table 1.

| Model | Variant | Self-BLEU (↓) | Perplexity (↓) | | BERTScore (↑) | | Relevance | | Multi-Hopness | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | train | test | train | test | 0-shot | 9-shot | 0-shot | 9-shot |
| Pythia (2.8B) | Base | 0.403 | 30.5 | 29.8 | 0.8412 | 0.8378 | 1.33 | | 1.50 | |
| | Fitness FT + OIG | 2.093 | 7.12 | 15.3 | 0.8643 | 0.8528 | 3.02 | 3.21 | 2.66 | 2.79 |
| Pythia (6.9B) | Base | 0.390 | 28.8 | 28.0 | 0.8418 | 0.8378 | 1.24 | | 1.06 | |
| | Fitness FT + OIG | 0.607 | 14.7 | 17.8 | 0.8541 | 0.8476 | 2.23 | **3.75** | 1.90 | **3.52** |
| Pythia-CB (7B) | Base | 11.55 | 67.8 | 64.8 | 0.8434 | 0.8380 | 1.86 | | 1.68 | |
| | Fitness FT + OIG | 1.561 | 14.3 | 16.8 | 0.8488 | 0.8446 | 2.80 | 3.21 | 2.44 | 2.72 |
| Pythia-CB (7B) (Desc. names) | Base | - | - | - | - | - | 2.09 | 3.03 | 1.71 | 2.78 |
| | Fitness FT + OIG | - | - | - | - | - | 2.94 | 3.38 | 2.81 | 2.79 |

Table 1. Evaluation of model generations from Pythia (2.8B and 6.9B) and Pythia-chat-base (shown as Pythia-CB). Self-BLEU [26], Perplexity, and BERTScore [8, 12, 25] are presented to represent model generation diversity and semantic similarity to ground truth responses. Relevance and Multi-hopness columns present the average ratings from 17 labellers using a 5-point likert scale; a multi-hopness score above three indicates contextualized multi-turn responses. Base: pre-trained model; Fitness FT + OIG: fine-tuned model regularized with OIG dataset. As a reference, the Self-BLEU for the ground truth responses is 8.678.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 1

[3] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023. 2

[4] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. 1

[5] Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vincent Y. Zhao, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. Lamda: Language models for dialog applications. In *arXiv*. 2022. 1

[6] Together Computer. OpenChatKit: An Open Toolkit and Base Model for Dialogue-style Applications. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[9] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2

[10] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 2

[11] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3

[13] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko.

Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023. 2

[14] Antoine Mercier, Guillaume Berger, Sunny Panchal, Florian Dietrichkeit, Cornelius Böhm, Ingo Bax, and Roland Memisevic. Is end-to-end learning enough for fitness activity recognition? 2023. 2

[15] Reiichiro Nakano, Jacob Hilton, S. Arun Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332, 2021. 1

[16] Huu Nguyen, Sameer Suri, and et al. Tsui Ken. The open instruction generalist (oig) dataset, 2023. 2

[17] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023. 1

[18] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023. 1

[19] Gabriel Recchia. Teaching autoregressive language models complex tasks by demonstration. *ArXiv*, abs/2109.02102, 2021. 1

[20] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 1

[21] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[22] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 1

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[24] Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. Human parity on commonsenseqa: Augmenting self-attention with external attention. *arXiv preprint arXiv:2112.03254*, 2021. 1

[25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 2, 3

[26] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018. 2, 3