

## Xufeng Duan

Email Google Scholar Personal Website

**Research interests:** Psycholinguistics, Neurolinguistics, Mechanistic Interpretability

### Working Experience

---

---

|                                            |                                     |
|--------------------------------------------|-------------------------------------|
| <b>The Chinese University of Hong Kong</b> | 2025–Present                        |
| Brain and Mind Institute                   | <i>Research Assistant Professor</i> |

### Education

---

---

|                                                |                             |
|------------------------------------------------|-----------------------------|
| <b>The Chinese University of Hong Kong</b>     | 2020–2025                   |
| Department of Linguistics and Modern Languages | <i>Ph.D.</i> in Linguistics |
| <b>The Chinese University of Hong Kong</b>     | 2019–2020                   |
| Department of Linguistics and Modern Languages | <i>M.A.</i> in Linguistics  |
| <b>Sun Yat-sen University</b>                  | 2015–2019                   |
| Neurolinguistics Laboratory                    | <i>B.A.</i> in Chinese      |

### Research Projects

---

---

Participate, General Research Fund 2021/22 (14613722), *Neurocognitive mechanisms of character amnesia in Chinese handwriting*, RGC, HK\$803,050, 01/01/2023 – 31/12/2025. PI: Zhenguang Cai (CUHK).

Participate, Strategic Partnership Award for Research Collaboration, *Chinese handwriting in the digital age: neurocognitive mechanisms of character amnesia*, The Chinese University of Hong Kong, HK\$150,000, 02/2022 – 01/2023. PIs: Zhenguang Cai (CUHK) and Joseph Devlin (UCL).

### Supervision and Mentorship

---

---

**Bei Xiao** — PhD Student, co-supervised with Prof. Zhenguang G. Cai, The Chinese University of Hong Kong (CUHK PhD Studentship; will visit the Ted Lab at MIT in 2027).

**Zhaoqian Yao** — PhD Student, co-supervised with Prof. Zhenguang G. Cai, The Chinese University of Hong Kong.

**Beijia Kan** — PhD Student, co-supervised with Prof. Zhenguang G. Cai, The Chinese University of Hong Kong.

### Publication

---

---

**Duan, X.**, Xiao, B., Kan, B., & Cai, Z. G. (in press). StimMAS: A Multi-Agent Framework for Automated Linguistic Stimulus Construction in Psychological Research. *Advances in Methods and Practices in Psychological Science*. [10.31234/osf.io/83zev\\_v1](https://doi.org/10.31234/osf.io/83zev_v1)

**Duan, X.**, Yao, Z., Zhou, X., Fu, Z., Kan, B., Wang, Y., Xiao, B., Zhang, H., Tang, X., Nie, E., & Cai, Z. G. (2026). Mechanistic Interpretability for Understanding Language Abilities in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2026*. Budapest, Hungary: Association for Computational Linguistics. (to appear) [10.20944/preprints202607.1441.v1](https://arxiv.org/abs/2026.07.1441)

**Duan, X.**, Ma, Z., Yao, Z., Li, J., & Cai, Z. G. (2026). Do Large Language Models Acquire Phrase-Based Processing? Evidence from Eye Movements and Model-Brain Alignment After Fine-Tuning. In *Proceedings of the Society for Computation in Linguistics 2026* (pp. 464–476). San Diego, CA: Association for Computational Linguistics. [10.18653/v1/2026.scil-main.44](https://arxiv.org/abs/2026.04.44)

Tang, X., **Duan, X.**, & Cai, Z. G. (2026). Do Large Language Models Plan Answer Positions? Position Bias in Multiple-Choice Question Generation. *arXiv preprint*. [arXiv:2605.01846](https://arxiv.org/abs/2605.01846)

- Zhang, H., Zhang, Z., Wang, M., Su, Z., Wang, Y., Wang, Q., Yuan, S., Nie, E., **Duan, X.**, Xue, Q., Yu, Z., Shang, C., Liang, X., Xiong, J., Shen, H., Tao, C., Liu, Z., Jin, S., Xi, Z., Zhang, D., Ananiadou, S., Gui, T., Xie, R., So, H. K.-H., Schütze, H., Huang, X., Zhang, Q., & Wong, N. (2026). Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models. *Computer Science Review*, 62, 101011. [10.1016/j.cosrev.2026.101011](https://doi.org/10.1016/j.cosrev.2026.101011)
- Zhang, H., Zhang, Z., Nie, E., Wang, M., Su, Z., Wang, Y., Wang, Q., Yuan, S., **Duan, X.**, Xue, Q., Yu, Z., Shang, C., Liang, X., Xiong, J., Shen, H., Tao, C., Liu, Z., Jin, S., Xi, Z., Zhang, D., Ananiadou, S., Gui, T., Xie, R., So, H. K.-H., Schuetze, H., Huang, X., Zhang, Q., & Wong, N. (2026). Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 10317–10362). San Diego, California, United States: Association for Computational Linguistics. [10.18653/v1/2026.findings-acl.502](https://doi.org/10.18653/v1/2026.findings-acl.502)
- Fu, Z., **Duan, X.**, & Cai, Z. G. (2026). SCALPEL: Selective Capability Ablation via Low-rank Parameter Editing for Large Language Model Interpretability Analysis. *arXiv preprint*. [arXiv:2601.07411](https://arxiv.org/abs/2601.07411)
- Tang, X., **Duan, X.**, & Cai, Z. G. (2025). Large Language Models for Automated Literature Review: An Evaluation of Reference Generation, Abstract Writing, and Review Composition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 1602–1617). Suzhou, China: Association for Computational Linguistics. [10.18653/v1/2025.emnlp-main.83](https://doi.org/10.18653/v1/2025.emnlp-main.83)
- Zhao, N., & **Duan, X.** (2025). Enhancing Conference Interpreting with Computer-assisted Interpreting Tools: A Multi-agent System. *PsyArXiv preprint*. [10.31234/osf.io/pc3ht\\_v1](https://doi.org/10.31234/osf.io/pc3ht_v1)
- Zhao, N., **Duan, X.**, & Cai, Z. G. (2025). The missing half of language learning in current developmental language models: Exogenous and endogenous linguistic input. *Open Mind*, 9, 1543–1549. [10.1162/OPMI.a.33](https://doi.org/10.1162/OPMI.a.33)
- Duan, X.**, Yao, Z., Zhang, Y., Wang, S., & Cai, Z. G. (2025). How syntax specialization emerges in language models. *arXiv preprint*. [arXiv:2505.19548](https://arxiv.org/abs/2505.19548)
- Lin, Y., **Duan, X.**, Haslett, D., Chen, Y., & Cai, Z. G. (2025). Information Loss in LLMs’ Multilingual Translation: The Role of Training Data, Language Proximity, and Language Family. *arXiv preprint*. [arXiv:2506.23340](https://arxiv.org/abs/2506.23340)
- Xiao, B., **Duan, X.**, Haslett, D. A., & Cai, Z. G. (2025). Human-likeness of LLMs in the Mental Lexicon. *Proceedings of the 29th Conference on Computational Natural Language Learning* (pp. 586–601). Vienna, Austria: Association for Computational Linguistics. [10.18653/v1/2025.conll-1.38](https://doi.org/10.18653/v1/2025.conll-1.38)
- Qiu, Z., **Duan, X.**, & Cai, Z. G. (2025). Grammaticality representation in ChatGPT as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1), 1–15. [10.1057/s41599-025-04907-8](https://doi.org/10.1057/s41599-025-04907-8)
- Wang, S., **Duan, X.**, & Cai, Z. (2025). What to predict? Exploring how sentence structure influences contrast predictions in humans and large language models. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, J. Li, & B.-D. Oh (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 244–252). NAACL 2025. [10.18653/v1/2025.cmcl-1.28](https://doi.org/10.18653/v1/2025.cmcl-1.28)
- Wu, H., **Duan, X.**, & Cai, Z. (2025). Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, J. Li, & B.-D. Oh (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 135–143). NAACL 2025. [10.18653/v1/2025.cmcl-1.18](https://doi.org/10.18653/v1/2025.cmcl-1.18)
- Duan, X.**, Li, S., & Cai, Z. G. (2025). MacBehaviour: An R package for behavioural experimentation on large language models. *Behavior Research Methods*, 57(1), 19. [10.3758/s13428-024-02524-y](https://doi.org/10.3758/s13428-024-02524-y)
- Duan, X.**, Zhou, X., Xiao, B., & Cai, Z. (2025). Unveiling language competence neurons: A psycholinguistic approach to model interpretability. In O. Rambow, L. Wanner, M. Apidianaki, H.

- Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10148–10157). Association for Computational Linguistics. [2025.coling-main.677](#)
- Zhou, X., Chen, D., Cahyawijaya, S., **Duan, X.**, & Cai, Z. G. (2025). Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6866–6888). Association for Computational Linguistics. [2025.coling-main.459](#)
- \*Wang, S., \***Duan, X.**, & Cai, Z. (2024). A Multimodal Large Language Model “Foresees” Objects Based on Verb Information but Not Gender. *Proceedings of the 28th Conference on Computational Natural Language Learning* (pp. 435–441). [10.18653/v1/2024.conll-1.32](#)
- Qiu, Z., **Duan, X.**, & Cai, Z. (2024). Evaluating Grammatical Well-Formedness in Large Language Models: A Comparative Study with Human Judgments. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 189–198). [10.18653/v1/2024.cmcl-1.16](#)
- Duan, X.**, Xiao, B., Tang, X., & Cai, Z. G. (2024). HLB: Benchmarking LLMs’ Humanlikeness in Language Use. *arXiv preprint*. [arXiv:2409.15890](#)
- Duan, X.**, & Cai, Z. G. (2024). Chinese Character Processing. *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. [9780199384655.013.1012](#)
- Wu, H., **Duan, X.**, & Cai, Z. G. (2024). Speaker Demographics Modulate Listeners’ Neural Correlates of Spoken Word Processing. *Journal of Cognitive Neuroscience*, 36(10), 2208–2226. [10.1162/jocn-a\\_02225](#)
- Cai, Z., **Duan, X.**, Haslett, D., Wang, S., & Pickering, M. (2024). Do large language models resemble humans in language use? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 37–56). [10.18653/v1/2024.cmcl-1.4](#)
- Qiu, Z., **Duan, X.**, & Cai, Z. (2023). Does ChatGPT Resemble Humans in Processing Implicatures? *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop* (pp. 25–34). [2023.naloma-1.3](#)
- Lu, S., **Duan, X.**, & Li, X. (2021). Prominence and Locality Principle in Chinese Possessive Topicalization Sentences and its Psychological Basis—Evidence from eye-movements. *Contemporary Linguistics*, 23(1), 56–73. [\[pdf\]](#)

## Conference Talk

- Duan, X.**, Yao, Z., Zhang, Y., Wang, S., & Cai, Z. G. (2026). How syntax specialization emerges in language models. Paper presented at the *39th Annual Conference on Human Sentence Processing*, Cambridge, MA, United States.
- Duan, X.** (2024). MacBehaviour: An R package for behavioural experimentation on large language models. Paper presented at *Architectures and Mechanisms for Language Processing Asia (AMLaP Asia)*, Singapore.
- Duan, X.** (2024). The neural basis of decline in written production: Evidence from Chinese handwriting. Paper presented at the *Annual Meeting of the Society for the Neurobiology of Language*, Brisbane, Australia.
- Duan, X.** (2024). Humanlike language use in LLMs. Talk presented at the *CogSci Hong Kong Meetup*, Hong Kong, China.
- Duan, X.** (2023). Chinese translations of English words: A comparison between human and ChatGPT translations. Paper presented at *Architectures and Mechanisms for Language Processing Asia (AMLaP Asia)*, Hong Kong, China.
- Cai, Z. G., **Duan, X.**, Sun, L., & Pickering, M. (2023). Structural priming in ChatGPT. Paper presented at *Architectures and Mechanisms for Language Processing Asia (AMLaP Asia)*, Hong Kong, China.

- Cai, Z. G., **Duan, X.**, Haslett, D., Wang, S., & Pickering, M. (2023). Do language models resemble humans in language use? Paper presented at *Architectures and Mechanisms for Language Processing (AMLaP)*, Donostia-San Sebastián, Spain.
- Qiu, Z., **Duan, X.**, & Cai, Z. G. (2023). Does ChatGPT resemble humans in processing implicatures? Paper presented at the *4th Natural Logic Meets Machine Learning Workshop*, Nancy, France.
- Duan, X.** (2018). Language rehabilitation training therapies for Chinese aphasia: Using Chinese characters as clues. Paper presented at *Beyond a Limited Scope: Chinese for Specific, Academic, and Professional Purposes*, Honolulu, HI, United States.
- Duan, X.** (2018). The dependence on different symbolic systems during word recognition–pronunciation tests by Chinese native speakers with aphasia: Evidence from eye-tracking. Paper presented at the *Annual Conference of the Chinese Society of Neurological Linguistics*, Xuzhou, China.
- Duan, X.** (2018). The rules of prominence and locality: Psychological reality. Paper presented at the *Annual Academic Conference of the Chinese Linguistic Society*, Guangzhou, China.

### Poster Presentation

- 
- Duan, X.**, Xiao, B., Kan, B., & Cai, Z. G. (2026). StimMAS: A multi-agent framework for automated linguistic stimulus construction in psychological research. Poster presented at the *39th Annual Conference on Human Sentence Processing*, Cambridge, MA, United States.
- Cai, Z. G., **Duan, X.**, Sun, L., & Pickering, M. J. (2026). Do syntax-like representations emerge in LLMs? Evidence from structural priming. Poster presented at the *39th Annual Conference on Human Sentence Processing*, Cambridge, MA, United States.
- Zhang, Y., **Duan, X.**, & Cai, Z. G. (2024). The causal role of left inferior frontal gyri in Chinese character handwriting: A transcranial magnetic stimulation study. Poster presented at the *Annual Meeting of the Society for the Neurobiology of Language*, Brisbane, Australia.
- Qiu, Z., **Duan, X.**, & Cai, Z. G. (2024). Do large language models resemble humans in grammaticality judgment? Poster presented at the *Annual Conference on Human Sentence Processing*, Ann Arbor, MI, United States.
- Duan, X.**, Cai, Z. G., & Yao, B. (2023). Electrophysiological responses associated with character amnesia in Chinese handwriting. Poster presented at *Architectures and Mechanisms for Language Processing (AMLaP)*, Donostia-San Sebastián, Spain.
- Wu, H., **Duan, X.**, & Cai, Z. G. (2023). Belief of speakers’ linguistic competence modulates the N400 effect elicited by inconsistent lexical use. Poster presented at *Architectures and Mechanisms for Language Processing (AMLaP)*, Donostia-San Sebastián, Spain.
- Qiu, Z., **Duan, X.**, Cai, Z. G., & Zhao, N. (2023). Pragmatic implicature processing in ChatGPT. Poster presented at *Architectures and Mechanisms for Language Processing (AMLaP)*, Donostia-San Sebastián, Spain.

### Academic Service

---

#### Conference / Workshop Committee

Program Committee member for the 2nd Workshop on LLM Agents for Social Simulation (LASS 2026), co-located with ACM CIKM 2026.

#### Journal Reviewing

*NLP / Artificial intelligence: Computational Linguistics; IEEE Intelligent Systems; IEEE Transactions on Affective Computing; International Journal of Applied Linguistics; Journal of Language Modelling; Information Sciences; Data; Applied Artificial Intelligence; International Journal of Human-Computer Interaction.*

*Neuro/cognitive science: PLOS ONE; Behavior Research Methods; Psychology Communications; Cognitive Science; Journal of Neurolinguistics; Acta Psychologica; Cognitive Processing; Social Sciences*

& Humanities Open.

### Community Outreach

Initiated the WeChat public account *Computational Neurolinguistics* for research commentaries in Chinese; run the [Computational Neurolinguistics Forum](#).

### Training and Internship

---

---

#### TMS Training

May 2023

APAC TMS clinic Workshop, University of Nurnich & MagMore Company.

#### Language Therapist

Jul. 2017 – Jul. 2019

The Third Affiliated Hospital of Sun Yat-sen University.

#### Skill

---

---

**LLMs:** Mechanistic interpretability, Model fine-tuning, Benchmark Construction, etc.

**Experimentation:** fMRI / TMS / EEG / fNIRS / Eye-tracker using PsychoPy & Experiment Builder & E-prime & Gorilla / Qualtrics.

**Coding:** Python, R and Matlab.

**Last updated:** 11/08/2026