

Conversational Search with Large Language Models



Svitlana Vakulenko



Conversational Search

"DEFINITION 1. A conversational search system is a **system for retrieving information that permits a mixed-initiative** back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user."

Radlinski, F., & Craswell, N. (2017). "A Theoretical Framework for Conversational Search" (CHIIR 2017)

Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands



"DEFINITION 1. A conversational search system is a **system for retrieving information that permits a mixed-initiative** back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user."

Radlinski, F., & Craswell, N. (2017). "A Theoretical Framework for Conversational Search" (CHIIR 2017)

Walker et al. (2003). "A multi-turn spoken dialogue system for information access" (HLT-NAACL BEA 2003)

Brin, S., & Page, L. (**1998**). "The Anatomy of a Large-Scale Hypertextual Web Search Engine." (WWW7)

Bates, M. J. (1989). "The design of browsing and berry-picking techniques for the online search interface" Online Review, vol. 13

Belkin et al. (1982). "ASK for Information Retrieval: Part I. Background and Theory" Journal of Documentation 38, 2

Bush, V. (1945). "As we may think." The Atlantic Monthly 176, 1

AI Overview

Google has not made its standalone Conversational AI interface the universal default for all searches. However, it heavily integrates AI into the core search page via AI Overviews and a dedicated AI Mode that users can easily click into. You can also access the pure chatbot experience directly on [Google AI](#).

Why the Shift?

- **Complex Queries:** Users increasingly ask longer, more conversational questions requiring synthesis rather than just simple keywords.
- **The Agentic Vision:** Google aims to transition from a list of blue links to an "AI assistant" that answers immediately and retains context for multi-step tasks. [blog.google +2](#)

The Reaction

- **General Users:** Opinions in discussions on platforms like [Reddit](#) are mixed. Many appreciate the convenience, while privacy advocates push back over default settings that feed user-uploaded media (like photos and voice inputs) into AI training models. [YouTube · WheezyNews +2](#)
- **The Open Web:** Publishers and content creators are concerned because scraping answers directly onto the search page limits click-through traffic to independent websites. [YouTube · WheezyNews +1](#)
- **Alternative Options:** Privacy-focused, no-AI alternatives like DuckDuckGo saw a surge in user traffic following Google's AI push. [YouTube · BitBiasedAI](#)

Are They Going to Roll It Back?

No. While Google Chrome briefly tested redirecting all address bar queries to AI Mode, executives confirmed this was an experimental error. The company is doubling down on this technology, viewing conversational AI as the future of search, but they are keeping traditional "blue link" options available. [Mashable +2](#)

AI responses may include mistakes. [Learn more](#)



Ask anything



[LinkedIn](#) 10 sites

Google's Going Full AI Mode. Here's...
25 May 2026 — hello the AI apocalypse is... [5:38](#)

[YouTube · WheezyNews](#)

Google AI Mode Shifts Organic Traffic to...
27 May 2026 — Google is officially rolling out its new "AI Mode" to... [Post on LinkedIn](#)

[LinkedIn · Sasha Reid](#)

Your Google Search bar isn't drastically changing....
22 May 2026 — Here's what is. ... Google Search is getting a faceli... [Search](#)

[USA Today](#)

Show all

AI responses may include mistakes. [Learn more](#)



is this part conversational does it take previous question and answer as history context into account?

15 June 2026

Yes, this specific part of the conversation is fully conversational and takes the previous history into account. 

As an AI, I analyze the earlier text in our chat to understand your follow-up questions. For example, when you just said *"is this part conversational"*, I knew instantly you were asking about our current interaction, building directly on our previous discussion about Google's AI models.

How the History Context Works

- **Memory Loop:** Every new message you send is bundled with the older questions and answers.
- **Context Awareness:** The system reads the entire bundle together to keep the conversation flowing naturally.
- **Pronoun Resolution:** You can use words like "it," "they," or "this" without re-explaining what you mean. 

When the Context Resets

- **New Chat:** If you open a completely new session, the system starts with a blank slate.
- **Topic Drift:** If you switch topics entirely, the system focuses on the new keywords but still holds the old data in the background until the session ends. 

If you would like to test this context memory further, let me know if you want to **deep dive into Google's specific AI models** or **explore how to turn off AI features** in your own browser.

AI responses may include mistakes. [Learn more](#)

 4 sites

AgentSession and Multi-Turn Conversations in...

Run this and you'll have a fully stateful conversation loop. The...



 Dev Leader

How many turns in conversation history relevant for new response? :...

To create a new response a human takes the past history of the conversations (the past turns) into...

 Reddit

Using "we" in American History : r/historyteachers

I think it's fine to say "they" when speaking about the past when you weren't alive. But i don't think...

 Reddit

Show all

Large Language Model

Large Language Model

"we train **GPT-3**, an autoregressive language model with **175 billion parameters, 10x more than any previous non-sparse language model**, and test its performance in the few-shot setting."

Brown et al. (2020). "Language Models are Few-Shot Learners" (NeurIPS 2020)

"we train **GPT-3**, an autoregressive language model with **175 billion parameters, 10x more than any previous non-sparse language model**, and test its performance in the few-shot setting."

Brown et al. (2020). "Language Models are Few-Shot Learners" (NeurIPS 2020)

Roberts et al. (2020). "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" (EMNLP 2020)

Radford et al. (**2019**). "Language Models are Unsupervised Multitask Learners." OpenAI Technical Report, 1(8), 9.

Vaswani et al. (2017). "Attention Is All You Need" (NeurIPS 2017)

Mikolov et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality" (NeurIPS 2013)

Bengio et al. (2000). "A Neural Probabilistic Language Model" (NIPS 2000)

Shannon, C. E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal, 27(3), 379–423.

- **Past** (2019-2025)
 - PhD intern @ Apple Siri (June - September **2019**)
 - PostDoc @ University of Amsterdam (2019-2021)
 - ML scientist @ Amazon Alexa (2021-2025)
- **Present** (2025-2026) @WU WWTF-VRG
- **Future** (2026-2033) @WU WWTF-VRG

Apple Internship

Question: Tell me about the benefits of **Yoga**?

Answer: Increased flexibility, muscle strength...

URL: <https://osteopathic.org/what-is-osteopathic-medicine/benefits-of-yoga>

Question: Does **it** help in reducing stress?

Rewrite: Does **Yoga** help in reducing stress?

Answer: Yoga may help reduce stress, lower blood pressure, and lower your heart rate.

URL: <https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/yoga/art-20044733>

Apple Internship

Question: Tell me about the benefits of **Yoga**?

Answer: Increased flexibility, muscle strength...

URL: <https://osteopathic.org/what-is-osteopathic-medicine/benefits-of-yoga>

Question: Does **it** help in reducing stress?

Rewrite: Does **Yoga** help in reducing stress?

Answer: Yoga may help reduce stress, lower blood pressure, and lower your heart rate.

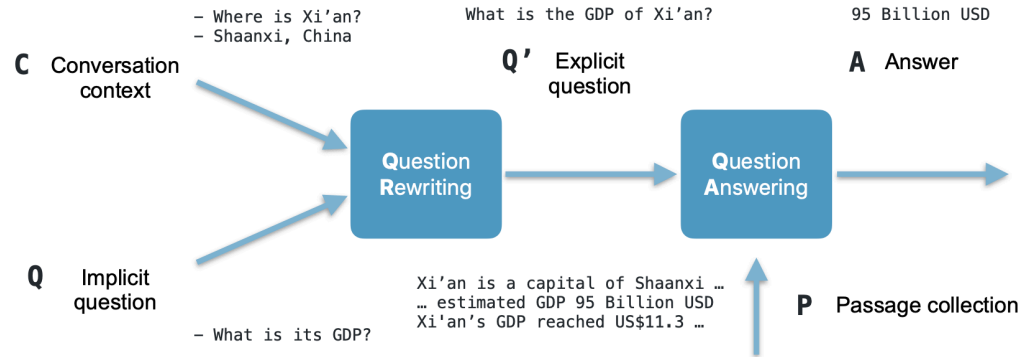
URL: <https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/yoga/art-20044733>

Question: What are some of the main types?

Rewrite: What are some of the main types **of Yoga**?

Answer: Hatha, Kundalini, Ashtanga, ...

URL: <https://www.mindbodygreen.com/articles/the-11-major-types-of-yoga-explained-simply>



1. **Svitlana Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: Question Rewriting for Conversational Question Answering. WSDM 2021: 355-363
2. Raviteja Anantha, **Svitlana Vakulenko**, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534
3. **Svitlana Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering. SCAI@EMNLP 2020: 10-17
4. **Svitlana Vakulenko**, Nikos Voskarides, Zhucheng Tu, Shayne Longpre: A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. ECIR 2021: 442-449
5. **Svitlana Vakulenko**, Nikos Voskarides, Zhucheng Tu, Shayne Longpre: Leveraging Query Resolution and Reading Comprehension for Conversational Passage Retrieval. TREC 2020

1. **Svitlana Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: Question Rewriting for Conversational Question Answering. WSDM 2021: 355-363
2. Raviteja Anantha, **Svitlana Vakulenko**, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534
3. **Svitlana Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering. **SCAI@EMNLP** 2020: 10-17
4. **Svitlana Vakulenko**, Nikos Voskarides, Zhucheng Tu, Shayne Longpre: A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. ECIR 2021: 442-449
5. **Svitlana Vakulenko**, Nikos Voskarides, Zhucheng Tu, Shayne Longpre: Leveraging Query Resolution and Reading Comprehension for Conversational Passage Retrieval. TREC 2020



Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands
2. 2018 EMNLP @Brussels, Belgium
3. 2019 WebConf @San Francisco, USA + IJCAI @Macao, China
4. **2020 EMNLP @Punta Cana, Dominican Republic**

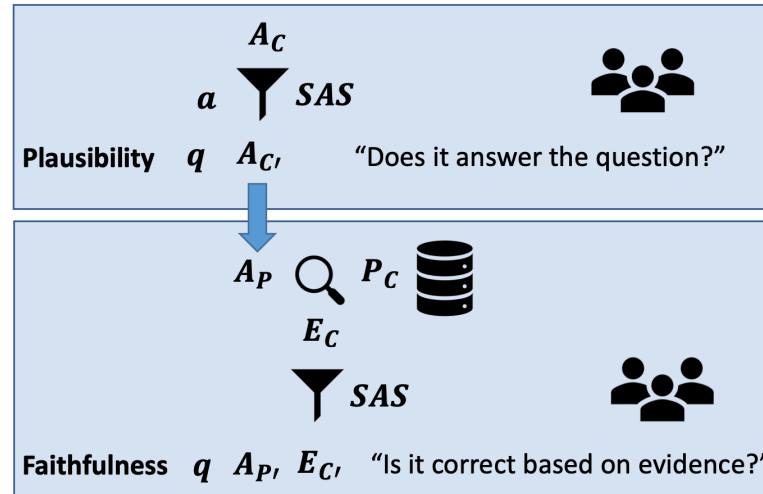


Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands
2. 2018 EMNLP @Brussels, Belgium
3. 2019 WebConf @San Francisco, USA + IJCAI @Macao, China
4. 2020 EMNLP @Punta Cana, Dominican Republic
5. **2021 Online Event @Zoom/GatherTown**



1. **Svitlana Vakulenko**, Johannes Kiesel, Maik Fröbe: SCAI-QReCC Shared Task on Conversational Question Answering. LREC 2022: 4913-4922



1. **Svitlana Vakulenko**, Evangelos Kanoulas, Maarten de Rijke: An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues. SIGIR 2020: 1109–1112
2. **Svitlana Vakulenko**, Evangelos Kanoulas, Maarten de Rijke: A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. ACM Trans. Inf. Syst. 39(4): 49:1-49:32 (2021)

1. Vera Provatorova, **Svitlana Vakulenko**, Evangelos Kanoulas, Karst Dercksen, Joeri M. van Hulst: Named Entity Recognition and Linking on Historical Newspapers: UvA. ILPS & REL at CLEF HIPE 2020. CLEF 2020: 1–13
2. Vera Provatorova, Samarth Bhargav, **Svitlana Vakulenko**, Evangelos Kanoulas: Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing. EMNLP (1) 2021: 10501-10510
3. Julien Rossi, **Svitlana Vakulenko**, Evangelos Kanoulas: VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law. CIKM 2021: 4554-4563
4. Georgios Sidiropoulos, Nikos Voskarides, **Svitlana Vakulenko**, Evangelos Kanoulas: Combining Lexical and Dense Retrieval for Computationally Efficient Multi-hop Question Answering. SustaiNLP@EMNLP 2021: 58–63
5. Niels J. Rouws, **Svitlana Vakulenko**, Sophia Katrenko: Dutch SQuAD and Ensemble Learning for Question Answering from Labour Agreements. BNAIC/BENELEARN 2021: 155-169
6. Georgios Sidiropoulos, **Svitlana Vakulenko**, Evangelos Kanoulas: On the Impact of Speech Recognition Errors in Passage Retrieval for Spoken Question Answering. CIKM 2022: 4496–4500
7. Mariya Hendriksen, Maurits J. R. Bleeker, **Svitlana Vakulenko**, Nanne van Noord, Ernst Kuiper, Maarten de Rijke: Extending CLIP for Category-to-Image Retrieval in E-Commerce. ECIR (1) 2022: 289-303
8. Mariya Hendriksen, **Svitlana Vakulenko**, Ernst Kuiper, Maarten de Rijke: Scene-Centric vs. Object-Centric Image-Text Cross-Modal Retrieval: A Reproducibility Study. ECIR (3) 2023: 68-85
9. Shaojie Jiang, **Svitlana Vakulenko**, Maarten de Rijke: Weakly Supervised Turn-level Engagingness Evaluator for Dialogues. CHIIR 2023: 258-268
10. Weijia Zhang, **Svitlana Vakulenko**, Thilina Rajapakse, Yumo Xu, Evangelos Kanoulas: Tackling Query-Focused Summarization as A Knowledge-Intensive Task: A Pilot Study. Gen-IR@SIGIR 2023
11. Weijia Zhang, Jia-Hong Huang, **Svitlana Vakulenko**, Yumo Xu, Thilina Rajapakse, Evangelos Kanoulas: Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models. ICPR (19) 2024: 89-104
12. Mariya Hendriksen, **Svitlana Vakulenko**, J. Massiah, G. Kazai, E. Yilmaz: Image-Seeking Intent Prediction for Cross-Device Product Search. GenAI4ECom@RecSys 2025
13. Shaojie Jiang, **Svitlana Vakulenko**, Maarten de Rijke: Orcheo: A Modular Full-Stack Platform for Conversational Search. SIGIR 2026

ML scientist @ Amazon Alexa

1. Xiaoyu Shen, **Svitlana Vakulenko**, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, Adrià de Gispert: Neural Ranking with Weak Supervision for Open-Domain Question Answering : A Survey. EACL (Findings) 2023: 1691-1705
2. **Svitlana Vakulenko**, Bill Byrne, Adrià de Gispert: Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation. ECAI 2023: 2378-2385
3. Philipp Christmann, **Svitlana Vakulenko**, Ionut-Teodor Sorodoc, Bill Byrne, Adrià de Gispert: Retrieving Contextual Information for Long-Form Question Answering Using Weak Supervision. Findings of the Association for Computational Linguistics: EMNLP 2024: 14301–14316

Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands
2. 2018 EMNLP @Brussels, Belgium
3. 2019 WebConf @San Francisco, USA + IJCAI @Macao, China
4. 2020 EMNLP @Punta Cana, Dominican Republic
5. 2021 Online Event @Zoom/GatherTown
6. **2022 SIGIR @Madrid, Spain**



Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands
2. 2018 EMNLP @Brussels, Belgium
3. 2019 WebConf @San Francisco, USA + IJCAI @Macao, China
4. 2020 EMNLP @Punta Cana, Dominican Republic
5. 2021 Online Event @Zoom/GatherTown
6. 2022 SIGIR @Madrid, Spain
7. 2023
8. 2024 CHIIR @Sheffield, UK



The 1st Vienna Research Group @ WU

VRG leader: [Svitlana Vakulenko](#)

Institution: Amazon

Proponent: [Axel Polleres](#)

Institution: WU - Vienna University of Economics and Business

Projekttitel: Knowledge Representation Learning for Large Language Models

Status: Laufend (01.06.2025 – 31.05.2033)

GrantID: 10.47379/VRG24013

Fördersumme: € 1.597.758

The 1st Vienna Research Group @ WU



Adrian Bracher



Svitlana Vakulenko



Omar Hamed



[first name].[last name]@wu.ac.at

BSc Thesis Topics 2025/2026

1. Takacs Richard. Analysis of Failure Modes in Generative Retrieval
2. Papushina Sofia. CoReL: Conversational Data for Retriever Optimization for Long-Form Question Answering
3. Schweitzer Vincent. Adoption of RAG Systems in the Legal Domain
4. Krupalek Philipp. Evaluation of RAG Systems in the Legal Domain
5. Barros Marco. LLM-as-a-Judge for Legal RAG Evaluation
6. Gotschim Tim. An Empirical Study of Hallucination Triggers

1. Richard Takacs, Adrian Bracher, **Svitlana Vakulenko**: Understanding and Debugging Failures in N-Gram-Based Generative Retrieval. VulGen@SIGIR 2026
2. Adrian Bracher, **Svitlana Vakulenko**: Generative Retrieval Overcomes Limitations of Dense Retrieval but Struggles with Identifier Ambiguity. arXiv preprint arXiv:2604.05764 (2026)
3. Vahid Sadiri Javadi, Fryderyk Róg, A. Aksa, Jenny R. Trippas, **Svitlana Vakulenko**, Lucie Flek: CHARISMA: Character-Based Interaction Simulation with Multi-LLM Agents Toward Computational Social Psychology. CHIIR 2026: 1–11
4. Philipp Christmann, Roxana Petcu, Shivani Singhania, Mohammad Aliannejadi, Marco Gohsen, **Svitlana Vakulenko**: The 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2026). SIGIR 2026
5. **Svitlana Vakulenko**, Philipp Christmann, Isabel Feustel, Vaibhav Adlakha, Vahid Sadiri Javadi, Patrícia Schmidtová: Report on the 9th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2025) at IJCAI 2025. SIGIR Forum 59(2): 1–12 (2025)

Search-Oriented Conversational AI (SCAI)

1. 2017 ICTIR @Amsterdam, Netherlands
2. 2018 EMNLP @Brussels, Belgium
3. 2019 WebConf @San Francisco, USA + IJCAI @Macao, China
4. 2020 EMNLP @Punta Cana, Dominican Republic
5. 2021 Online Event @Zoom/GatherTown
6. 2022 SIGIR @Madrid, Spain
7. 2023
8. 2024 CHIIR @Sheffield, UK
9. 2025 IJCAI @Montreal, Canada
- 10. 2026 SIGIR @Melbourne, Australia**



Future (2026+)

-> **WWTF** Empirical Social Sciences (ESS) 2026 „Resilience & Crises”

*University of Vienna (PI)

*Vienna University of Economics and Business (**Co-PI**)

*Danube Private University (Co-PI)

*Paris Lodron University of Salzburg

Future (2026+)

-> **WWTF** Empirical Social Sciences (ESS) 2026 „Resilience & Crises“

*University of Vienna (PI)

*Vienna University of Economics and Business (**Co-PI**)

*Danube Private University (Co-PI)

*Paris Lodron University of Salzburg

[Anbahnungsfinanzierung ERC-2027-StG](#)

[Vakulenko, S.](#) (PI - Project lead)

WU intern - Globalbudget: €15,000.00

15/06/26 → 31/12/26

Award date: 12/06/26

Award: Internal Projects > internal projects - application FOS

-> **ERC 2027 Starting Grant (StG)**

Title: Tracking Information Provenance in Large Language Models

Project duration: 5 years

Funding amount: EUR 1.5M (WU share 100%)