

Labadain: Language Technology Resources and Tools for Tetun

Gabriel de Jesus, PhD

Affiliated Researcher at **INESC TEC** (Porto, Portugal)

Founder of **Timor News** (Dili, Timor-Leste)

Creator of **Labadain**

2nd ASEAN-NEXUS ECRs Conclave

June 2–6, 2026

Singapore



Labadain Research

Labadain is a set of language technology resources and tools created for the **Tetun language**



Datasets

- ✓ Labadain-30k+
- ✓ Labadain-Avaliadór
- ✓ Labadain-Stopwords
- ✓ LabadainLog-17k+
- ✓ Labadain-ZSRuns



Software

- ✓ Labadain Crawler



Algorithms and tools

- ✓ Labadain-Stemmer
- ✓ Tetun Tokenizer
- ✓ Tetun LID



Prototypes

- ✓ Labadain Search
- ✓ Labadain Chat

Latest Research

Labadain Chat

Jan 2026: Labadain **LIX-R361** – Agentic AI
Evaluated task success and user satisfaction

Labadain Chat: A Conversational Agent for the Tetun Language

Gabriel de Jesus
Institute for Systems and Computer Engineering,
Technology and Science (INESC TEC)
Porto, Portugal
gabriel.jesus@inesctec.pt

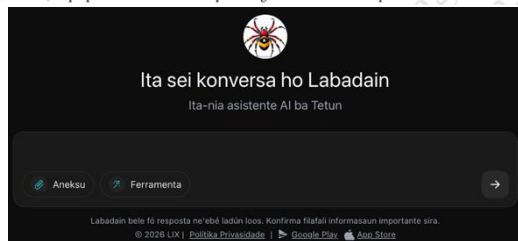
Sérgio Nunes
INESC TEC / University of Porto
Porto, Portugal
sergio.nunes@fe.up.pt

Abstract

Large language model (LLM)-based conversational assistants are designed for general-purpose conversation tasks and are primarily optimized for high-resource languages. Although these systems support some low-resource languages (LRLs), their responses often fall short of user expectations. Conversely, the openness of LLMs presents significant challenges for targeted, language-specific solutions that can effectively serve underrepresented language communities. This study presents Labadain Chat, a conversational agent for Tetun, a low-resource language spoken by over 932,000 people in Timor-Leste. We adapt existing LLMs to Tetun

1 Introduction

Large language model (LLM)-powered conversational assistants, such as ChatGPT, Gemini, Claude, and Perplexity, have fundamentally reshaped information access, enabling tasks such as question answering, summarization, multi-step reasoning, and complex task execution. These systems improve access to information and present significant challenges for speakers of low-resource languages (LRLs) due to inconsistent performance, data scarcity, and linguistic bias [3, 18]. Consequently, these communities remain marginalized, emphasizing the need for targeted, language-specific solutions to enable equitable access to information.



Platform: <https://www.labadain.com>

Labadain Timor-Leste

Hafasil ita-nia moris ho Labadain, asistente intelijénsia artifisiál hosi Timor News.

Ita-boot presija login hodi bele utiliza Labadain:

- Introdús ita-nia utilizadór no password ho loloos.
- Karik ita-nia konta seidaok iha, karrega butaun rejistu nian hodi kria ida.
- Karik ita haluha password, karrega iha butaun reset password hodi hetan fail foun ida.

Utilizadór

Password

Login

Platform: <https://old.labadain.com>

Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior

Gabriel de Jesus
INESC TEC / University of Porto
Porto, Portugal
gabriel.jesus@inesctec.pt

Sérgio Nunes
INESC TEC / University of Porto
Porto, Portugal
sergio.nunes@fe.up.pt

Abstract

Advancements in large language model (LLM)-based conversational assistants have transformed search experiences into more natural and context-aware dialogues that resemble human conversation. However, limited access to interaction log data hinders a deeper understanding of their real-world usage. To address this gap, we analyzed 16,952 prompt logs from 904 unique users of Labadain Chat, an LLM-based conversational assistant designed for Tetun speakers, to uncover patterns in user search behavior, engagement, and intent. Our findings show that most users (98.87%) spent between one and five minutes per session, with an average of 43 unique daily users. The majority (93.97%) submitted multiple prompts per session, with an average session duration of 16.9 minutes. Most users (95.22%) were based in Timor-Leste, with education and science (28.75%) and health (28.00%) being the most searched topics. We compared our findings with a study on Google Bard logs in English, revealing similar search characteristics—including engagement duration, command-based instructions, and requests for specific assistance.

1 Introduction

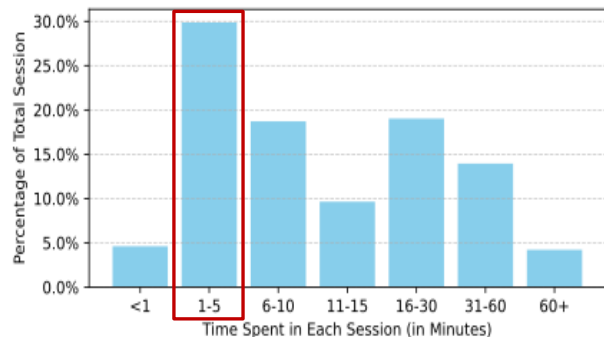
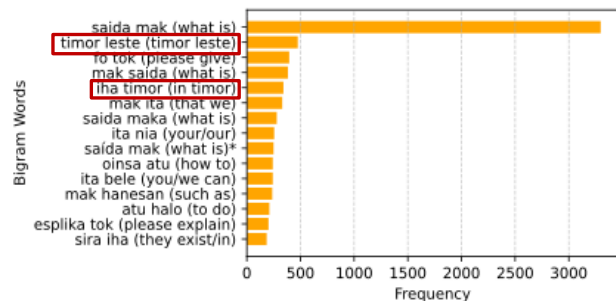
Conversational search refers to user-friendly dialogues between humans and machines, whether spoken or written, aimed at satisfying information needs [3, 4, 21, 29]. With the advent of LLM-based conversational assistants, such as OpenAI's ChatGPT,¹ Google's Gemini,² Anthropic's Claude,³ and similar AI tools, such dialogues have become more natural and context-aware, closely resembling human conversation. These systems' ability to process unstructured language, handle complex queries, and support multi-turn interactions has made search more seamless and intuitive. Furthermore, these developments demand a deeper understanding of user behavior, providing deeper insights into search behavior, engagement, and intent. These insights are crucial for refining LLM-based conversational systems to provide more personalized and effective interactions, ultimately enhancing the search experience [19, 31]. Despite these advancements, limited access to interaction log data hinders understanding of how these systems are used [25]. This challenge is especially pronounced in underrepre-

Research publication available (DOI: [10.1145/3731120.3744596](https://doi.org/10.1145/3731120.3744596))

Jan 2024: Labadain **R361** – Conversational AI Assistant
Studied user's search behavior, engagement, and Intent

Latest Research

Labadain R361



Paper: <https://dl.acm.org/doi/abs/10.1145/3731120.3744596>

Dataset: <https://doi.org/10.25747/X2DK-5Y06>

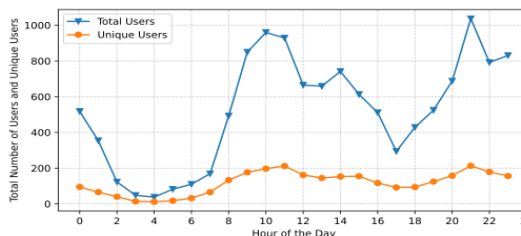
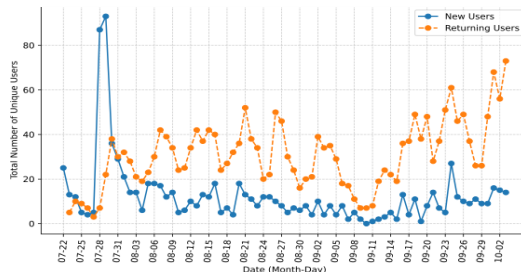
Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior

Gabriel de Jesus
INESC TEC / University of Porto
Porto, Portugal
gabriel.jesus@inesctec.pt

Sérgio Nunes
INESC TEC / University of Porto
Porto, Portugal
sergio.nunes@fe.up.pt

Abstract

Advancements in large language model (LLM)-based conversational assistants have transformed search experiences into more natural and context-aware dialogues that resemble human conversation. However, limited access to interaction log data hinders a deeper understanding of their real-world usage. To address this gap, we analyzed 16,952 prompt logs from 904 unique users of Labadain Chat, an LLM-based conversational assistant designed for Tetun speakers, to uncover patterns in user search behavior, engagement, and intent. Our findings show that most users (29.87%) spent between one and five minutes per session, with an average of 43 unique daily users. The majority (93.97%) submitted multiple prompts per session, with an average session duration of 16.9 minutes. Most users (95.22%) were based in Timor-Leste, with education and science (28.75%)



1 Introduction

Conversational search refers to user-friendly dialogues between humans and machines, whether spoken or written, aimed at satisfying information needs [3, 4, 21, 29]. With the advent of LLM-based conversational assistants, such as OpenAI's ChatGPT,¹ Google's Gemini,² Anthropic's Claude,³ and similar AI tools, such dialogues have become more natural and context-aware, closely resembling human conversation. These systems' ability to process unstructured language, handle complex queries, and support multi-turn interactions has made search more seamless and intuitive.

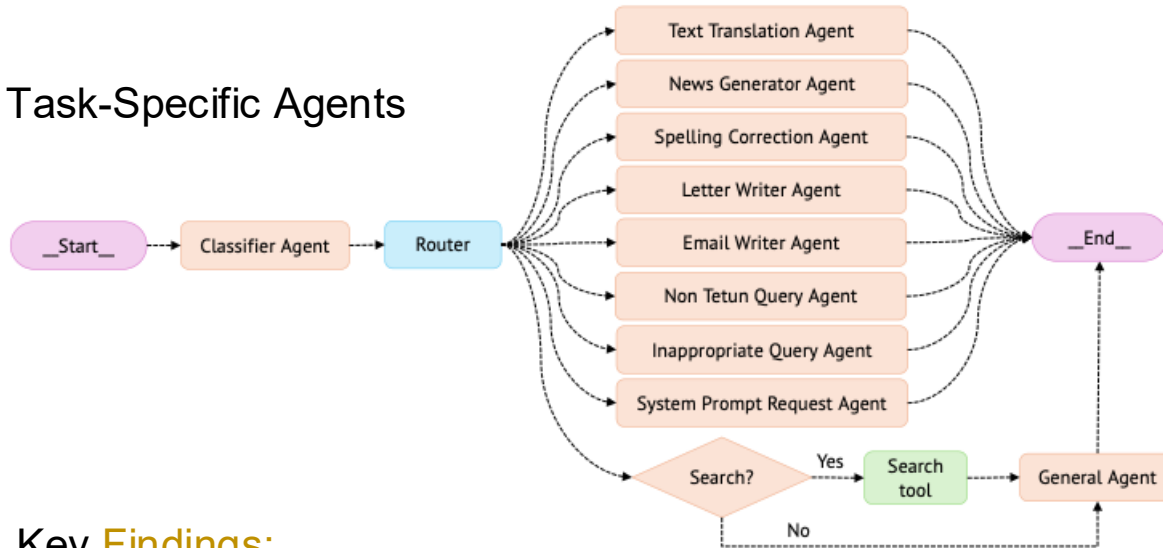
Furthermore, these developments demand a deeper understanding of user behavior, providing deeper insights into search behavior, engagement, and intent. These insights are crucial for refining LLM-based conversational systems to provide more personalized

Topic	L. Chat	L. Search	Timor News
Education/Science	28.75%	16.75%	11.25%
Health	28.00%	16.50%	12.25%
Social/Culture	13.75%	11.75%	3.25%
Research/Academia	6.75%	4.00%	3.75%
Politic/Government	2.50%	7.50%	6.00%
Job vacancy	0.00%	3.50%	7.25%
Economy/Finance	1.00%	6.75%	2.75%
Agriculture/Environment	6.50%	2.00%	1.75%
Law/Justice	0.75%	3.75%	1.75%
Language/Translation	2.00%	4.25%	1.50%
Computer science	3.75%	3.00%	1.25%
Personal	0.50%	2.50%	3.25%
Porn	0.00%	0.25%	2.00%
Sport	0.75%	1.00%	0.50%
Holiday/Travel	0.25%	0.50%	0.75%
Entertainment	0.00%	0.50%	0.75%
Other	4.75%	15.75%	40.00%

Latest Research

Labadain LIX-R361

Task-Specific Agents



Key Findings:

Language-specific LLM customization enhances both response effectiveness and user satisfaction for Tetun-speaking users



Ongoing Research

Labadain-HealthAI

A general-purpose conversational AI system for healthcare in Tetun



The Asia Foundation



Duration
1 Year



Funded by
TAF Timor-Leste



To develop a safe and culturally appropriate general-purpose conversational AI system that enable access to healthcare information in the Tetun language.



Including:

- ✓ Medical terminology in Tetun
- ✓ Tetun text healthcare corpus

Expected **outcomes**

- ✓ A functional conversational AI
- ✓ Healthcare datasets in Tetun
- ✓ Improve access to health information in Tetun

Research **Interest**

Open to research collaboration on advancing language technology for low-resource languages (LRLs), with applications in health and other high-impact domains



Areas of **interest**



Information Retrieval
and NLP for LRLs



Conversational AI



Retrieval-Augmented
Generation (RAG)



AI for Health and Legal
domains

Get in Touch

Google

All Images Videos News Forums Maps More ▾ Tools ▾

 **labadain.com**
<https://gabrieldejesus.labadain.com> ⋮

Gabriel de Jesus - Information Retrieval Researcher

I am an affiliated researcher at INESC TEC Porto (Portugal) and the creator of Labadain. My research focuses on language inclusion and digital preservation of ... [Read more](#)

 **Google Scholar**
<https://scholar.google.com/citations> ⋮

Gabriel de Jesus

Labadain-30k+: a monolingual Tetun document-level audited dataset. G de Jesus, S Nunes. Proceedings of the 3rd Annual Meeting of the Special Interest Group on ... [Read more](#)

 **labadain.com**
<https://gabrieldejesus.labadain.com/publications> ⋮

Publications - Gabriel's Personal Page

by G de Jesus · Cited by 2 — Information Retrieval Researcher and Creator of Labadain, a set of language-technology resources and tools for the Tetun language. Follow. Timor-Leste; Email ... [Read...](#)

 **GitHub**
<https://github.com/gabriel-de-jesus-labadain-crawler> ⋮

[gabriel-de-jesus/labadain-crawler](#)

Labadain crawler is a data collection pipeline for low-resource languages designed to automate and optimize the process of constructing comprehensive textual ... [Read more](#)

Scan here to view my
professional profile





Thank You

Labadain: Resources and Tools for the Tetun Language

Gabriel de Jesus, PhD

Affiliated Researcher at **INESC TEC** (Porto, Portugal)

Founder of **Timor News** (Dili, Timor-Leste)

Creator of **Labadain**

2nd ASEAN-NEXUS ECRs Conclave

June 2–6, 2026

Singapore

