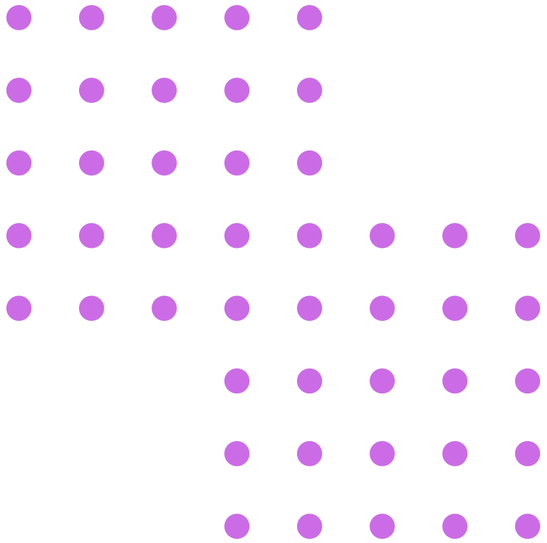
MCIF: Multimodal Crosslingual Instruction-Following Benchmark from Scientific Talks

Sara Papi, Maike Züfle, Marco Gaido,
Beatrice Savoldi, Danni Liu, Ioannis Douros,
Luisa Bentivogli, Jan Niehues

Why Instruction Following?



Modern LLMs operate via **natural instructions**, often in **multiple languages**

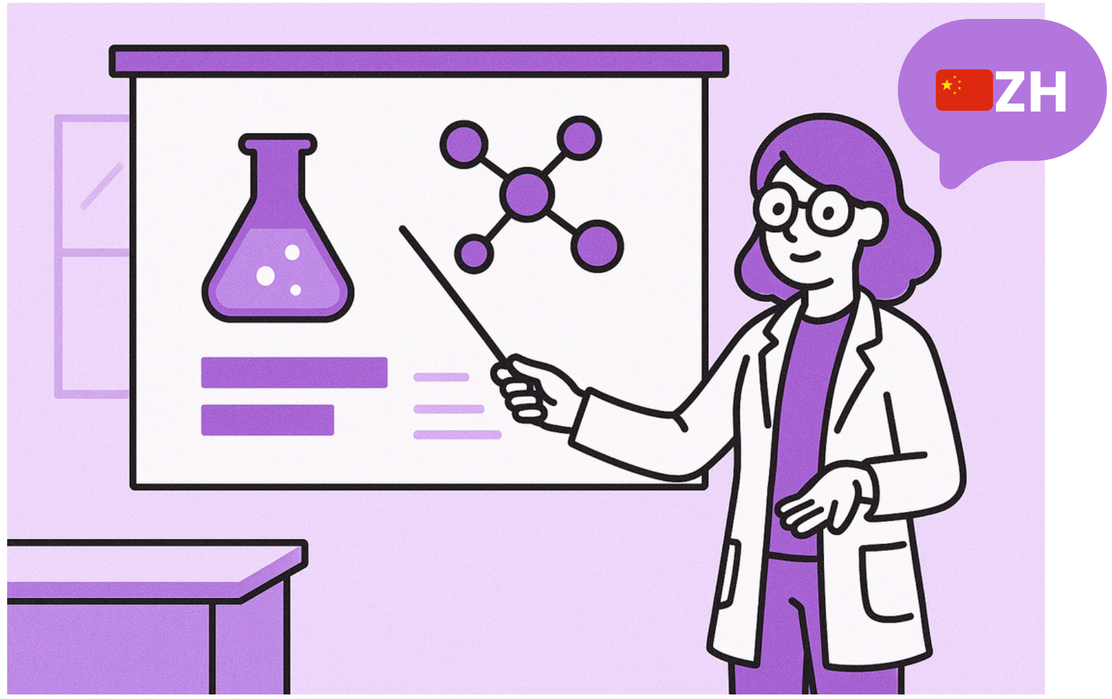
A model is expected to handle **text, speech, and video inputs**

Instruction following becomes a **unifying interface** for **multiple tasks**

Our Goal:

Assessing **Crosslingual and Multimodal** capabilities of **General-Purpose Models**

EN: “Based on the chart on the slide, explain the conclusion of the work.”



Current Evaluation...

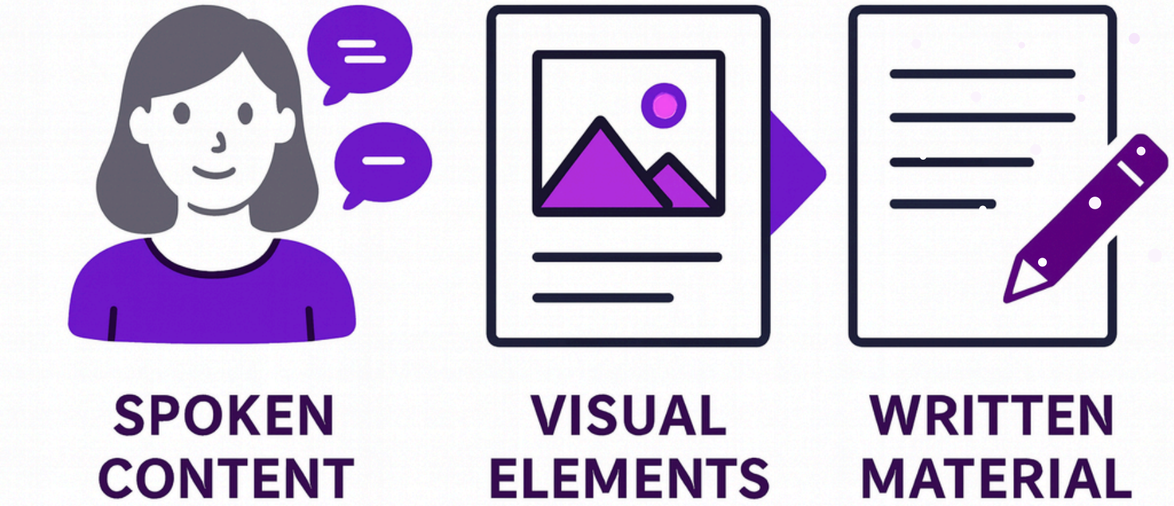
1. Often **single modality**
2. Primarily **monolingual**
3. Often **single task**
4. Ignore **long-context** inputs

...but Real World is: Multimodal

Instructions may refer to:

- **Spoken** content
- **Visual** elements
- **Written** material

Models must ***integrate*** information across **modalities**



...but Real World is: Multilingual

Multinguality: input and output in multiple languages

Crosslinguality: input and output languages are different

Models must ***interpret*** from the prompt the **correct language** with which to reply

...but Real World is: Multitask

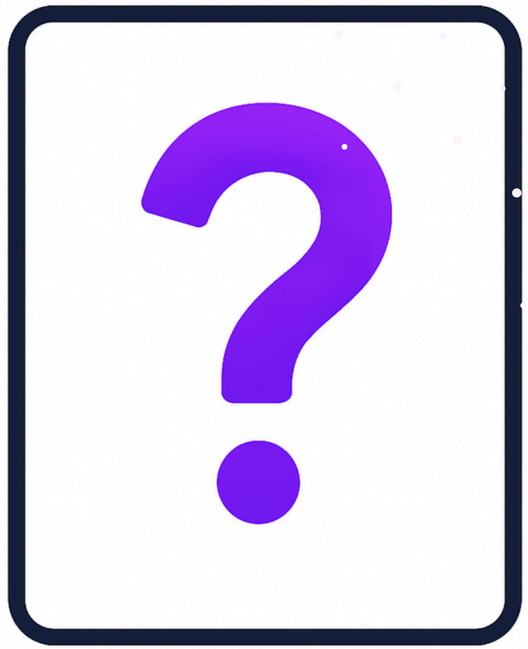
No explicit information on the task that should be executed

transcription

translation

summarization

question answering



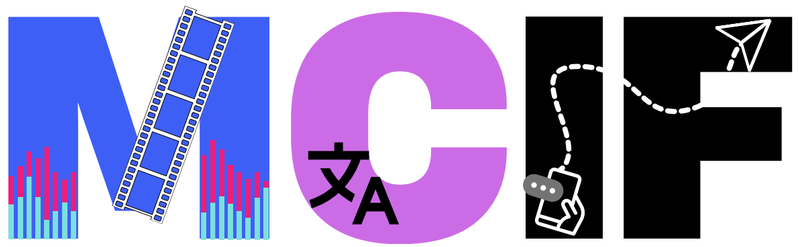
Models must ***interpret*** the **task** from the instruction

...but Real World is: Long-form

The input context can be a **long stream** of text, speech, or video

Models must ***identify*** the **relevant content** to fulfill the instruction request

A new benchmark



- Multimodal
- Crosslingual
- Instruction Following

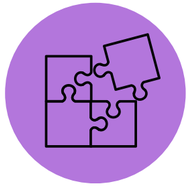
Core Design Principles

Human Annotation and Natural Instructions

fix and **mix** prompts

Native Multimodal Long-form content

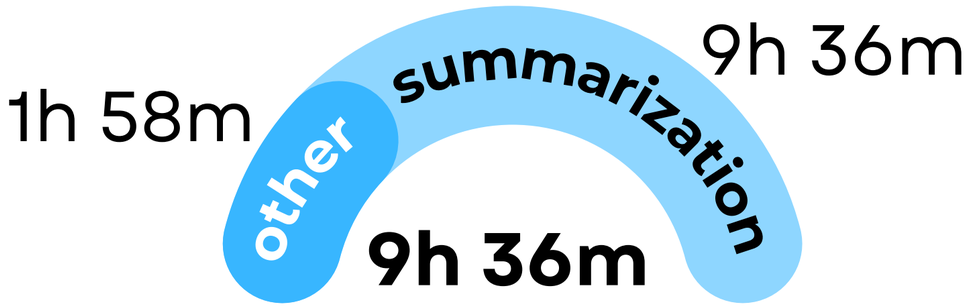
from **Scientific Talks**



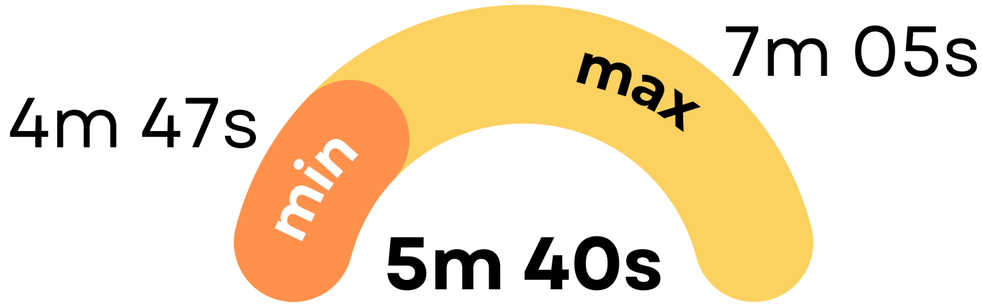
Parallel across modalities, languages, and context types

short- and **long**-form

Data Overview



Total Duration



Average Duration

Natively **long-form**, automatically segmented for **short-form** (~16 seconds)

Which Models?

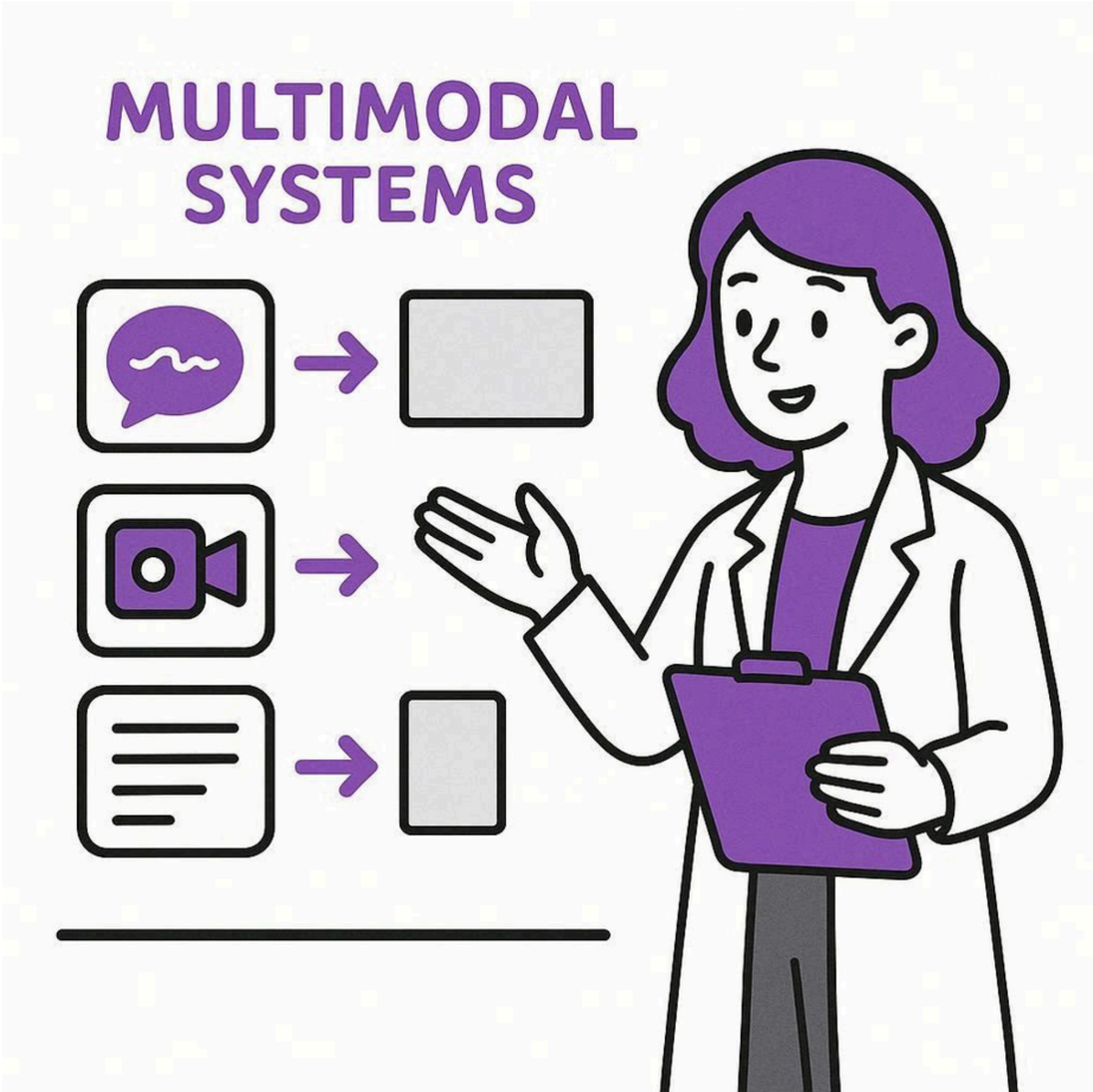
23 state-of-the-art systems

Models with:

- Open weights on HuggingFace
- <20B parameters

+ Gemini 2.5 Flash

7 LLMs, 5 SpeechLLMs, 5 VideoLLMs, 6 MLLMs



Which Models?

Curious about the results? Visit our poster!

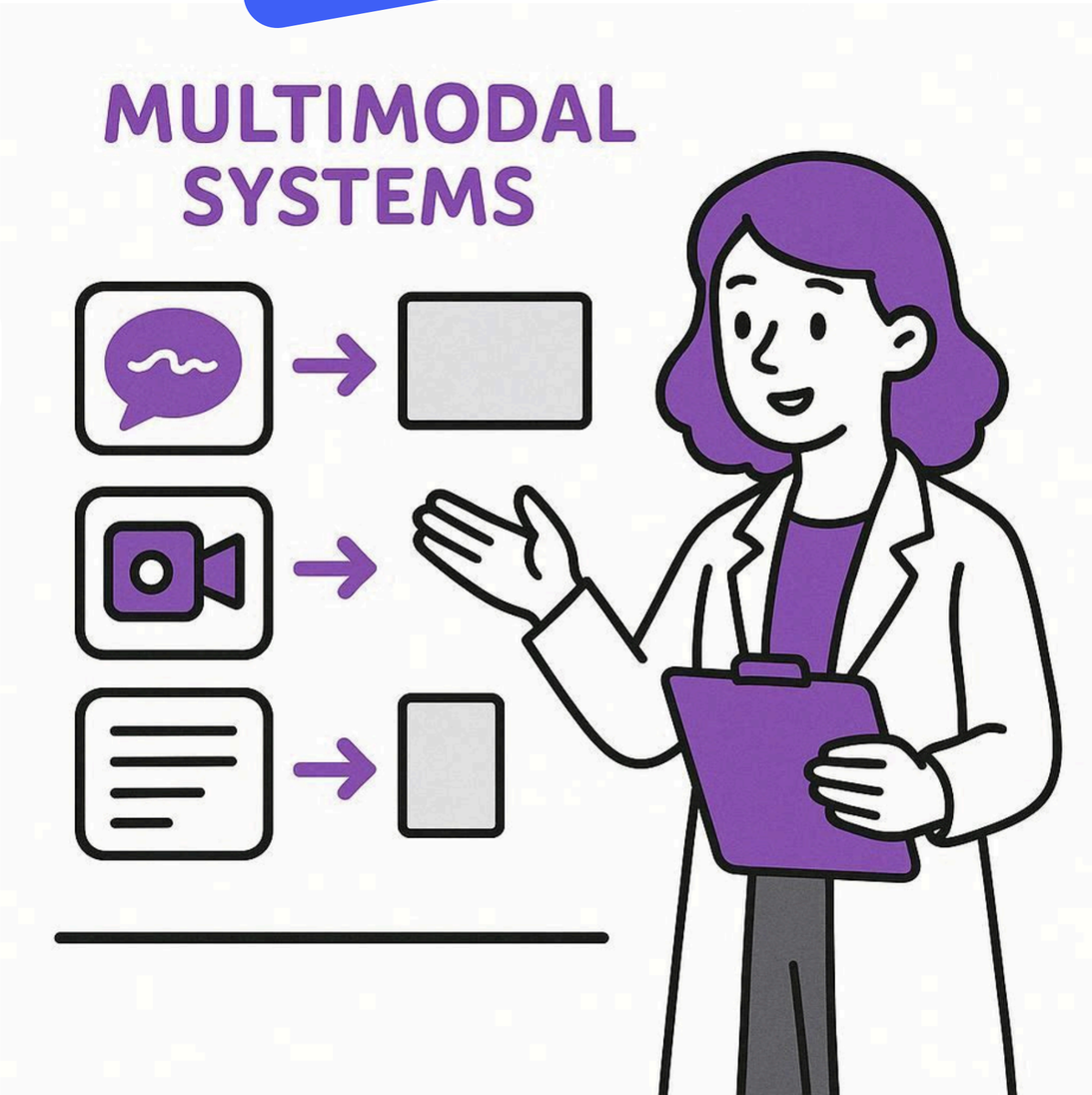
23 state-of-the-art systems

Models with:

- Open weights on HuggingFace
- <20B parameters

+ Gemini 2.5 Flash

7 LLMs, 5 SpeechLLMs, 5 VideoLLMs, 6 MLLMs



Thanks for your attention

hf.co/datasets/FBK-MT/MCIF

github.com/hlt-mt/mcif