SESEMMI FOR LinkedMusic: DEMOCRATIZING ACCESS TO MUSICAL ARCHIVES VIA LARGE LANGUAGE MODELS
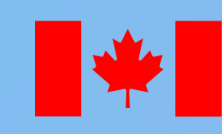
Liam Pond[1, 2], Simon Ngassam[1]*, Linnea Kirby[1]*, Sichen Meng[1]*, Sebastien Chow[1]*, Dylan Hillerbrand[1], Ichiro Fujinaga[1, 2]

[1] Schulich School of Music, McGill University
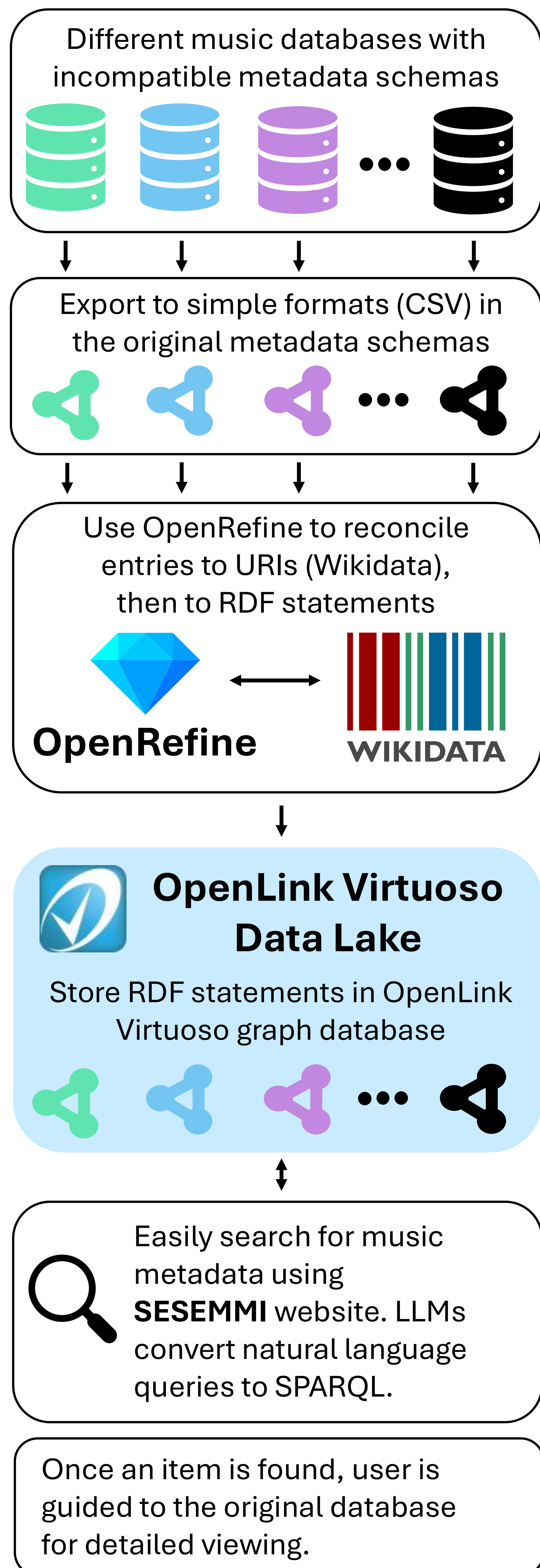[2] The Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
*Denotes equal contribution

Introduction

- There are over 100 music metadata databases online, making large-scale, cross-cultural, or longitudinal research time-consuming
- The LinkedMusic project aims to merge each of these in a single RDF data lake, queryable using SPARQL
- To facilitate querying for non-technical users, this study investigates the ability for LLMs to translate musical natural language queries (NLQ) into SPARQL queries

LinkedMusic Process

Different music databases with incompatible metadata schemas

Export to simple formats (CSV) in the original metadata schemas

Use OpenRefine to reconcile entries to URIs (Wikidata), then to RDF statements

OpenRefine ↔ **WIKIDATA**

OpenLink Virtuoso Data Lake

Store RDF statements in OpenLink Virtuoso graph database

Easily search for music metadata using **SESEMMI** website. LLMs convert natural language queries to SPARQL.

Once an item is found, user is guided to the original database for detailed viewing.

Methodology

- Built custom dataset of 20 NLQ/SPARQL pairs with ground truth SPARQL
- Designed general all-purpose prompt to guide the LLM to convert NLQ to SPARQL
- Five LLMs tested: Claude Sonnet 4, Gemini 2.5 Flash, Gemini 2.5 Pro, GPT-4o, OpenAI o4-mini
- One-shot and zero-shot tests
- Ontology provided in Turtle format
- Each evaluated three times in the browser

Four challenge types (five questions each):

Challenge 1:
- Retrieve information that can be found on a single sub-database's website
- **Example:** Find all compositions by William Byrd in DIAMM

Challenge 2
- Retrieve information that is stored in a single sub-database but cannot be found through the website
- **Example:** Find all different time signatures for jigs in The Session
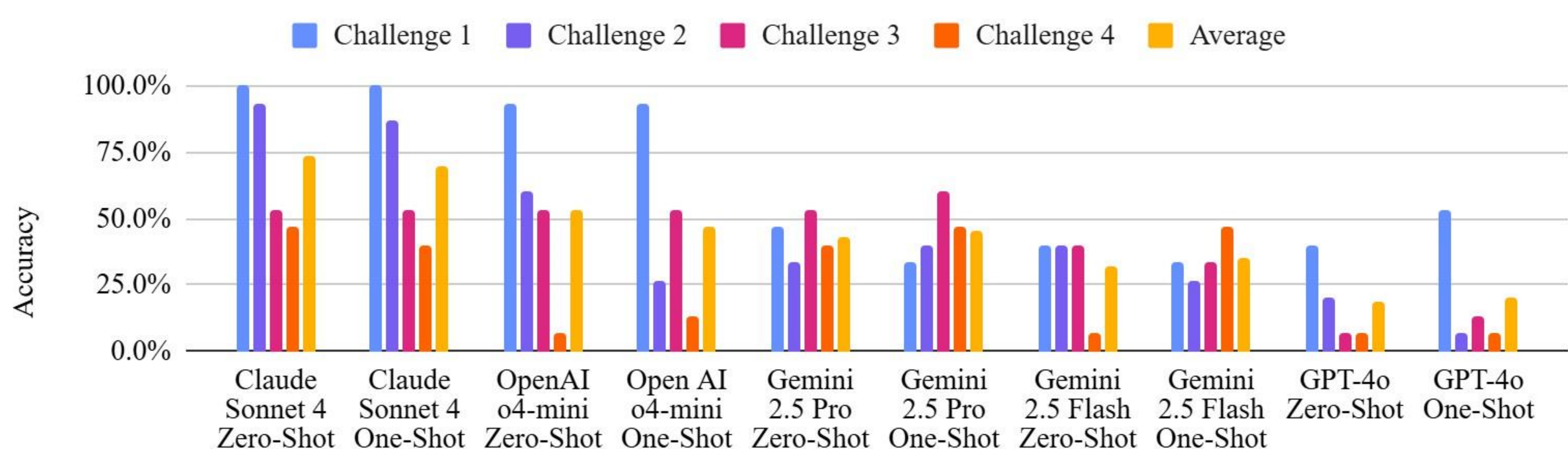
Challenge 3
- Retrieve anything that can be found on a single sub-database plus Wikidata
- **Example:** Find the average number of record labels that female singers in MusicBrainz have signed with

Challenge 4
- Retrieve any information in the entire LinkedMusic data lake and Wikidata
- **Example:** Find all works commissioned by Isabella d'Este that have a surviving manuscript and a recording made after 1980

Results

Databases

Over **352 million** RDF triples in five databases:

MusicBrainz

DIAMM **THE SESSION**

The Global Jukebox

Dig That Lick 1000

Three databases added since submission:

RISM Répertoire International des Sources Musicales

SIMSSA Database

cantus

Ontology

Difficulties

- Errors traversing sub-graphs
- LLMs often misunderstood how entities were reconciled
- Limitations applying Wikidata schema to sub-databases
- Wikidata "Q" identifiers often incorrect
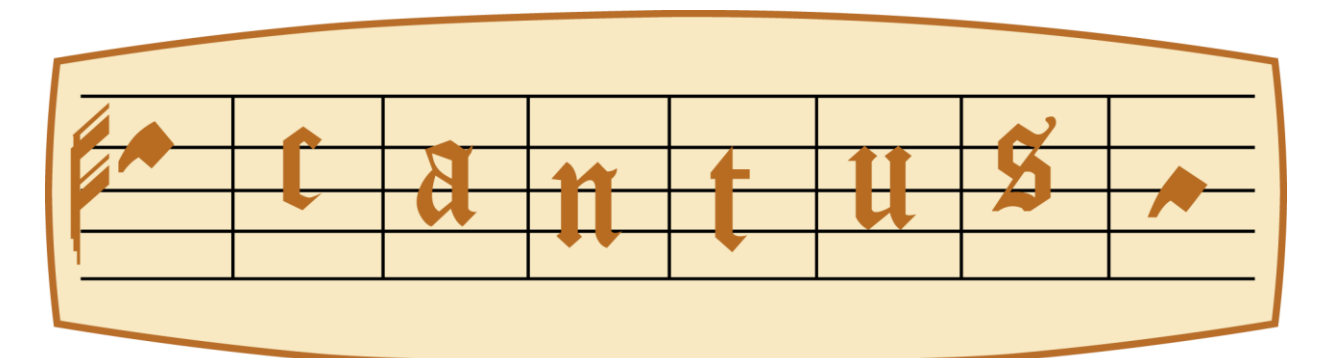- Moderate variance between attempts
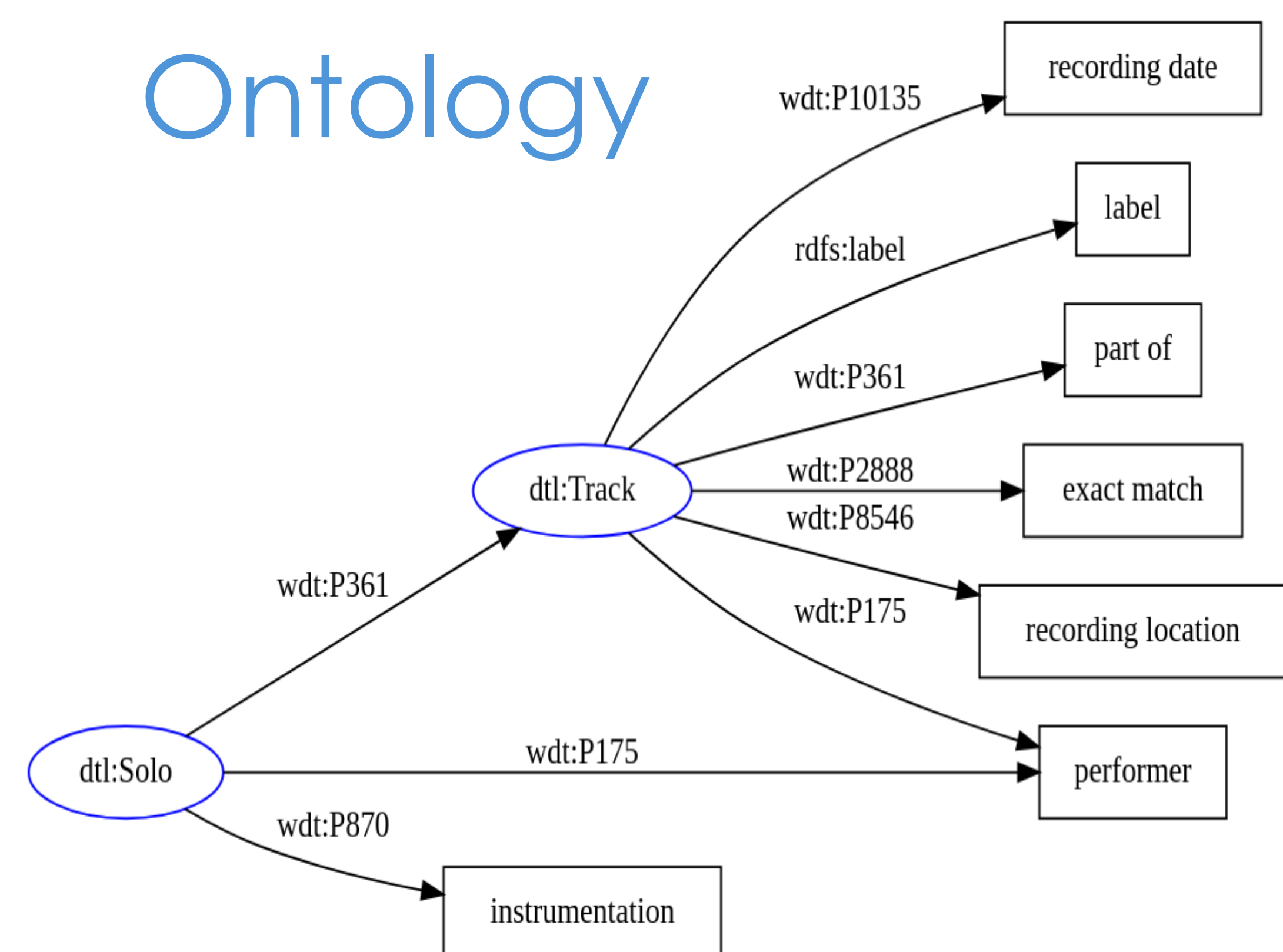
Conclusion

- First systematic investigation of NLQ to SPARQL in the music domain
- Performance decreases significantly with query complexity
- One-shot prompting did not consistently improve results
- Highest accuracy: Claude Sonnet 4 zero-shot (73.3% average, 100.0% for challenge 1 questions, 46.7% for challenge 4 questions)