Multimodal Large Language Model-Assisted Metadata Extraction from Historical Concert Programmes (1872–1928)

Sebastian Oliver Eck
Faculty of Music
University of Oxford
Oxford, United Kingdom
sebastian.eck@music.ox.ac.uk

Kevin R. Page
Oxford e-Research Centre
University of Oxford
Oxford, United Kingdom
kevin.page@oerc.ox.ac.uk

Abstract

Historical concert programmes are widely recognised as valuable sources for musicological historiography; yet, although many collections are increasingly becoming available online, identifying relevant programmes within larger archival holdings remains difficult. When present, descriptive metadata is often provided only at the collection level rather than for individual items. Even then, the absence of granular metadata at the level of individual programmes continues to preclude systematic analysis of their contents at scale. This paper reports the results of an experiment, testing whether general-purpose multimodal large language models (MLLMs) can assist in preparing semi-structured concert programme metadata as a first-pass metadata record ready for subsequent expert verification and reconciliation. Using a sample of 100 programmes (1872–1928) from three Oxford student music societies held at the Bodleian Libraries, we implement a lightweight workflow comprising low-cost image capture using everyday consumer hardware, minimal preprocessing, schema-constrained JSON output, and repeated sampling using a simple consensus strategy of extracted metadata fields. We compare MLLM outputs against a manually curated reference dataset using Levenshtein distance as a character-level error measure, treating extraction correctness primarily in terms of sufficient intelligibility and data consistency for future reconciliation. We find that MLLM output quality depends strongly on model choice and on expert-controlled design decisions (including hierarchical JSON schema definition and prompt specification). We conclude with practical recommendations for institutions and researchers who wish to treat MLLMs as assistive components in ephemera metadata creation workflows, while retaining expert authority over final metadata records.

CCS Concepts

• **Applied computing** → **Document metadata**; *Annotation*; **Digital libraries and archives**.

Keywords

concert programmes, music ephemera, metadata extraction, MLLMs

DLfM 2026, Thessaloniki, Greece

ACM ISBN 979-8-4007-2369-8/26/07
https://doi.org/10.1145/3815723.3815726

ACM Reference Format:
Sebastian Oliver Eck and Kevin R. Page. 2026. Multimodal Large Language Model-Assisted Metadata Extraction from Historical Concert Programmes (1872–1928). In *13th Annual Conference on Digital Libraries for Musicology (DLfM 2026), July 02, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3815723.3815726

1 Background and Motivation

Ephemera, described by the poet and clergyman George Crabbe (1754–1832) as something "born [/] To die before the next revolving morn" [31], have long been recognised as "documentary evidence of the culture of the everyday" [41]. Similarly, *music ephemera* preserve "a wealth of information for historical research" [39] and have supported historiographic work on performance and concert programming practices, institutional history, and canonic formation. Exemplary studies drawing extensively on concert ephemera as primary sources are [10, 11, 14, 26–28, 32, 42, 46, 47].

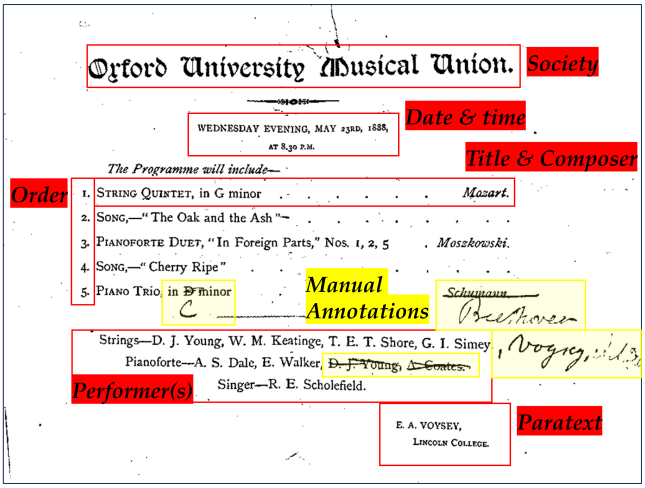

Figure 1: Manually Annotated Concert Programme Metadata: *Oxford University Musical Union*; image derived from: G.A. Oxon 4° 291/2 [8], © Bodleian Libraries, University of Oxford.

This study focuses on *concert programmes*, a prominent form of such ephemera: although printed in large numbers and widely disseminated from the early modern period onward, they were "things never intended to be preserved for posterity" [13] and typically discarded once their immediate function had been fulfilled. However, where concert programmes survive, they record repertory choice and order, name performers, or document unforeseen

changes (cf. an attendee's corrections in Figure 1) at a granularity often absent from official archival records *intended* for long-term preservation. At the same time, this heterogeneous information is rarely expressed in machine-explicit (semi-)structured or tabular form that would facilitate *computational extraction, classification,* and *analysis.* Rather, as shown in Figures 1 and 3, programmes rely on well-established human-readable layout conventions that encode structure and semantics *implicitly*, e.g., through typographic hierarchy, spatial arrangement, positional cues, and recurrent formatting patterns – including manual annotations (explored in §1.3).

1.1 Challenges for Music Ephemera Scholarship

Despite this recognised value, the informational contents of concert ephemera remain difficult to access and use within music historiography. Already in 1981, Fuld stated that such materials "do not seem to have achieved a status in music libraries comparable to their usefulness" and were often "not [..] well organized, preserved, or completely catalogued" [17]; almost three decades later, Bashford, in her essay *'Writing (British) Concert History: The Blessing and Curse of Ephemera'* (2008), confirmed these as "a source of almost unimaginable richness" [4]; yet, emphasised once more the persistent difficulty of locating and working with this source type, especially at the level of individual concert programme items [4]. Three factors, in combination, contribute to these issues:

(1) **Schema Mismatch** — the lack of specialised standards and the poor semantic fit of bibliographic schemas, e.g., MARC21 [23], for concert programme metadata, which tends to collapse named entities and roles, e.g., composer vs. performer, into generic note fields without semantic differentiation;
(2) **Insufficient Metadata** — digitisation at scale without corresponding item-level metadata, producing digital surrogates of concert ephemera that are hard to retrieve or access [12], often reinforcing the default archival option of simply "putting it in a box" [37]; and
(3) **Lack of Unified Discovery Systems** — a fragmented landscape of concert programme metadata discovery systems that remain limited in scope and interoperability (Table 1).

Although more generalisable concert programme platforms operating at the collection [34, 35, 39] or individual concert event level [49] as well as recent projects centred on linked open data modelling [3, 30] have provided partial solutions to these issues, the integration of item-level metadata at scale still largely depends on manual expert annotation. In practice, this often requires substantial high-expertise intervention before data can be integrated into authoritative digital archives [2, 13]. The result is a fragmented landscape in which only a fraction of historical concert programmes can be queried or analysed computationally at scale.

1.2 MLLMs in Cultural Heritage Applications

MLLMs [9] refer to transformer-based **large language models (LLMs)** [45] extended with the ability to "receive, reason and output with **multimodal** information." [52] As such, MLLMs can produce textual output, e.g., *object* or *document descriptions,* based on multimodal input, e.g., *textual instructions* and *images,* that they receive (Figure 2(c)). As *general-purpose models*, their application is not limited to a single document type or task and can be extended to

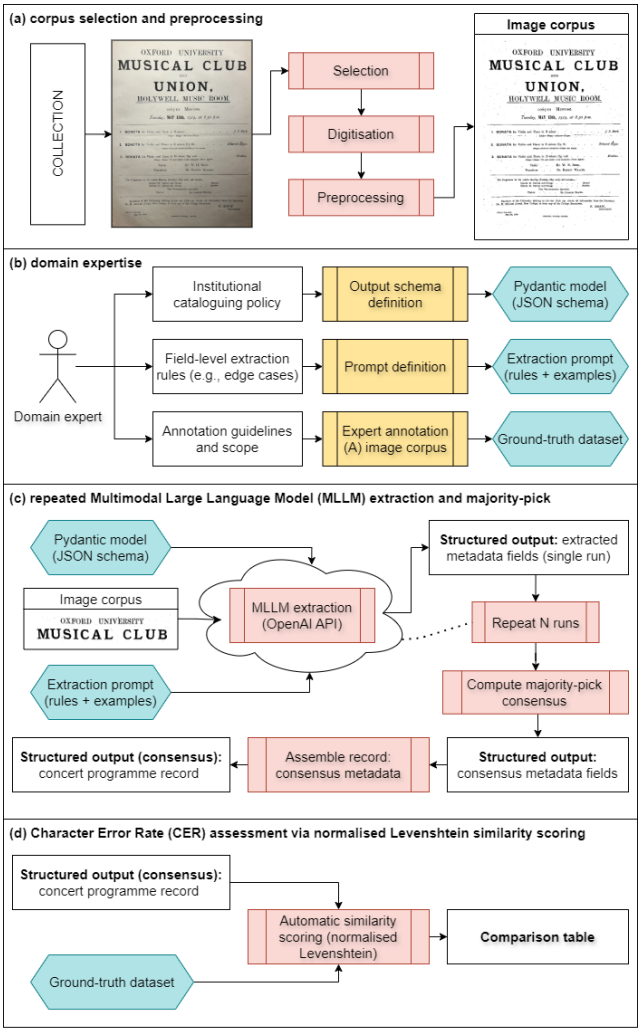

Figure 2: Workflow Overview: MLLM-assisted concert programme metadata extraction and assessment. Yellow: processes explicitly requiring *domain expertise*; image in (a): G.A. Oxon c.225/1 [6], © Bodleian Libraries, University of Oxford.

diverse sources. Adapting model behaviour to a specific use case, e.g., *metadata extraction,* generally requires either *fine-tuning* or delivering *extraction rules, constraints,* and *annotation examples* via *targeted prompting.* In-context prompting (ICP) relies entirely on such *carefully designed instructions – 'prompts'* – to steer MLLMs towards the desired outputs. By contrast, (supervised) fine-tuning [1] on annotated examples can improve annotation accuracy without *necessarily* relying on explicit extraction rules via targeted prompting, but it is time-consuming and computationally expensive.

Against this background, recent work within the Cultural Heritage (CH) domain has begun to explore how the interpretive and generative capabilities of (M)LLMs can support archival tasks that have traditionally required human high domain expertise [29]. Successful applications of MLLMs to heterogeneous historical texts

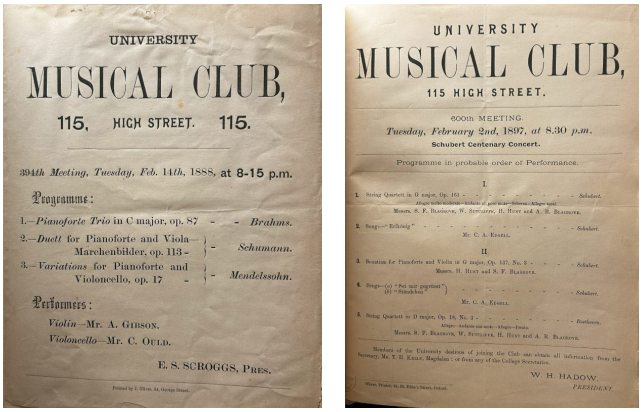

Figure 3: Two programme covers from the OUMC with a layout change in the mid-1880s (OUMC-I and OUMC-II); G.A. Oxon 4° 292/3 [7], © Bodleian Libraries, University of Oxford.

[18, 36, 43] indicate the potential of these general-purpose models for interpretive metadata extraction, particularly in settings where structure and semantics must be inferred rather than read from pre-delimited fields. Within this body of literature, Reusens et al. [33] demonstrated MLLMs' ability to derive keywords from curator-authored descriptions, while Groppe et al. [20] recently developed an agentic architecture that coordinated multiple MLLMs through federated intelligence [24] for automated metadata generation. Closest to the present use case, Xie et al. demonstrated schema-constrained extraction from digitised Swedish patent cards (1945–1975) using OpenAI's GPT-4o, combining structured output with error-flagging to support subsequent human verification [50].

1.3 Data Challenges when Applying MLLMs to Concert Programmes

While these studies establish the general potential of MLLMs for archival metadata extraction, most focus on sources with relatively regular or semi-structured layouts. Historic concert programmes, by contrast, represent a visually much more heterogeneous and structurally complex document type: unlike semi-structured patent cards as used by Xie et al. [50] – which consist largely of pre-delimited fields for *dates*, *names*, and *statuses* – concert programmes contain dense textual content in which semantics are conveyed visually through implicit layout conventions and typographic organisation rather than through pre-classified fields, as seen in Figures 1 and 3.

In this context, correctly identifying implicit hierarchical relationships (e.g., movements nested within works), inferring performer roles (e.g., pianist, violinist) of named entities, and disambiguating printed content from handwritten annotations requires visual and structural interpretation that integrates textual analysis with domain-specific musicological knowledge and familiarity with concert programme layout conventions. This combination seems well-matched to the capabilities of MLLMs, motivating our assessment through practical experimentation in the following sections.

2 Overview of the Experiment

This paper reports an experiment testing whether general-purpose multimodal large language models (MLLMs) can assist in extracting semi-structured *concert programme metadata* from digitised historical concert programmes.

Section 3 describes the selection, digitisation, and preprocessing of a controlled test corpus of historical English concert programmes from the Bodleian Libraries, University of Oxford (shown within the overall workflow in Figure 2(a)). Section 4 presents the MLLM-assisted extraction pipeline, including expert-guided schema-restricted model output, prompt specification, ground truth dataset creation (Figure 2(b)), and repeated sampling with majority-pick consensus (Figure 2(c)). Section 5 details the assessment methodology, comparing MLLM outputs against a manually curated ground truth dataset of 100 manually annotated concert programmes using normalised Levenshtein distance as a character-level error measure (Figure 2(d)). Finally, Section 6 analyses common error sources, identifies stages at which domain expertise remains decisive (e.g., schema design, image preparation, prompt specification, list-handling rules, and validation), and distils practical insights for institutions and researchers considering the applications of MLLMs on this source type.

Importantly, this study is not designed as a comprehensive benchmark of MLLMs, but as an exploratory investigation into their suitability for structured metadata extraction from concert programmes. Accordingly, the goal of our assessment was not broad model coverage or generalisation, but to provide a compact, repeatable experimental pipeline and an empirically grounded account of where, and under what conditions, MLLM-assisted concert programme metadata extraction and classification succeeds or fails in this task setting. Beyond reporting this investigation, our study introduces an hierarchically structured JSON schema (*concert-level* vs. *work-level fields*) as a reusable starting point for comparable experiments on different concert programme collections.[1]

3 Image Data Selection and Preparation

In this section we describe the source materials which, captured as digital images, form the basis for our experimental metadata extraction. As summarised in Figure 2(a), programmes were manually selected, digitised, and preprocessed before model inference (§4).

3.1 Historic Concert Programmes (1872–1928)

The study uses historic concert programmes (1872–1928) held at the Bodleian Libraries, Oxford, documenting programming practices of three student music societies, once central to Oxford University's late-19th and early-20th-century musical life [21]:

(1) *Oxford University Musical Club* (OUMC; 1872–1916) [7],
(2) *Oxford Univ. Musical Union* (OUMU; 1884–1916) [8], and
(3) *Oxford Univ. Musical Club & Union* (OUMCU; 1916–1928) [6].

Like many concert-ephemera collections, these materials are currently catalogued only at shelf-mark level with minimal metadata,

[1] Reusable code (and sample data) is available for adaptation at: https://github.com/oerc-music/concert_programmes_MLLM while a snapshot also including the experimental data used for the assessment described in this paper (inc. source data, ground truth, MLLM outputs, and assessment statistics) is archived at: https://dx.doi.org/10.5287/ora-n6yg5jgqy

```
ConcertProgramme
   concerts   [1...n]
      {
        c_position     int|null,
        c_title        str|null,
        c_date         str|null,
        c_time         str|null,
        c_venue        str|null,
        c_society      str|null,
        c_series_no    str|null,
        c_comp_list    [str]|null,
        c_perf_list    [{name, roles?}]|null,
        c_works        [1...n]
             {
               w_position  int|null,
               w_title     str|null,
               w_comp      str|null,
               w_movements [{m_position, m_title}]|null,
               w_perf_list [{name, roles?}]|null
             }
      }
```

Figure 4: Simplified hierarchical Pydantic output schema for schema-constrained metadata extraction (fields per Table 2) using Multimodal Large Language Models (MLLMs), separating *concert-level* (`c_*`) and *work-level* (`w_*`) metadata.

primarily for inventory purposes [6–8, 12]. This lack of granular metadata limits discoverability and usability of these materials and motivated the present experiment.

The programmes survive in at least twelve bound volumes within the *Gough Adds.* collection; nine selected volumes contain ≈2,593 printed, one-sided programmes with generally consistent typography and good visual condition (Figure 1). Across the corpus, layout is sufficiently regular to support controlled assessment, while still exhibiting meaningful variation: OUMC shows a clear mid-period layout shift (OUMC-I vs. OUMC-II); together with OUMU and OUMCU this yields four distinct layout groups (Figure 3).

3.2 Digitisation and Preprocessing

Programmes were captured using a smartphone camera at a native resolution of 4032 × 3024 px in fast succession; images that were blurred or exhibited cropped content were excluded. To obtain a controlled test set while retaining edge cases – e.g., *manual annotations* or *atypical layouts* – 25 programmes per layout group were manually sampled (100 total). Images were binarised using adaptive Gaussian thresholding via `opencv-python 4.12.0.88`, resized to 2048 × 768 px, and converted to PNG (1-bit) to comply with OpenAI batch processing file size limits (200 MB per batch; 1,000 requests for 100 images × 10 variants; cf. §4.3); in some cases, binarisation also improved readability of manual annotations (Figure 5).

3.3 Final Corpus and Ground Truth

Preprocessing reduced typical file sizes from 2-3 MB to 10-40 kB, enabling efficient batch processing and inclusion of the binarised image (base64 encoding) alongside extracted metadata in the final deliverable. All 100 images were manually annotated by a trained musicologist against the custom `Pydantic` JSON schema (4.1), producing a ground truth dataset of 3,939 reference values used for subsequent automatic MLLM-output assessment (§5).

4 Methodology and Pipeline Design

The experiment was designed to test whether general-purpose MLLMs can produce usable first-pass, item-level concert programme metadata at scale. The pipeline enforces structured output via a custom hierarchical Pydantic schema, requiring MLLMs to express their interpretation and classification of concert programme metadata as schema-valid JSON.

4.1 Schema-Constrained Structured Output

Concert programmes encode implicit hierarchical semantics through explicit layout: a document describes an event; an event contains an ordered sequence of works; works may include internal structure (e.g., movements) and performer attributions (Table 2). These hierarchical relationships can be modelled as a JSON (JavaScript Object Notation) object (Figure 4). *High-level concert elements* (`c_*`) contain structured information on *lower-level work elements* (`w_*`), extending the *programme–event distinction* formulated by Lee [22]:

- **Concert-level** (`c_*`): metadata describing ≥ 1 distinct *concert events* (e.g., `c_date`) in each *programme document;* and
- **Work-level** (`w_*`): metadata describing ≥ 1 individual works listed within each distinct *concert event* (e.g., `w_movements`).

Defining a structured output schema and delivering it to the MLLM during inference, alongside the prompt and image input (Figure 2(c)), forces the model to produce data in the desired hierarchical format, rather than unstructured text. In this experiment, JSON fields were derived from a comparative survey of eight concert programme databases [15, 19, 25, 38, 40, 44, 48, 49]. Field selection balanced general applicability with semantic granularity, excluding rarely occurring fields, such as `attendees` (Table 1). In our experimental pipeline, each extracted metadata value is represented by a unique quadruple (*image, concert, work, field*), as shown in Table 3. This output schema also allowed for automatic assessment against the manually annotated ground truth dataset (§3.3).

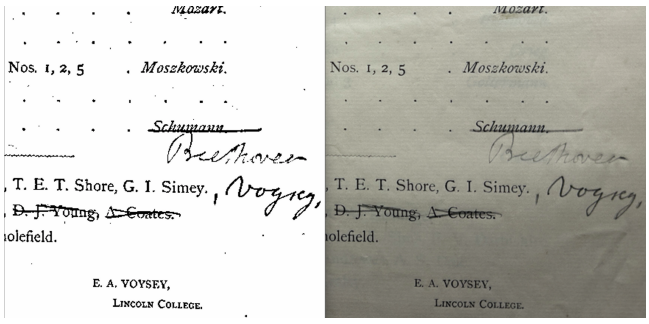

Figure 5: Preprocessing of an Annotated Concert Programme: original (right; cf. Figure 1); pre-processed, compressed (left).

4.2 Prompt Specification as Domain Encoding

We used in-context prompting (ICP) with few-shot examples [51]. In this study, each extraction request combined an input image, a `Pydantic` model specifying the required structured output (Figure 2(c)), and a textual prompt that consisted of these three elements: (i) field-specific metadata extraction and classification rules; (ii) the schema definition; and (iii) annotated examples.

Table 1: Metadata Field Coverage Across *Selected Concert Programme Platforms* – Comparison of Field Availability and Selection.
TRUE **= field present;** ***FALSE*** **= absent;** ***UNKN*** **= unclear (data cut-off: 08/2025).** [a] **Venue fields combine venue & location.** [b] **Organiser, e.g., 'Königliches Konservatorium der Musik zu Leipzig', stored as corporate body [= Körperschaft] rather than "society".** [c] **Performer listed under work titles** [d] **Includes soloists & ensembles, e.g., orchestras.** [e] **Performer list recorded at work-level.** [f] **Images only partially available / linked.** [g] **Source files are included in the final deliverable as base64 images.**

Database/Project	c_title	c_season	c_ID	c_date	c_time	c_venue	c_society	c_perf_list[c]	c_pers_role
Prague Concert Life 1850–1881 [40]	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
Felix Meritis 1832–1888 [44]	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
The Carnegie Hall Archives [19]	FALSE	FALSE	UNKN	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
The NYPhil Digital Archives [38]	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE[a]	FALSE	TRUE[d]	TRUE
musiconn.performance [49]	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE[b]	TRUE	TRUE
Wiener Phil. Concert Archive [48]	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE[d]	TRUE
Hallé Digital Repertoire Database [25]	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
Archiv des Konzertlebens [15]	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
Selected Fields	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Database/Project	c_attendees	w_order	w_title	w_comp	w_perf_list[e]	notes/comm	SOURCES	IMAGES	PROJECTS
Prague Concert Life 1850–1881 [40]	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
Felix Meritis 1832–1888 [44]	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
The Carnegie Hall Archives [19]	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
The NYPhil Digital Archives [38]	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
musiconn.performance [49]	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
Wiener Phil. Concert Archive [48]	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Hallé Digital Repertoire Database [25]	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Archiv des Konzertlebens [15]	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE[f]	FALSE
Selected Fields	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE[g]	FALSE

Table 2: Hierarchical JSON Schema for Schema-Constrained MLLM Output (some examples drawn from Figure 3: right). Field-wise extraction rules (verbatim capture vs. normalisation) were specified in the extraction prompt (Figure 2(c)).

Field	Type	Example	Description / rule
Document-level			
`concerts`	List[ConcertEvent]	`[{...}]`	Concert events on the programme.
Concert-level			
`c_position`	Int	`1`	Concert order (≥ 1).
`c_title`	String	`"Schubert Centenary Concert"`	Event title (if present; as printed).
`c_date`	String	`"1897-02-02"`	Date (ISO 8601: YYYY-MM-DD).
`c_time`	String	`"20:30"`	Start time (ISO 8601: hh:MM).
`c_venue`	String	`"115 High Street"`	Venue / address line.
`c_society`	String	`"University Musical Club"`	Organiser / society name.
`c_series_no`	String	`"600th"`	Series/meeting number.
`c_comp_list`	List[String]	`["Schubert", "Beethoven", "J. S. Bach"]`	Concert-level composer block.
`c_perf_list`	List[Performer]	`[{"name": name, "roles": roles}, ...]`	Concert-level performer block.
`name`	String	`"S. F. Blagrove"`	Performer name.
`roles`	List[String]	`["Violin", "Voice"]`	Instrument/voice/role label(s).
`c_works`	List[Work]	`[{w_position: 1, ...}]`	Works in performance order (≥ 1).
Work-level			
`w_position`	Int	`1`	Work order within concert.
`w_title`	String	`"String Quartett in G major, Op. 161"`	Work title line (incl. key/Op./cat.).
`w_comp`	String	`"Schubert"`	Composer attached to work.
`w_movements`	List[Movement]	`[{"m_position": m_position, "m_title": m_title}, ...]`	Movements/sections printed under the work.
`w_perf_list`	List[Performer]	`[{"name": name, "roles": roles}, ...]`	Performers printed under the work.
Movement-level			
`m_position`	Int	`1`	Movement order within work.
`m_title`	String	`"Allegro molto moderato"`	Movement/section heading.

Table 3: Ground Truth Metadata Record and GPT-5 Annotation for G.A. Oxon 4° 292/1-5 [7] – Assessment file (excerpt), showing structured fields (`c_*`, `w_*`) for `gpt-5-2025-08-07`; corresponding to the manually annotated concert programme (cf. Figure 1).

ID	run_id	image	concert	work	field	GT_value	GT_ID	pred_value	pred_ID	pred_share	sim_score	match_100	match_90	match_80	match_70
1	09_gpt-5- [...]	IMG_7853	1		c_position	1	680	1	693	1.00	1	TRUE	TRUE	TRUE	TRUE
2	09_gpt-5- [...]	IMG_7853	1		c_date	1888-05-23	681	1888-05-23	694	1.00	1	TRUE	TRUE	TRUE	TRUE
3	09_gpt-5- [...]	IMG_7853	1		c_time	20:30	682	20:30	695	1.00	1	TRUE	TRUE	TRUE	TRUE
4	09_gpt-5- [...]	IMG_7853	1		c_society	Oxford University Musical Union	683	Oxford University Musical Union	696	0.90	1	TRUE	TRUE	TRUE	TRUE
5	09_gpt-5- [...]	IMG_7853	1		c_perf_list[1]	D. J. Young, Strings	684	D. J. Young, Strings	697	1.00	1	TRUE	TRUE	TRUE	TRUE
6	09_gpt-5- [...]	IMG_7853	1		c_perf_list[2]	W. M. Keatinge, Strings	685	W. M. Keatinge, Strings	698	0.90	1	TRUE	TRUE	TRUE	TRUE
7	09_gpt-5- [...]	IMG_7853	1		c_perf_list[3]	T. E. T. Shore, Strings	686	T. E. T. Shore, Strings	699	1.00	1	TRUE	TRUE	TRUE	TRUE
8	09_gpt-5- [...]	IMG_7853	1		c_perf_list[4]	G. I. Simey, Strings	687	G. I. Simey, Strings	700	1.00	1	TRUE	TRUE	TRUE	TRUE
9	09_gpt-5- [...]	IMG_7853	1		c_perf_list[5]	Voysey, Strings	688	A. S. Dale, Pianoforte	701	1.00	0.303	FALSE	FALSE	FALSE	FALSE
10	09_gpt-5- [...]	IMG_7853	1		c_perf_list[6]	A. S. Dale, Pianoforte	689	E. Walker, Pianoforte	702	1.00	0.789	FALSE	FALSE	FALSE	TRUE
11	09_gpt-5- [...]	IMG_7853	1		c_perf_list[7]	E. Walker, Pianoforte	690	E. F. Young, Pianoforte	703	0.30	0.667	FALSE	FALSE	FALSE	FALSE
12	09_gpt-5- [...]	IMG_7853	1		c_perf_list[8]	R. E. Scholefield, Singer	691	A. Coates, Pianoforte	704	0.40	0.39	FALSE	FALSE	FALSE	FALSE
13	09_gpt-5- [...]	IMG_7853	1		c_perf_list[9]	$MISSING$		R. E. Scholefield, Singer	705	0.80	0	FALSE	FALSE	FALSE	FALSE
14	09_gpt-5- [...]	IMG_7853	1	1	w_position	1	692	1	706	1.00	1	TRUE	TRUE	TRUE	TRUE
15	09_gpt-5- [...]	IMG_7853	1	1	w_title	String Quintet, in G minor	693	String Quintet, in G minor	707	1.00	1	TRUE	TRUE	TRUE	TRUE
16	09_gpt-5- [...]	IMG_7853	1	1	w_comp	Mozart	694	Mozart	708	1.00	1	TRUE	TRUE	TRUE	TRUE
17	09_gpt-5- [...]	IMG_7853	1	2	w_position	2	695	2	709	1.00	1	TRUE	TRUE	TRUE	TRUE
18	09_gpt-5- [...]	IMG_7853	1	2	w_title	Song,—"The Oak and the Ash"	696	Song,— "The Oak and the Ash"	710	0.90	0.98	FALSE	TRUE	TRUE	TRUE
19	09_gpt-5- [...]	IMG_7853	1	3	w_position	3	697	3	711	1.00	1	TRUE	TRUE	TRUE	TRUE
20	09_gpt-5- [...]	IMG_7853	1	3	w_title	Pianoforte Duet, "In Foreign Parts," Nos. 1, 2, 5	698	Pianoforte Duet, "In Foreign Parts," No. 1	712	0.60	0.937	FALSE	TRUE	TRUE	TRUE
21	09_gpt-5- [...]	IMG_7853	1	3	w_comp	Moszkowski	699	Moszkowski	713	1.00	1	TRUE	TRUE	TRUE	TRUE
22	09_gpt-5- [...]	IMG_7853	1	4	w_position	4	700	4	714	0.80	1	TRUE	TRUE	TRUE	TRUE
23	09_gpt-5- [...]	IMG_7853	1	4	w_title	Song,—"Cherry Ripe"	701	Pianoforte Duet, "In Foreign Parts," No. 2	715	0.60	0.226	FALSE	FALSE	FALSE	FALSE
24	09_gpt-5- [...]	IMG_7853	1	4	w_comp	$MISSING$		Moszkowski	716	0.60	0	FALSE	FALSE	FALSE	FALSE
25	09_gpt-5- [...]	IMG_7853	1	5	w_position	5	702	5	717	0.80	1	TRUE	TRUE	TRUE	TRUE
26	09_gpt-5- [...]	IMG_7853	1	5	w_title	Piano Trio, in C minor	703	Pianoforte Duet, "In Foreign Parts," No. 5	718	0.60	0.517	FALSE	FALSE	FALSE	FALSE
27	09_gpt-5- [...]	IMG_7853	1	5	w_comp	Beethoven	704	Moszkowski	719	0.60	0.105	FALSE	FALSE	FALSE	FALSE
28	09_gpt-5- [...]	IMG_7853	1	6	w_position	$MISSING$		6	720	0.67	0	FALSE	FALSE	FALSE	FALSE
29	09_gpt-5- [...]	IMG_7853	1	6	w_title	$MISSING$		Song,— "Cherry Ripe"	721	1.00	0	FALSE	FALSE	FALSE	FALSE
30	09_gpt-5- [...]	IMG_7853	1	7	w_position	$MISSING$		7	722	0.67	0	FALSE	FALSE	FALSE	FALSE
31	09_gpt-5- [...]	IMG_7853	1	7	w_title	$MISSING$		Piano Trio, in D minor	723	1.00	0	FALSE	FALSE	FALSE	FALSE
32	09_gpt-5- [...]	IMG_7853	1	7	w_comp	$MISSING$		Schumann	724	1.00	0	FALSE	FALSE	FALSE	FALSE

A trained musicologist familiar with cataloguing conventions (i) translated these conventions into field-wise extraction and classification rules, (ii) abstracted concert programme semantics into a hierarchically structured JSON schema, and (iii) curated mock examples covering representative edge cases to steer MLLMs towards the intended output via in-context prompting (as explained in §1.2).

4.3 Repeated Sampling and "Majority-Pick"

Because MLLM outputs are inherently non-deterministic, each model was run ten times per programme image under almost identical runtime settings (temperature = 0 where supported; reasoning-models `o3`, `o4-mini`, and `GPT-5` require = 1, limiting comparability). For each unique metadata quadruple (§4.1), we selected the most frequent value as the consensus ("majority pick"; Figure 2(c)) and recorded its prediction share as a proxy for MLLM agreement:

$$\texttt{pred\_share} = \frac{\text{count}(n_{\text{maj}})}{n_{\text{variants}}} \tag{1}$$

This procedure was intended to stabilise MLLM outputs. However, in practice, `pred_share` might also provide a lightweight triage signal, where low-share fields can be prioritised for manual expert-guided review and reconciliation.

5 Comparative Assessment of Models

Metadata extraction was executed for eleven off-the-shelf OpenAI MLLM snapshots (Figure 6) that supported multimodal input and schema-constrained JSON output at the experiment's conclusion (August 2025). We conducted an informed assessment of whether MLLM-extracted metadata was *sufficiently intelligible* for first-pass, schema-valid candidate records for expert verification and reconciliation. Minor syntactic deviations (e.g., *"Bethoven"* instead of *"Beethoven"*) were thus considered intelligible annotations that a skilled human annotator could plausibly correct. Following standard NLP practice [5, 16], we operationalised intelligibility using *normalised Levenshtein distance* [53] via character error rate (CER).[2]

5.1 CER-Based Similarity Scoring

We compared each model's majority-pick quadruple against the manually curated ground truth dataset using a normalised Levenshtein distance [53] as a CER measure (Figure 2(d)). This was intended to tolerate minor character-level spelling errors while penalising more substantive deviations. For each unique quadruple $r = (\textit{image}, \textit{concert}, \textit{work}, \textit{field})$ (Table 3), we computed a similarity score between ground truth values x_r and MLLM predictions y_r:[3]

$$\text{sim\_score}(x, y) = \begin{cases} 0, & \text{if } x \text{ or } y \text{ is } \texttt{\$MISSING\$}, \\ 1, & \text{if } \text{norm}(x) = \text{norm}(y), \\ 1 - \dfrac{\text{Lev\_dist}(\text{norm}(x),\ \text{norm}(y))}{\max(|\text{norm}(x)|,\ |\text{norm}(y)|)}, & \text{otherwise.} \end{cases} \tag{2}$$

Lev_dist denotes Levenshtein edit distance (implemented via `python-Levenshtein 0.27.1`). A custom implemented norm(·) function lowercased values and removed non-informative punctuation and repeated whitespace. The model was prompted to perform dates and times normalisation to ISO-8601 (`YYYY-MM-DD` and `hh:MM`) during extraction; both were scored as exact-match fields (any non-exact match resulted in a sim_score = 0).

[2] Yujian and Bo defined *Generalised Levenshtein Distance* (GLD) "as the minimum cost of transforming one string into another through a sequence of weighted edit operations", i.e, "the deletion, insertion, and substitution of individual symbols." [53]

[3] For list-typed fields, r was expanded element-wise with an index i, i.e., $(\ldots, \textit{field}, i)$ (e.g., `c_perf_list[1]`), and scores were computed per element (cf. Table 3: lines 5–13).

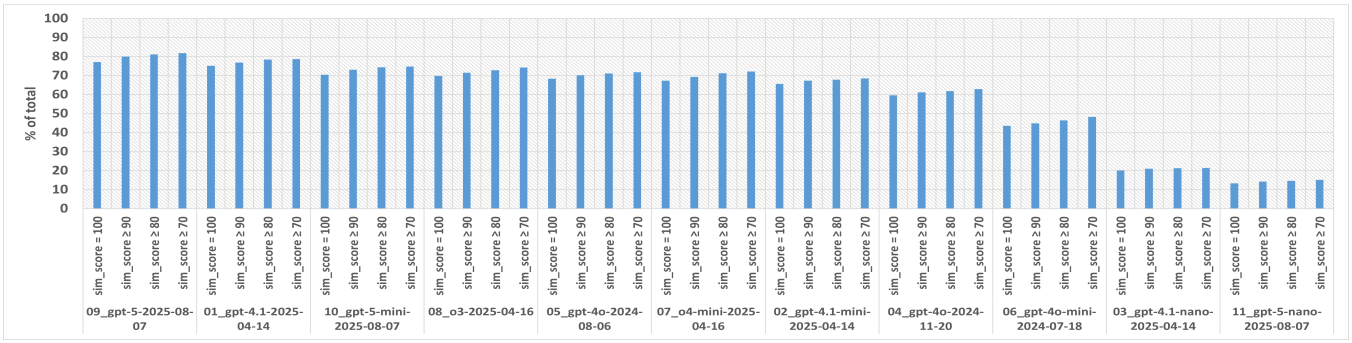

Figure 6: Model-level accuracy across assessed MLLMs, reported as the macro-average percentage of extracted metadata values whose normalised Levenshtein similarity score (`sim_score`) meets or exceeds each threshold ($t \in \{1.00, 0.90, 0.80, 0.70\}$).

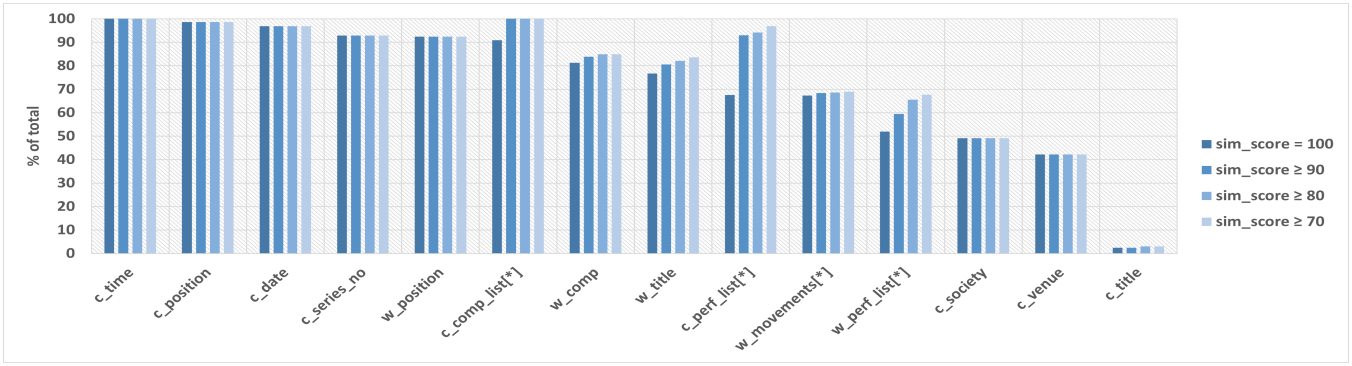

Figure 7: Field-level accuracy for `gpt-5-2025-08-07`, expressed as the percentage of extracted metadata values per field whose normalised Levenshtein similarity score (`sim_score`) meets or exceeds each threshold ($t \in \{1.00, 0.90, 0.80, 0.70\}$).

5.2 Aggregation and Reporting

Unless stated otherwise, values reported in Figures 6 and 7 denote the *percentage of extracted metadata values whose normalised similarity score (`sim_score`) meets or exceeds a given threshold* $t \in \{1.00, 0.90, 0.80, 0.70\}$. An MLLM prediction is counted as correct whenever $\text{sim\_score}(x_r, y_r) \geq t$.

Individual field types. For each metadata field type (e.g., `c_date`, `w_title`), field-level accuracy was computed as the proportion of all row-level comparisons for that field type that met the respective threshold, relative to the total number of comparisons for that field across the full set of 100 annotated programmes (Figure 7).

List aggregation. For list-typed fields, similarity scores were first computed element-wise over indices i and then aggregated under the corresponding list wildcard (e.g., `c_perf_list[*]`), yielding a single accuracy value per list-typed field.

MLLM comparison. Accuracy was assessed by row-wise aggregation of all field scores, reporting the proportion of rows whose similarity score meets the threshold relative to the total rows in the evaluation table (Figure 6). For list-type entries, only aggregated list wildcard labels (e.g., `c_perf_list[*]`) were used.

5.3 Indicative Performance of Selected MLLMs

Figure 6 reports aggregated field-level accuracies at the four similarity thresholds $t \in \{1.00, 0.90, 0.80, 0.70\}$. Across all eleven assessed models, `gpt-5-2025-08-07` achieved the highest accuracy at every threshold. Within the assessed set, this might indicate that selecting the largest model from the most recent MLLM generation maximises the likelihood of producing usable first-pass, schema-valid metadata. Figure 7 disaggregates GPT-5 performance by metadata field type. Semantically explicit concert-level fields (e.g., `c_date`, `c_time`) were extracted with consistently high accuracy. By contrast, structurally complex fields – most notably list-based work-level fields such as `w_movement[*]` and `w_perf_list[*]` – showed substantially lower accuracies. This seemed to be caused by the chosen automatic assessment strategy (§5), depending on exact index alignment for list-based metadata fields.

6 Error Sources and Practical Recommendations

Table 3 presents an excerpt from the assessment dataset for a single concert programme from the OUMU collection (source shown in Figure 1). The table compares manually curated ground truth metadata with schema-constrained MLLM annotations produced by `gpt-5-2025-08-07`. For each extracted field, it reports the majority-pick prediction (`pred_value`), its prediction share across repeated runs (`pred_share`), the normalised Levenshtein similarity score (`sim_score`), and the resulting threshold matches (`match_100`, etc.).

This example was deliberately selected for closer inspection as it exhibits several error modes observed across the dataset, largely stemming from *prompt underspecification, human oversight in ground truth annotation,* and *assessment limitations* – as detailed in the following sections – rather than incorrect model behaviour.

Prompt underspecification. A first class of mismatches arose from insufficient prompt specification. Handwritten corrections, e.g., corrected composer names and work titles (e.g., *"Piano Trio in D minor Schumann"* → *"Piano Trio in C minor Beethoven"*) or performer names (e.g., *"Young"*, *"Coates"*; cf. Figure 1), were consistently ignored during MLLM extraction due to the prompt's implicit prioritisation of printed text over manual annotations. Future applications should adjust extraction rules through targeted prompt refinement to explicitly define how such edge-cases, i.e., handwritten corrections, are to be interpreted for a given collection, with affected cases reprocessed once error sources have been identified.

Prompt–ground truth misalignment. Errors also arose where prompt-level extraction rules and human annotation diverged. Both annotator and MLLMs were instructed to distinguish works from movements as follows:

```
6. **Pieces vs. movements**
* If a title contains two or more works with **different opus
/ catalogue / part numbers** each (e.g. "BWV 871; Op. 45 No.
2"), always split into separate 'works', even if the composer
and performer are identical.
* Do not treat a movement heading (e.g. "Allegro") as a
standalone work.
* If ambiguity arises (e.g. individual songs inside a cycle),
treat each printed item as its own work.
```

Under this specification, Moszkowski's *Pianoforte Duet, "In Foreign Parts", Nos. 1, 2, and 5* was correctly decomposed by the MLLM (Table 3: lines 19–27), whereas the ground truth annotator collapsed these into a single entry (ibid.: lines 19–21), leading to false negatives. A similar issue affected `c_title`, causing consistently low accuracy scores (Figure 7): an in-context prompt example (see definition in §1.2) treated a meeting number (*"1100th Meeting"*) as the `c_title`, applied consistently by the MLLM but inconsistently by the annotator, producing systematic assessment mismatches. These two cases highlight the need to co-develop prompt specifications and annotation guidelines, ideally with inter-annotator validation, to avoid penalising correct model behaviour. Accordingly, until such refinements are made, this study's ground truth should be treated as a practical reference for indicative MLLM-output assessment rather than a ready-to-use benchmark dataset.

Positional assessment effects in list-typed fields. Finally, some mismatches are attributable to the positional automatic assessment of list-typed fields (§5). Elements i in `w_movement[i]` and `w_perf_list[i]`, for instance, were assessed index-wise; single omissions or insertions can shift indices and propagate apparent errors across otherwise correct entries (cf. Table 3: lines 9–13 and 23–31). This example represents an edge case: such effects were not observed consistently across the dataset, but occurred primarily in cases of incorrect work-splitting or list-type fields. Manual inspection indicates inflated error reporting rates for affected fields in the range of ca. 5–10%, although the exact number of cases is not known. This is amplified by the use of normalised Levenshtein distance: while appropriate for scalar fields (e.g., `c_date`), it is less suited to hierarchical or list-based data, where deviations can reflect misalignment rather than character-level differences. Accordingly, scores for list-type fields should be interpreted as indicative; set-based comparison (e.g., Jaccard similarity) would be a useful future refinement.

7 Conclusions and Future Work

Using a controlled sample of 100 Bodleian concert programmes, this experiment indicates that general-purpose MLLMs can assist with structured concert programme metadata creation. For a subset of fields, especially temporal and typographically stable concert-level fields, the assessed `gpt-5-2025-08-07` produced largely accurate metadata annotations (Figure 7). Persistent errors centre on list-structured semantics or stem from insufficient prompt design, ground truth annotation inconsistencies, and assessment artefacts rather than systematic model failure; even with these limitations, the majority of extracted records remain usable as basic item-level descriptions (e.g., date, time, works performed, performer names) and, in most cases, are suitable as first-pass candidate records for subsequent expert verification and reconciliation.

Practical implications. Batch-processing costs were moderate even for the most expensive model, GPT-5 ($65.21 for 1,000 requests). Combining low-cost image capture, preprocessing, schema-constrained output, with repeated sampling and majority-pick consensus (Figure 2) shifts effort from manual transcription to expert verification and correction. In practice, rather than fully automated extraction across entire concert programme collections, the greatest coverage gains are likely to come from pre-populating high-confidence fields and routing low-confidence or structurally complex cases to targeted expert review. Even before correction, such first-pass records can improve collection discoverability, provided they are clearly labelled as AI-generated metadata.

Reusable outputs and generalisation. This study has also created a hierarchically structured JSON schema (concert-level vs. work-level fields; Figure 4 & Table 3) intended as a reusable starting point; however, its applicability to concert programme collections with substantially different layout conventions is yet to be tested. Importantly, we did not measure time-to-reconcile or inter-annotator agreement; we thereby treat these results as indicative for this specific corpus and prompt specification. In future applications of this technique, we would expect local experts to define extraction rules (Figure 2(b)) following individual institutions' cataloguing policy. Extending the workflow to providers beyond OpenAI (e.g., Anthropic, Google) or to locally hosted open-weight models would provide a useful test case for assessing the generalisability of these findings, as well as the viability of non-commercial alternatives.

Domain-expertise remains decisive. Crucially, the experiment demonstrated that extraction quality depends not only on conscious MLLM selection but also on expert-guided design decisions in *output schema definition, prompt specification,* and *expert annotation of the image corpus for ground truth generation* (Figure 2(b)). These findings point towards the creation of hybrid workflows in which MLLMs pre-populate candidate records for already digitised collections, while experts retain authority over verification, correction, and inclusion into authoritative digital archives.

Acknowledgments

The authors thank Tom Halvarsson and Martin Holmes of the Bodleian Libraries for their generous help and inspiration: drawing our attention to the collection of historical concert programmes, facilitating access to the materials, and seeding the initial question of whether MLLMs might be applicable to this source material.

This work was extended and continued through the Leverhulme Trust Research Project Grant *Elgar's Themes: New Pathways for Analysis, Interpretation and Engagement* (project number RPG-2024-080, PI Daniel Grimley). We remain ever grateful to our project colleagues for their advice and encouragement, especially Frankie Perry whose expertise was invaluable in understanding the needs and practicalities of cataloguing.

The experiment work reported here was made possible by OpenAI API access granted to the authors by the University of Oxford AI Competency Centre.

The first author's studies are supported by the Clarendon Fund at the University of Oxford, the Hélène La Rue Scholarship in Music at St Cross College University of Oxford, and the Studienstiftung des Deutschen Volkes; with additional support for conference attendance from St Cross College and the Academic Support Fund of the Department of Engineering Science, University of Oxford.

Finally, we thank the anonymous reviewers for their constructive comments and suggestions which helped us clarify and improve the reporting of this study for an interdisciplinary audience.

References

[1] D. M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zachi I. Attia. 2025. Fine-Tuning Large Language Models for Specialized Use Cases. *Mayo Clinic Proceedings: Digital Health* 3, 1 (2025), 100184. doi:10.1016/j.mcpdig.2024.11.005

[2] Charlotte Armstrong, Rachel Cowgill, Alan Dix, Christina Bashford, Rupert Ridgewell, Maureen Reagan, Michael Twidale, and J. Stephen Downie. 2023. Reframing Ephemera: Digitisation, Community Music-making, and Archival Value(s). In *Digital Approaches to Inclusion and Participation in Cultural Heritage*, Danilo Giglitto, Luigina Ciolfi, Eleanor Lockley, and Eirini Kaldeli (Eds.). Routledge, London, 160–180. doi:10.4324/9781003277606-9

[3] David Bainbridge, Rachel Cowgill, Frankie Perry, John Stephen Downie, Alan J. Dix, and Michael B. Twidale. 2023. Collaborative Musicology: Designing a Digital Library of Musical Events Ephemera. In *Proceedings of the 10th International Conference on Digital Libraries for Musicology (DLfM '23)*, Martha E. Thomae (Ed.). Association for Computing Machinery, New York, NY, USA, 119–127. doi:10.1145/3625135.3625147

[4] Christina Bashford. 2008. Writing (British) Concert History: The Blessing and Curse of Ephemera. *Notes, Second Series* 64, 3 (2008), 458–473. doi:10.1353/not.2008.0023

[5] Karin Beijering, Charlotte Gooskens, and Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands* 25, 1 (2008), 13–24. doi:10.1075/avt.25.05bei

[6] Bodleian Libraries. 2025. Oxford University Musical Club and Union [Programmes. 1916-]: "SOLO" Catalogue Record. https://solo.bodleian.ox.ac.uk/permalink/44OXF_INST/ogbd98/alma990143522580107026

[7] Bodleian Libraries. 2025. Oxford University Musical Club [Programmes &c., 1872-1916]: "SOLO" Catalogue Record. https://solo.bodleian.ox.ac.uk/permalink/44OXF_INST/ogbd98/alma990143523700107026

[8] Bodleian Libraries. 2025. Oxford University Musical Union [Programmes &c. 1884-1916]: "SOLO" Catalogue Record. https://solo.bodleian.ox.ac.uk/permalink/44OXF_INST/ogbd98/alma990143565930107026

[9] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13590–13618. doi:10.18653/v1/2024.findings-acl.807

[10] John Carnelley (Ed.). 2015. *George Smart and Nineteenth-Century London Concert Life*. Music in Britain, 1600-2000, Vol. 12. Boydell & Brewer, Woodbridge. doi:10.1017/9781782045922

[11] Rachel Cowgill. 1998. The London Apollonicon Recitals, 1817–32: A Case-Study in Bach, Mozart and Haydn Reception. *Journal of the Royal Musical Association* 123, 2 (1998), 190–228. doi:10.1093/JRMA/123.2.190

[12] Jasmine Darlington-Rielly. 2019. Music Ephemera within Library Collections: A Review of the Literature. *Journal of the Australian Library and Information Association* 68, 2 (2019), 194–205. doi:10.1080/24750158.2019.1609338

[13] Alan Dix, Rachel Cowgill, Christina Bashford, Simon McVeigh, and Rupert Ridgewell. 2019. Crowdsourcing and Scholarly Culture: Understanding Expertise in an Age of Popularism. In *Macrotask Crowdsourcing: Engaging the Crowds to Address Complex Problems*, Vassillis-Javed Khan, Konstantinos Papangelis, Ioanna Lykourentzou, and Panos Markopoulos (Eds.). Springer International Publishing, Cham, 189–214. doi:10.1007/978-3-030-12334-5_7

[14] Helen English. 2014. Music-making in the Colonial City: Benefit Concerts in Newcastle, NSW in the 1870s. *Musicology Australia* 36, 1 (2014), 53–73. doi:10.1080/08145857.2014.896071

[15] Anja Fischer, Julia Heimerdinger, Ralf Kwasny, and Sabrina Radatz. 2016. Archiv des Konzertlebens. https://www.simpk.de/en/forschung/themen/interpretationsforschung/archive-of-concert-life.html

[16] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility. In *Proceedings of INTERSPEECH 2016: Understanding Speech Processing in Humans and Machines*, Nelson Morgan, Georgiou Panayiotis, Shrikanth Narayanan, and Florian Metze (Eds.). Causal Productions Pty Ltd, Rundle Mall, Australia, 650–654. doi:10.21437/Interspeech.2016-431

[17] James J. Fuld. 1981. Music Programs and Posters: The Need for an Inventory. *Notes* 37, 3 (1981), 520–532. doi:10.2307/940313

[18] Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. doi:10.48550/arXiv.2504.00414

[19] Kathryn Gronsbell, Lisa Barrier, and Rob Hudson. 2020. The Evolution and Impacts of Modeling Performance History as Data in the Carnegie Hall Archives. *Performing Arts Resources* 35 (2020), 47–73, XIV–XV. https://www.proquest.com/docview/2544916505

[20] Jinghua Groppe, Andreas Marquet, Annabel Walz, and Sven Groppe. 2026. Automated Archival Descriptions with Federated Intelligence of LLMs. In *Database and Expert Systems Applications: Proceedings of the International Conference on Database and Expert Systems Applications*, Robert Wrembel, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (Eds.). Springer-Verlag, Berlin, Heidelberg, 53–67. doi:10.1007/978-3-032-02049-9_4

[21] Christopher Hibbert and Edward Hibbert. 1988. *The Encyclopaedia of Oxford*. Macmillan, London.

[22] Deborah Lee. 2011. Classifying Musical Performance: The Application of Classification Theories to Concert Programmes. *Knowledge Organisation* 38, 6 (2011), 530–540. https://www.imrpress.com/journal/ko/38/6/10.5771/0943-7444-2011-6-530

[23] Library of Congress. 1999. MARC 21 Format for Bibliographic Data. https://www.loc.gov/marc/bibliographic/

[24] Guodong Long. 2024. The Rise of Federated Intelligence: From Federated Foundation Models Toward Collective Intelligence. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 8547–8552. doi:10.24963/ijcai.2024/980

[25] Charles Edward McGuire. 2025. Hallé Archive. *Nineteenth-Century Music Review* FirstView article, n.n. (2025), 1–7. doi:10.1017/S1479409825000035

[26] S. McVeigh. 1979. *The violinist in London's concert life, 1750-1784: Felice Giardini and his contemporaries*. Ph. D. Dissertation. University of Oxford. https://ora.ox.ac.uk/objects/uuid:f9e0fac8-89ae-4f30-a528-51b59f8a8970

[27] Simon McVeigh. 1989. The Professional Concert and Rival Subscription Series in London, 1783–1793. *Royal Musical Association Research Chronicle* 22 (1989), 1–135. doi:10.1080/14723808.1989.10540933

[28] Simon McVeigh. 2010. 'As the sand on the sea shore': Women Violinists in London's Concert Life around 1900. In *Essays on the History of English Music in Honour of John Caldwell: Sources, Style, Performance, Historiography*, Emma Hornby and David Maw (Eds.). Boydell & Brewer, Woodbridge, 232–258. doi:10.1017/9781846158018.013

[29] Sander Münster, Ferdinand Maiwald, Isabella Di Lenardo, Juha Henriksson, Antoine Isaac, Manuela M. Graf, Clemens Beck, and Johan Oomen. 2024. Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe. *Heritage* 7, 2 (2024), 794–816. doi:10.3390/heritage7020038

[30] Terhi Nurmikko-Fuller, Alan Dix, David M. Weigl, and Kevin R. Page. 2016. In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera. In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology (DLfM '16)*, Ben Fields and Kevin Page (Eds.). Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/2970044.2970049

[31] Oxford English Dictionary. 2024. ephemera, n.^{2}, sense 2. doi:10.1093/OED/6647230068

[32] Jann Pasler. 1993. Concert Programs and their Narratives as Emblems of Ideology. *International Journal of Musicology* 2 (1993), 249–308. http://www.jstor.org/

stable/24617987

[33] Manon Reusens, Amy Adams, and Bart Baesens. 2025. Large Language Models to make museum archive collections more accessible. *AI & SOCIETY* 40, 6 (2025), 4485–4497. doi:10.1007/s00146-025-02227-8

[34] Rupert Ridgewell. 2010. The Concert Programmes Project: History, Progress and Future Directions. *Fontes Artis Musicae* 57, 1 (2010), 50–64. http://www.jstor.org/stable/23512083

[35] Rupert M. Ridgewell, International Association of Music Libraries., and Music Libraries Trust. 2003. *Concert programmes in the UK and Ireland : a preliminary report.* IAML UK & Irl and the Music Libraries Trust, London.

[36] Andrea Schimmenti, Valentina Pasqual, Francesca Tomasi, Fabio Vitali, and Marieke van Erp. 2024. Structuring Authenticity Assessments on Historical Documents using LLMs. In *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024: Proceedings (Quaderni di Umanistica Digitale)*, Antonio Di Silvestro and Daria Spampinato (Eds.). AIUCD, Catania, 463–468. doi:10.6092/unibo/amsacta/7927

[37] Christopher Scobie. 2016. Ephemeral Music?: – The 'Secondary Music' Collection at the British Library. *Fontes Artis Musicae* 63, 1 (2016), 21–32. doi:10.1353/fam.2016.0005

[38] Misti Shaw. 2012. The New York Philharmonic Digital Archives. *Music Reference Services Quarterly* 15, 4 (2012), 276–280. doi:10.1080/10588167.2012.728087

[39] Anne Shelley and Veronica A. Wells. 2011. Reviewed Work: Concert Programmes Database. *Notes* 68, 1 (2011), 145–147. http://www.jstor.org/stable/23012883

[40] Karl Stapleton. 2010. Prague Concert Life 1850-1881: An Annotated Database. *Fontes Artis Musicae* 57, 1 (2010), 1–22. http://www.jstor.org/stable/23512080

[41] Richard Stone. 1998. Junk mail: Printed ephemera and preservation of the everyday. *Journal of Australian Studies* 22, 58 (1998), 99–106. doi:10.1080/14443059809387406

[42] Ian Taylor. 2005. 'A Period of Orchestral Destitution'?: Symphonic Performance in London, 1795–1813. *Nineteenth-Century Music Review* 2, 1 (2005), 139–168. doi:10.1017/S1479409800001592

[43] Gabor Mihaly Toth, Richard Albrecht, and Cedric Pruski. 2025. Explainable AI, LLM, and digitized archival cultural heritage: a case study of the Grand Ducal Archive of the Medici. *AI & SOCIETY* 40, 6 (2025), 4561–4573. doi:10.1007/s00146-025-02238-5

[44] Mascha van Nieuwkerk, Harm Nijboer, and Ivan Kisjes. 2020. The Felix Meritis Concert Programs Database, 1832–1888: From Archival Ephemera to Searchable Performance Data: Arts and Media. *Research Data Journal for the Humanities and Social Sciences* 5, 2 (2020), 62–78. doi:10.1163/24523666-00502006

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[46] William Weber. 2000. Miscellany vs. Homogeneity: Concert Programmes at the Royal Academy of Music and the Royal College of Music in the 1880s. In *Music and British Culture, 1785–1914: Essays in Honour of Cyril Ehrlich*, Christina Bashford and Leanne Langley (Eds.). Oxford University Press, 299–320. doi:10.1093/oso/9780198167303.003.0014

[47] William Weber. 2008. *The Great Transformation of Musical Taste: Concert Programming from Haydn to Brahms.* Cambridge University Press, Cambridge.

[48] Wiener Philharmoniker. n.d. Concert Archive. https://www.wienerphilharmoniker.at/en/konzert-archiv

[49] Barbara Wiermann. 2018. Musiconn.performance – Musikalische Ereignisdaten im Fachinformationsdienst Musikwissenschaft. In *Kooperative Informationsinfrastrukturen als Chance und Herausforderung: Festschrift für Thomas Bürger zum 65. Geburtstag*, Achim Bonte and Juliane Rehnolt (Eds.). De Gruyter Saur, Munich, 398–415. https://www.jstor.org/stable/j.ctvbkk3p5.43

[50] Yunting Xie, Matti La Mela, and Fredrik Tell. 2025. Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT. In *Proceedings of the 9th Digital Humanities in the Nordic and Baltic Countries Conference*, Mari Väina (Ed.). 1–15. doi:10.5617/dhnbpub.12294

[51] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More Samples or More Prompts? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1772–1790. doi:10.18653/v1/2024.findings-naacl.115

[52] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), 1–20. doi:10.1093/nsr/nwae403

[53] Li Yujian and Liu Bo. 2007. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 1091–1095. doi:10.1109/TPAMI.2007.1078