MultiSumm - Multimodal Summarisation Task at MediaEval 2026

Anastasiia Potyagalova, Hyunji Cho and Gareth J. F. Jones

ADAPT Centre, School of Computing, Dublin City University, Ireland
Department of Geography, School of Natural Sciences, Trinity College Dublin, Ireland
ADAPT Centre, Maynooth University, Ireland

Abstract

We describe the MediaEval MultiSumm task. This tasks requests participants to use the online content relating to Food Sharing Initiative (FSIs) for each of a small numnber of cities accessed via a list of FSI URLs for each city. Evaluation of submissions explored the user of an "LLM-as-Judge" approach to assessment of the quality of submissions. Although this was the first edition of this task, participants submissions provided valuable initial insights into effective methodologies for the creation of multidocument summaries from diverse online content.

1. Introduction

Multidocument summarization of textual content has long been an active area of research [1], [2]. Traditionally, this has been a complex and relatively inflexible process in terms of both output style and task formulation, often requiring the integration of multiple natural language processing (NLP) tools and carefully specified summarization pipelines [3]. The emergence of large language models (LLMs) has significantly transformed many NLP tasks, including summarization [4]. More recently, multimodal LLMs have begun to exert a similar influence on tasks involving multimedia content [5].

Although the MultiSumm tasks could in principle be addressed using more traditional NLP and multimedia processing approaches, we expect that most participants will approach them using multimodal LLM-based methods.

To the best of our knowledge, this is the first benchmark task to focus specifically on this problem, offering a valuable opportunity to explore both the potential and the challenges of applying multimodal LLMs to tasks of this kind.

2. Task Definition

The MultiSumm task at MediaEval 2026 focuses on the creation of multimodal summaries from multiple heterogeneous web content sources. In this task, participants are asked to produce concise and informative summaries that combine textual and visual information derived from websites describing Food Sharing Initiatives (FSIs) across cities worldwide [6], [7]. The task is based on resources developed within the H2020 CULTIVATE project [1] , which seeks to investigate

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
anastasia.potyagalova@adaptcentre.ie (A. Potyagalova); hcho@tcd.ie (H. Cho); Gareth.Jones@mu.ie (G. J. F. Jones)

[1] [https://cultivate-project.eu/

and support urban and peri-urban food sharing through the development of large-scale open data resources [8].

A key resource supporting the task is the ShareCity200 database, an automatically crawled and curated collection of FSIs identified in 200 cities. This database extends the earlier ShareCity100 dataset [2] by covering both European and selected international cities [9], [10]. Each city-specific subset of ShareCity200 contains verified web pages, metadata, and multimedia material related to local food sharing activities.

Participants are provided with the URLs of FSIs identified within ShareCity200 for a small number of cities and are asked to access the online content associated with these URLs in order to produce a multimodal summary that captures the FSI landscape of each city. The summary is expected to represent several dimensions, including:

- Geographical distribution of initiatives across city districts
- Categories and types of initiatives (e.g., sharing, swapping, gifting)
- Operational level (government-funded, district-supported, community-led)
- Popularity or activity level, as reflected in web and social signals
- Public sentiment or community feedback
- Representative visual content (e.g., photographs illustrating major FSIs)

The summaries must be produced in English and presented in a structured multimodal format that combines textual and visual elements to improve clarity and engagement. Participants will be provided with a reference schema describing the expected summary structure, example outputs, and detailed evaluation criteria.

2.1. Research Questions

The task also aims to promote reflection on broader research issues related to multimodal summarization with LLMs, including:

- What are the key challenges in generating summaries from multi-source web content?
- How can LLMs be effectively applied in multimodal summarization?
- What open research problems remain in multidocument and multimodal summarization?
- How effective are LLM-based evaluation methods for assessing multimodal summaries?

By addressing these questions, the MultiSumm task seeks to advance understanding of LLM-driven summarization techniques [11], [12] and their potential for automatically generating rich, cross-media representations of real-world phenomena such as urban food sharing.

3. Dataset and Evaluation

The dataset provided for the MultiSumm 2026 task originates from the ShareCity200 collection developed within the H2020 CULTIVATE project. It consists of manually verified web links to online resources describing Food Sharing Initiatives (FSIs) in selected cities [9]. Each entry corresponds to an identified FSI website that forms part of the broader ShareCity200 database of global urban food sharing activity [13].

[2] [https://sharecity.ie/research/sharecity100-database/

3.1. Data Splits

The MultiSumm datasets are organized by city and divided according to the main and subtask configurations. The dataset for Cork (Ireland) dataset is provided as the development and training material. It contains verified URLs of FSIs within City city and serves as an example for link structure, file format, and summary schema.

The main evaluation focuses on Dublin (Ireland) and Brighton and Hove (U.K.), representing English-speaking cities. Participants are required to use the verified URLs to collect content and generate multimodal summaries describing the food sharing landscape in each city.

The subtask expands the evaluation to London (U.K.), Milan (Italy), and Barcelona (Spain). These datasets introduce the challenges of a much larger city and iadditional linguistic and cultural variation, enabling the assessment of summarization systems under multilingual and cross-domain conditions.

As an optional extension, participants may produce summaries that reflect the spatial distribution of Food Sharing Initiatives (FSIs) across administrative districts or boroughs, where such information is available. In this setting, FSIs can be grouped by district and described using a simple qualitative density scheme, with green indicating a high concentration, yellow a medium concentration, and red a low concentration. No map generation is required; instead, participants may capture these patterns either through textual description, such as noting that FSIs are concentrated in inner-city areas, or through a lightweight structured indication of district-level density categories. This extension is entirely optional, is primarily intended for the subtask cities of London, Barcelona, and Milan, and is not required for participation or formal evaluation. However, participants are also welcome to apply the same district-level analysis to the main task cities of Dublin and Brighton, and such submissions will be viewed positively in the qualitative assessment.

Participants are responsible for retrieving relevant textual and visual content from the listed web resources in accordance with the summary schema and ethical web use guidelines. All datasets are released under the CULTIVATE research data sharing policy, and may be expanded with additional city links on request to support multilingual or region-specific experimentation.

3.2. Evaluation Methodology

Systems are assessed using a hybrid automatic–LLM judging framework. The textual component of each submission is evaluated by large language models acting as evaluators (LLM-as-Judge), following recent methods proposed for generative IR and summarization quality assessment. The visual component (images accompanying summaries) is evaluated using a combination of automatic metadata verification and LLM-based multimodal reasoning (via GPT-4V-like models), allowing both factual and contextual assessment of relevance.

3.3. LLM-as-Judge Implementation

Evaluation is conducted using prompting templates in which the LLM acts as an expert assessor [14], [15]. A representative prompt is structured as follows:

"You are an expert evaluator reviewing a multimodal report about Food Sharing Initiatives (FSIs) in [City]. Evaluate the report using the following criteria—Informational Coverage, Accuracy and Factual Consistency, Clarity and Structure, Use of Visuals, Local Relevance, and Practical Usefulness—assigning a score from 1 to 5 for each and providing justifications for scores."

This framework enables consistent large-scale evaluation with manual scoring, while maintaining transparency and flexibility for future iterations. The methodology also facilitates later fine-tuning of smaller evaluators using the collected LLM-as-Judge data [16].

3.4. Evaluation Output and Reporting

Each evaluated submission produces a structured JSON report containing per-dimension scores and textual explanations. These can be aggregated to compute mean performance across cities and dimension-specific leaderboards. The evaluation schema also supports inclusion of human reviewer annotations for calibration and cross-validation of LLM-based judgments.

4. Discussion and Insights

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. Further, the author(s) used X-AI-IMG for figures 3 and 4 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Anonymous, Survey on multi-document summarization: Systematic literature review, arXiv preprint arXiv:2312.12915 (2023). URL: https://arxiv.org/abs/2312.12915.

[2] Supriyono, A. P. Wibawa, Suyono, F. Kurniawan, A survey of text summarization: Techniques, evaluation and challenges, Natural Language Processing Journal 7 (2024) 100070. URL: https://www.sciencedirect.com/science/article/pii/S2949719124000189. doi:https://doi.org/10.1016/j.nlp.2024.100070.

[3] Z. Sheng, K. Yang, et al., Multi-document summarization via deep learning techniques, ACM Transactions on Information Systems (2021). doi:10.1145/3529754.

[4] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, 2024. URL: https://arxiv.org/abs/2307.06435. arXiv:2307.06435.

[5] C. X. Liang, P. Tian, C. H. Yin, Y. Yua, W. An-Hou, L. Ming, T. Wang, Z. Bi, M. Liu, A comprehensive survey and guide to multimodal large language models in vision-language tasks, 2024. URL: https://arxiv.org/abs/2411.06284. arXiv:2411.06284.

[6] A. R. Davies, Urban food sharing: Rules, tools and networks, Policy Press, 2019.

[7] A. R. Davies, A. Cretella, V. Franck, Food sharing initiatives and food democracy: Practice and policy in three european cities, Politics and Governance 7 (2019) 8–20.

[8] H. Wu, H. Cho, A. R. Davies, G. J. F. Jones, Llm-based automated web retrieval and text classification of food sharing initiatives, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24, ACM, 2024, p. 4983–4990. URL: https://doi.org/10.1145/3627673.3680090. doi:10.1145/3627673.3680090.

[9] A. Davies, H. Cho, A.-M. Gatejel, R. Martinez Varderi, M. Vedoa, CULTIVATE Briefing note - Food sharing landscapes in Hub city locations, 2024. URL: https://doi.org/10.5281/zenodo.11030355. doi:10.5281/zenodo.11030355.

[10] D. Phelan, A. Davies, N. Gomboli, The european food sharing dictionary, 2023. URL: https://doi.org/10.5281/zenodo.10160274. doi:10.5281/zenodo.10160274.
[11] Y. Zhang, M. Wang, C. Ren, Q. Li, P. Tiwari, B. Wang, J. Qin, Pushing the limit of llm capacity for text classification, arXiv preprint arXiv:2402.07470 (2024).
[12] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, G. Wang, Text classification via large language models, arXiv preprint arXiv:2305.08377 (2023).
[13] A. Potyagalova, H. Cho, I. Bacher, H. Wu, P. Buffini, A. R. Davies, G. J. F. Jones, An application for development and interactive visual engagement with the SHARECITY 200 food sharing initiative (FSI) database in the CULTIVATE project, in: Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining, WSDM '26, Association for Computing Machinery, New York, NY, USA, 2026, pp. 1331–1334. URL: https://doi.org/10.1145/3773966.3779409. doi:10.1145/3773966.3779409.
[14] Y. Lu, X. Yang, X. Li, et al., Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation, arXiv preprint arXiv:2305.11116 (2023). URL: https://arxiv.org/abs/2305.11116.
[15] J. Gu, X. Jiang, Z. Shi, et al., A survey on llm-as-a-judge, arXiv preprint arXiv:2411.15594 (2024). URL: https://arxiv.org/html/2411.15594v1.
[16] H. Wei, S. He, T. Xia, F. Liu, A. Wong, J. Lin, M. Han, Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, 2025. URL: https://arxiv.org/abs/2408.13006. arXiv:2408.13006.