EBU

OPERATING EUROVISION AND EURORADIO

USAGE AND EVALUATION OF LLM IN MEDIA ORGANIZATIONS

TECHNICAL REPORT

TR 083

BBC BR DW NHK Rai

SRG SSR SWR TVP vrt yle

ZDF radiofrance

JANUARY 2025

Document History

EBU Committee	TC	
Drafting Group	Smart Media Production – Metadata & AI	
First published	2025-01-06	
Revised		

Acknowledgement

EBU technical publications are the work of experts from EBU Members, Associates, Approved Participants and third parties consisting of international standards bodies, industry partners, academic institutions and independent consultants. Their contribution to EBU technical publications is a very generous act by the individuals concerned and by their employers. The EBU appreciates their efforts and thanks them most sincerely.

To access the full collection of our Technical Reports, please see: https://tech.ebu.ch/publications

If you are interested in submitting a topic for an EBU Technical Report, please contact: tech@ebu.ch

This publication has been produced with the assistance of the following entities:

Members:

RAI, Radio France, VRT, BBC, DW, BR, SWR, TVP, ZDF, SRG/SSR RSI, NHK, YLE

EBU Co-ordinator:

Alexandre Rouxel

Scope

This Technical Report examines the use of Generative AI for media and details of EBU Member activities providing an overview of the test and projects they are working on.

Key words

Generative AI, LLM, benchmarking, content tagging, transcript chaptering, LLM fine-tuning, writing assistant, newsroom, radio programmes.

Contents

Abstract

Large Language Models (LLMs) offer transformative potential for the media, provided that their use is guided by robust and adaptive evaluation methods. Experimental activities carried out by Members demonstrate the capabilities of LLMs for tasks such as content tagging, summarization, segmentation, title generation and translation.

Although LLMs are promising, their evaluation remains complex and it's difficult to accurately evaluate the content generated such as summaries or chapters. This report explores the use and evaluation of LLM in media organizations, focusing on applications, evaluation methods and integration challenges, it also addresses the issue of scalability and suggests cost-effective strategies for certain tasks such classification or tagging.

List of Tables

List of Tools and Processes

The following terms describe the tools noted throughout this report:

Note: Each tool name is also a bookmark reference for the main body of the report. A good source of information is Microsoft Learn https://learn.microsoft.com/en-us/

Tool	Description
Adobe Firefly	A generative AI system that helps creatives generate images, create content, and improve their workflows
AIVA	Artificial Intelligence Virtual Artist. An AI-powered music composition tool that uses deep learning algorithms to create original music across various genres.
BARTScore	Text-generation evaluation metric that treats model evaluation as a text-generation task
BertScore	Metric that relies on contextualized embeddings to measure the similarity between two texts
BLEU	**BiLingual Evaluation Understudy** - is a measurement of the difference between an automatic translation and human-created reference translations of the same source sentence
BLEURT	An evaluation metric for natural language generation that takes a pair of sentences and it returns a score indicating fluency and meaning between a reference and the generated text.
ChatGPT	Generative Pre-trained Transformer that '*interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests*"
Claude 3.5	A family of LLMs developed by Anthropic
CLIP	Contrastive Language-Image Pre-training - a model developed by OpenAI that connects images and text
Colossyan	A cloud-based AI video platform that allows users to create videos using AI actor

Tool	Description
DALL-E 3	DALL·E 3 uses a combination of deep learning and neural networks to create images based on natural language prompts.
DeepEval	open-source LLM evaluation framework, for evaluating and testing large-language model systems
FALCON	An autoregressive decoder-only LLM released by the UAE's Technology Innovation Institute (TII)
Flamingo	A visual language model (VLM) developed by Google DeepMind that can solve complex problems using just a few task-specific examples
Fliki	A text-to-video platform that uses artificial intelligence (AI) to create videos from text
Gemini1.5Pro	LLM that "*utilizes Mixture-of-Experts (MoE) architecture for increased efficiency, allowing it to handle complex tasks more adeptly*"
GPT-4	Generative Pre-trained Transformer 4 is a large language model (LLM) that can generate text from textual and visual input:
GPT-4o	"o" for "omni" - a multilingual, multimodal generative pre-trained transformer developed by OpenAI., designed to process and integrate text, visuals and audio all on one neural network
ImageFX	A free AI-powered tool from Google that generates images from text descriptions using uses Imagen 2, Google DeepMind's text-to-image model
Kosmos-2	A Multimodal Large Language Model (MLLM) - designed to handle text and images simultaneously, trained on a vast dataset of grounded image-text pairs (GRIT)
LangChain	A software framework designed to help create applications that utilize large language models.
Llama3.1	**Large Language Model Meta AI** - a family of autoregressive LLMs released by Meta AI
LLaVA	Large Language and Vision Assistant - a multimodal model that combines visual and language understanding
Llamantino 7B	An Italian-adapted version of Meta's LLaMA 2 that aims to provide Italian NLP researchers with a tool to tackle tasks such as information extraction and closed qa. (Optimised for the Italian language by the University of Bari)
LM Evaluation Harness	LLM evaluation tool designed to integrate with common datasets, with customizable settings for a variety of tasks
LoRA	Low rank adaptation – a Parameter efficient fine-tuning (PEFT) method commonly used to customize large language models for new tasks.
METEOR	Metric for Evaluation of Translation with Explicit ORdering - score grader that evaluates generated text by comparing it to reference texts, focusing on precision, recall, and content alignment
Midjourney	An independent research lab - a generative artificial intelligence (AI) program that creates images from text prompts
Mistral large 2	Mistral AI's Mistral Large 2 (24.07) foundation model (FM) is now generally available in Amazon Bedrock.
MusicLM	An experimental AI tool from Google that can turn your text descriptions into music
MusicGen	A free AI music generation tool developed by Meta using a single Language Model (LM) to create music based on either text descriptions or melodies

Tool	Description
Pika Lab	**An AI video generator** that allows users to create videos using either text or image prompts
Pyannote diarisation	An open-source Python toolkit that uses speaker diarization to partition audio files into speakers
ROUGE	**Recall-Oriented Understudy for Gisting Evaluation** - evaluates the similarity between the generated text and reference text based on n-gram overlap, including ROUGE-N (unigram, bigram, etc.), and ROUGE-L (longest common subsequence)
SDV	Stable Diffusion Videos - Image-to-Video is a diffusion model that takes in a still image as a conditioning frame and generates a video from it.
Stable Audio	An open-source model optimized for generating short audio samples, sound effects, and production elements using text prompts
Stable Diffusion	A generative AI model that produces photorealistic images from text and image prompts.
Synthesia	A synthetic media generation company that develops software used to create AI generated video content
Vid2Seq	A multi-modal single-stage dense event captioning model pretrained on narrated videos which are readily-available at scale.

1. INTRODUCTION

Generative AI is a type of machine learning in which a model is trained to generate new data. The goal is to create a representation of the data that can be used as an engine to generate other data. For example, if the data is text, a language model is generated. Other models can be trained for tasks such as generating realistic images, audio and video.

2. THE USE OF GENERATIVE AI MODELS IN MEDIA

Generative AI applications have the potential to produce new and creative solutions to problems that are difficult or impossible to solve using traditional rule-based programming methods or other machine learning approaches. In the media domain, Large Generative AI Models (LGAIM) are used to create and analyse images, video clips, text, music and creative works (Table 1).

Table 1
Examples of LGAM

Data generated	Training	Example	Capabilities
Text	Text	GPT-4, Claude 3.5, Gemini 1.5, Llama3.1, FALCON, Mistral large 2	NLP, Natural Language Generation, Translation, Text classification
Images	Images with text caption	DALL-E 3, Midjourney, Stable Diffusion, Adobe Firefly, ImageFX	Image Generation, Creative Design, Logo Creation
Music	Audio waveforms of recorded music with text annotations	MusicLM, MusicGen, Stable Audio, AIVA	Generate music from text, Background music creation, Music composition
Video	Annotated video	Synthesia, Pika Lab, Colossyan,Fliki, SDV	Video Clips
Text	Annotated images	CLIP, LLaVA, Vid2Seq	Image and Video Annotation, Search and generate captions from images and videos, Visual understanding via text
Text	Text, images, videos, audio	GPT-4o, Flamingo, CLIP, Kosmos-2	Analyse and interpret multiple data types (text, images, video, audio); generate text-based outputs such as descriptions, captions, or summaries.

3. LARGE LANGUAGE MODELS

Large Language Models are a class of LGAIMs that are trained on text to generate text. In terms of training strategy, self-supervised learning has been a key factor in the success of the LGAIMS.

Self-supervised learning has several advantages over traditional supervised learning as it does not require labelled data for training. By using unlabelled data, self-supervised learning can exploit vast amounts of data without the need for costly human annotation. It also allows the model to learn from diverse and unstructured data, which can be difficult or impossible to label manually.

In practice, the most popular LLMs are trained on large amounts of raw text data using a language modelling task. In the case of an auto-regressive model, this task aims to predict the next most likely word in a sequence based on the preceding words. By training this way, the model learns to grasp the complex statistical structure of natural language including syntax and semantics which enables it to produce high-quality text.

Although this is a very active area of research, off-the-shelf products that leverage LLMs are now available to end users e.g. ChatGPT which is a popular LLM commercialised by OpenAI.

LLMs are the foundation for almost all major language technologies, but their capabilities, limitations, and risks are not well understood. As the training of the LLM is performed on a pretext task such as prediction of next words, it makes evaluation a complex problem. Performance on a particular task does not necessarily reflect the performances on another task.

3.1. Datasets and Metrics

The evaluation of LLMs is important, and clear test cases need to be identified for broadcasters. The need for benchmark test sets is emphasised, especially for tasks like tagging, title generation, translation, and classification.

One possibility is to use a layered approach to assess LLMs, similar to assessing a candidate for a company, where the model's general knowledge, values, and fit for different department’s activities are evaluated.

For summarisation for example, there’s a need for developing semantic metrics different from BLEU or ROUGE to measure the distance between summaries and reference summaries for summarisation evaluation.

Benchmarking and evaluation of LLMs in the context of broadcasting include creating benchmark test sets for different tasks such as tagging, title generation, translation, news writing, summarisation, simplification, classification, fake news detection, and reasoning. The challenge is to determine the minimum benchmark test sets needed or find metrics that do not require complex and broad test sets for evaluation.

3.2. Benchmarking

The development of various benchmarks has been instrumental in assessing the performance of LLMs.

Evaluation is a complex and rapidly evolving area of reasearch. No single comprehensive method is currently universally accepted for evaluating all models. The diversity of benchmarking techniques reflects the range of tasks where LLMs are applied. Making direct comparisons between models is difficult due to differences in scope, evaluation criteria and methodology.

These are among the current most prominent benchmarking frameworks in LLM research include:

- **Massive Multitask Language Understanding (MMLU):** A comprehensive benchmark that tests models across 57 subjects, from STEM to the humanities, using multiple-choice questions. MMLU is widely used to assess a model's general knowledge and reasoning skills across different domains [1]
- **GLUE and SuperGLUE:** These benchmarks focus on natural language understanding (NLU). GLUE evaluates models on standard NLU tasks such as sentiment analysis and text entailment, while SuperGLUE provides a more challenging extension with harder tasks and additional requirements such as common-sense reasoning and multi-sentence reasoning [2]
- **BIG-Bench Hard (BBH):** A subset of 23 particularly difficult tasks selected from the broader BIG-Bench benchmark, designed to challenge the upper limits of LLM skills. BBH assesses advanced reasoning and comprehension in complex scenarios.
- **HellaSwag:** This benchmark assesses common sense reasoning by testing the ability of models to complete narrative sequences in a way that demonstrates an understanding of typical events and human behaviour.
- **HELM (Holistic Evaluation of Language Models)**: Developed by Stanford, HELM aims to provide a comprehensive evaluation of language models by considering various factors such as accuracy, calibration, robustness, fairness, bias, toxicity and efficiency, providing a broad perspective on model performance beyond task accuracy [3]
- **The Pile:** A large, diverse dataset used for both training and evaluation of language models. The Pile contains 800 GB of English text from a variety of domains (e.g. literature, academic papers and internet forums) and is a key resource for assessing model generalisation across a wide range of data sources [4]
- **LAMBADA:** This benchmark tests a model's ability to predict the last word of a sentence, focusing on understanding long-range dependencies and contextual understanding in text.

In addition to these benchmarks, tools such as **DeepEval** and **LM Evaluation Harness** (both open-source), have emerged to facilitate the implementation of benchmark framework evaluations. These tools allow researchers and developers to customise evaluation metrics and evaluate LLMs across different dimensions, including breadth of knowledge, reasoning skills, and ethical considerations such as bias and fairness.

To overcome the problem of the increasing complexity and number of tasks to be evaluated, another approach to benchmarking is to apply a minimalist approach as proposed in the **LMentry** [5] framework. A minimalist approach involves assessing the basics of the performance before extrapolating to more complex tasks, using simple language tasks that a primary school pupil might answer.

Despite advances in LLM evaluation, the processes are very dynamic, with ongoing discussion about the most effective and comprehensive way to assess a model’s performance. As LLMs continue to evolve and expand into new areas, evaluation methods must also adapt so they reflect the complexity of language understanding, reasoning and generation.

3.3. Understanding benchmarking

Understanding benchmarking is an essential first step. NHK has shared recent research on benchmarking models for overall performance [6], specific tasks [7], and the limitations of using human feedback to assess and train LLMs [8].

3.3.1. World knowledge

The Knowledge-oriented LLM Assessment (KoLA) benchmark [6], is designed to evaluate the world knowledge capabilities of LLMs. It underscores the importance of thoughtful design for comprehensive, unbiased, and relevant evaluations of LLMs. The KoLA Benchmark considers three essential factors:

- **Ability Modelling**: evaluates LLMs using a four-level taxonomy of knowledge-related abilities: Knowledge Memorisation, Knowledge Understanding, Knowledge Application, and Knowledge Creation.
- **Data**: utilises both established data sources (such as Wikipedia) and continuously collected, evolving data to assess LLMs' ability to handle new and dynamic knowledge.
- **Evaluation Criteria**: proposes a contrastive evaluation system with standardised overall scores for better comparability across tasks and models, along with a metric to automatically evaluate knowledge hallucination.

Using the KoLA benchmark, 21 open-source and commercial LLMs were evaluated. The outcomes indicated that larger models generally excel in knowledge memorization. Alignment techniques may improve higher-level abilities but can impair memorization, and open-source models tend to perform less well than commercial models. The KoLA dataset and leaderboard have been released publicly offering a valuable resource for advancing LLMs and knowledge-related systems.

3.3.2. News summarization

A 2023 paper, "Benchmarking large language models for news summarisation" [7] presents the performance of LLMs in news summarisation tasks. It raises several important points:

1. **Instruction tuning**: is more important than model size for achieving strong summarization performance in LLMs. Instruction-tuned LLMs, especially Curie Instruct and Davinci Instruct, perform the best overall.
2. **Low-quality reference**: the poor quality of reference summaries in existing benchmark datasets can lead to misleading evaluations and hinder the performance of systems.
3. **Evaluation with freelance writers:** existing reference summaries are unreliable, but even summaries written by well-paid freelance writers may not be significantly better than LLM summaries. Therefore, defining reference summaries as ground truth may be too restrictive, as LLMs approach or even exceed average human performance.
4. **Reference-based metrics:** correlation between automated metrics like ROUGE, METEOR, BertScore, BARTScore, and BLEURT for evaluating LLM-generated summaries, and human judgments depends heavily on the quality of the reference summaries.

The paper emphasises the importance of instruction tuning and the limitations of current benchmark datasets and evaluation metrics in assessing the quality of LLM-generated summaries. It highlights the need for high-quality references and alternative evaluation approaches to further improve summarisation evaluation.

3.3.3. Content tagging

The BBC has been using a tagging system for news stories for over seven years and has clear guidelines for journalists manually tagging articles based.

The question is whether these guidelines can be fed into a generative model to automatically generate tags and whether the quality of the tags generated is satisfactory

This system suggests tags to journalists, who then confirm or adapt them. Over time, the tags generated by the system have influenced the tags created by the journalists. The system has gained enough confidence that it is now used to automatically tag news articles. Although the performance of the system is not perfect, human tagging is also inconsistent.

BBC R&D has been investigating the use of LLMs for automated tagging. LLMs can generate many more tags than journalists typically use. This was initially seen as a disadvantage but as LLMs can provide a rich list of tags, not only allowing journalists to select the most relevant but LLMs can also produce tags journalists may not have initially thought of.

There are challenges in training LLMs for tagging tasks. Journalists may implicitly filter out or include tags when manually tagging articles and training the model on the final result may not capture the full range of possible tags. It would be interesting to explore training the model on journalists' initial thoughts, including the tags they did not select.

A taxonomy could be used to guide LLMs in the tagging process. However, there may be limitations to this approach, as the provision of a comprehensive taxonomy may limit the amount of text that can be generated by the LLM due to token limitations

3.3.4. Subtitles (captions)

Subtitles used in the past cannot be used as benchmarks as they tend to reinforce previous subtitling practices. This highlights the need for new benchmarks that take into account the use of LLMs and new rules. The SRG is investigating the use of LLMs for subtitling.

NerStar tool is being considered as an evaluation tool for LLM-generated subtitles. This tool focuses on evaluating the meaning of the entire sentence rather than just word errors. Similar tools based on the capabilities and correctness of LLMs are also being explored. The main focus is on ensuring that the meaning of the subtitles is correct, but this requires the introduction of new Key Performance Indicators (KPIs) for evaluation.

Broadcasters have their guidelines and rules for subtitling, leading to variations in subtitle creation across countries and languages, some prefer verbatim subtitles, while others prioritise accuracy and timeliness. In the AI-driven world, there is a need to harmonise these rules and establish common guidelines for all EBU Members. This harmonisation can facilitate the evaluation of AI tools and the benchmarking of LLMs in the specific context of subtitling.

3.3.5. Prompts

The prompting strategy for each task is tailored to the requirements of each type of channel and platform. This is achieved by creating a way to easily take existing task components and customise them with platform-specific styles. This process is operationalised with middleware that converts these tasks into an NLP pipeline using **LangChain**.

The strategy also depends on the platforms on which the articles are published. The Smart News Assistant focuses on adapting the content to the platform on which it will be distributed, such as TikTok or Instagram. It uses LangChain and is agnostic to the types of models used, allowing it to interact with different foundation models.

3.3.6. Human feedback

The paper "Human Feedback is Not Gold Standard" [8] critically examines the limitations of using human feedback to evaluate and train LLMs. It highlights that human preference scores, which are commonly used as a metric, often fail to capture essential factors such as factuality and consistency, instead favouring superficial features such as fluency and assertiveness. The study shows that assertive, confident output is often rated higher by human annotators, even when it contains factual errors, leading to a bias in scoring. Cautious but accurate responses tend to be rated lower, illustrating the subjectivity of human judgement

3.4. Platform choice

Platform choice is a strategic decision and a concern for PSM. The choice of platform is largely pragmatic, based on existing frameworks and capabilities as an example, Azure could be considered due to its enterprise capabilities which provides guarantees for data security and compliance.

4. MEMBER ACTIVITES

The following sections provide an overview of various activities and testing being carried out.

4.1. Rai

Rai has not conducted a general, non-application-specific benchmarking of LLMs. It has focused on evaluation based on specific applications to increase its understanding of their performance and functionality.

News summarization

All applications were based on transcriptions. In general the results were excellent for summarizing news and analyzing incoming feeds, as well as for in-depth analysis of talk shows and news content to enhance understanding and generate summaries.

Various LLMs, both commercial and open-source were tried. The results suggest commercial models generally outperformed the open-source LLMs. This could be due to continuous updates and user feedback. Open-source LLMs can still offer viable performance and, in some cases, can be used where a slightly lower quality and consistency is acceptable.

Despite occasional low transcription quality, the LLMs have demonstrated resilience by accurately interpreting content. The summaries produced are clear and consistently written in correct Italian, highlighting the models' ability to handle transcription errors effectively

RAI's R&D team (CRITS) has demonstrated that for "simple" and well-defined tasks, a well-tuned open-source model can sometimes be more effective than its proprietary counterparts

Experiments and results

As part Rai's activities, there was a focus on three key tasks relevant to a media company:

- **Title suggestion**: Generating a title for a news article based on its content.
- **Semantic tagging**: Suggesting a set of semantic tags for a news item.
- **Topic change detection**: Identifying whether a topic change occurs within a given text.

Rai created specific datasets using its own data, which were then used to fine-tune an LLM and benchmark it against GPT-4 which is widely considered to be the state-of-the-art proprietary model. The title and tagging datasets were compiled from Rai's own published news articles, while the topic change detection dataset was manually labelled specifically for this task.

The model chosen for fine-tuning was **Llamantino 7B** The fine-tuning process for each task used the **LoRA** technique, which allows efficient fine-tuning of large models with fewer computational resources. RAI then benchmarked the fine-tuned models against GPT-4 (appropriately prompted) using task-specific test sets. The evaluation metrics were defined as follows:

- **Topic change detection**: Standard classification metrics, including precision, recall, accuracy and F1 score.
- **Title generation**: BLEU and ROUGE scores, and cosine similarity between vector representations of the actual and generated titles.
- **Tag generation**: Accuracy, precision and recall based on the overlap between real and generated tags, and a modified Levenshtein distance to measure word set differences.

The results were very encouraging. In all three tasks, the fine-tuned models produced better results, i.e. responses that were closer to the ground truth than those generated by GPT-4.

4.2. BBC

The BBC is taking a fine-tuning approach, experimenting with GPT-J for various tasks such as article generation, tagging, title generation, translation, summarisation, simplification and classification. The results of using LLMs for these tasks are described as good, although not as good as existing systems.

- The evaluation of LLMs for tagging and titling articles at the BBC has shown promising, though not exceptional, results. Fine-tuning LLMs on a large dataset of previously tagged articles has been effective in improving performance. In addition, the direct use of LLMs allows a large number of tags to be suggested for journalists to consider
- The BBC focused on evaluating specific applications of LLMs using them for automatic podcast segmentation. The models were effective in summarizing and labeling transcript segments.
- The BBC has also applied LLMs to classification tasks, refining them with its existing news articles and tags. This approach has yielded positive results, though slightly below those of current production workflows. With relatively little effort and access to a large dataset, fine-tuned LLMs have delivered promising outcomes.

It is important to note that in addition to classification and summarization, LLMs have potential in other areas, particularly in archive searches and content discovery. In these cases, even imperfect summaries can give users a general sense of content, aiding navigation. Additionally, LLM-driven segmentation can enhance navigation within large archives, helping users find specific pieces of content more efficiently.

4.3. VRT

LLMs are being integrated into VRT's newsroom workflows through a proof of concept (POC) that is currently under evaluation. (See Annex 1)

The models are used to automatically summarize text, extracting key details such as "*who, what and where*" from articles. This automation is use not only to tag content, but also to support the creation of new content based on the core elements of a news story.

These LLMs are incorporated into operations with prompts and are augmented with news articles. Currently there is no integration of additional context from internal databases, though this may be useful in the future.

Generative AI also enables new interactive ways to explore content, allowing journalists to engage in a more conversational approach to news stories, enhancing understanding and increasing engagement.

Generative AI tools have also been used to automate the generation of headlines, descriptions and bullet points, reducing the workload for journalists. These tools are currently in an evaluation phase but are already accessible by all VRT involved in journalism.

Recognizing the potential privacy risks of sending draft articles to external providers for summarization, VRT is developing a POC bot for in-house news, allowing active user engagement and a better understanding of audience needs.

While they are exploring and fine-tuning the open source LLMs for specific tasks, the costs and benefits trail do not currently justify the development investment. VRT sees the fine-tuning of open source models as a positive direction, but notes that at this stage of workflow evaluation, the costs and risks involved for individual organizations remain high.

4.4. NHK

NHK reported on its experience using Llama-2-7B for English-to-Japanese and Japanese-to-English translation. Using just 10,000 English-to-Japanese sentences was sufficient to improve the model's translation quality (BLEU score) beyond existing models. This improvement was achieved with the 7-billion-parameter model and required only an NVIDIA A100 GPU running for two days. The findings suggest that larger models, such as ChatGPT, may not be necessary for specific tasks like this. The trend, therefore, points toward smaller, fine-tuned models, which offer advantages in computational efficiency, cost-effectiveness, and potentially better task-specific performance

4.5. SRG

The SRG is investigating the use of LLMs for subtitling. Subtitling relies on visual segmentation rather than sentence-based segmentation, which is a challenge for LLMs in this context. In addition, human-generated subtitles are often not perfectly aligned with the spoken content, resulting in a mismatch between the spoken timecode and the written subtitle. This discrepancy makes it difficult to evaluate subtitles generated by LLMs.

The SRG is also exploring the use of LLM to generate easy-to-read and simplified language for subtitling. Both approaches are being tested. The Whisper model is being used as a comparison.

Initial results of using LLMs for subtitling are often positive, leading to a sense of satisfaction and enthusiasm. However, when accessibility experts evaluate LLM-generated subtitles they often provide critical feedback. Significant effort is required to correct errors to make the subtitles usable in production and compliant with their guidelines.

4.6. Radio France

Radio France explored the capabilities of LLMs across several projects to enhance radio content:

- **Podcast chapterization**: Segmentation of podcast transcriptions into topic-based chapters.
- **Speaker identification**: Determining the identity of speakers in audio content.
- **Highlight detection**: Identify key moments in podcasts.

For chapterisation, Radio France first transcribes audio files to text using the Whisper-Large-V3, and then applies one of three custom algorithms:

- BERT-based segmentation: This approach uses BERT, a small language model, to identify semantic breaks within the transcription, marking each break as the start of a new chapter.
- Journalist script alignment: By aligning the script text written by journalists with the transcription of the corresponding news episode, this method provides chapters and their timecodes based on the journalists' own segmentation.
- LLM Prompt Engineering: Using GPT-4 or **Gemini1.5Pro**, a custom prompt is designed to describe the chapterisation task and integrate the transcription.

Although the LLM-based method achieves the best performance, its high cost limits its use on Radio France's extensive catalogue (over 300,000 new episodes per year). The script alignment method performs well but is only applicable to news programmes.

For speaker identification, Radio France uses the **Pyannote diarisation** tool, which segments speakers within the transcription but labels them generically (e.g. SPEAKER_00). Identifying specific speakers traditionally requires voiceprints and dedicated audio analysis tools, a setup that poses legal and privacy challenges. However, LLMs offer an alternative by identifying speakers based on transcription alone via a dedicated prompt in models such as Gemini1.5Pro.

Highlight detection focuses on identifying impactful moments in podcasts - 30 to 60 second segments that capture emotional, inspiring or powerful content. Radio France has developed a prompt for this task and plans to apply it to a large number of podcast episodes. The identified highlights will serve as a training dataset to fine-tune a smaller language model capable of performing this task at a lower cost.

This approach exemplifies a process of 'distilling' a large language model into a smaller one, with the aim of achieving similar performance at a lower cost of execution

5. CONCLUSION

Generative AI and LLMs offer transformative potential for media organizations, enabling innovative solutions for content management and production across tasks such as classification, annotation, segmentation, summarization, and content generation.

LLMs have been explored and evaluated in media organisations for various applications, including classification, summarisation and search. Both open-source and commercial models, along with tools developed by PSM, have shown promising results, boosting productivity, enhancing content search and reuse, and addressing information overload. Journalists particularly value these models for their ability to streamline workflows, freeing up time for more creative work.

However, the evaluation process is still ongoing and no formal benchmarking has been carried out. The focus has been on exploring the actual applications where LLMs can be used effectively.

The BBC and Rai have found LLMs to be powerful tools for a variety of media applications, particularly for news summarization and classification tasks. Although commercial LLMs generally outperform open-source LLMs, the latter can still be used effectively. Both organisations recognise the need for a more general benchmarking of LLMs, not just focused on specific applications.

- Extractive methods of summarisation, which involve selecting representative sentences rather than generating new text, have also been explored and have shown promising results.
- For classification, LLMs have shown potential but may not yet be on par with state-of-the-art solutions. It should be noted that GPT-4 is more effective than previous iterations, so progress is inevitable.

While LLMs may appear impressive to informal human evaluators, there is still a need to assess their accuracy and reliability before using them in production. Evaluating performance is challenging and has been largely informal without ongoing R&D efforts aimed at understanding capabilities and limitations, especially when comparing LLM-generated output with human-generated abstracts or summaries.

Formal benchmarking needs to be carried out. Evaluation is likely to involve comparing summaries produced by LLMs with reference summaries using metrics such as BLEU and ROUGE and more advanced semantic metrics.

As these models advance, so must evaluation methods. The LLM evaluation landscape is dynamic, with ongoing discussions about the most effective ways to assess language understanding, reasoning, and language generation. Refining LLM capabilities for media applications involves strategies like fine-tuning and prompt engineering, tailored to organizational needs using proprietary data. High-quality, context-sensitive evaluation metrics will be essential as LLMs become more deeply integrated into workflows.

Transitioning from experimentation to production also raises cost considerations for scaling. A promising approach is to use outputs from complex, high-cost LLMs to train simpler, cost-effective models, enabling efficient, large-scale deployment without compromising performance.

In summary, while LLMs have shown promise in media organisations for tasks such as classification and summarisation, their evaluation and benchmarking are ongoing. To date, the focus has been on exploring their usefulness and applicability in real-world scenarios. It is clear from these experiments and trials that continued collaboration, data sharing and support will help offset these challenges by pooling computing resources and coordinating development efforts.

6. REFERENCES

[1] Hendrycks, D., Burns, C., Basart, S. et al. (2020). Measuring Massive Multitask Language Understanding. arXiv. https://doi.org/https://arxiv.org/abs/2009.03300v3

[2] Wang, A., Pruksachatkun, Y., Nangia, et al. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv. https://doi.org/https://arxiv.org/abs/1905.00537v3

[3] Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic Evaluation of Language Models. https://arxiv.org/pdf/2211.09110

[4] Gao, L., Biderman, S., Black, S., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. https://arxiv.org/pdf/2101.00027

[5] Efrat, A., Honovich, O., Levy, O. et al. (2022). LMentry: A Language Model Benchmark of Elementary Language Tasks. https://arxiv.org/pdf/2211.02069

[6] Yu, Jifan, Xiaozhi Wang, Shangqing Tu, et al. (2024) "KoLA: Carefully Benchmarking World Knowledge of Large Language Models." https://arxiv.org/pdf/2306.09296

[7] Zhang, Tianyi, Faisal Ladhak, Esin Durmus, et al. (2023). "Benchmarking large language models for news summarisation." https://arxiv.org/pdf/2301.13848

[8] Hosking, Tom, Phil Blunsom, and Max Bartolo. (2024) "Human feedback is not gold standard." https://arxiv.org/pdf/2309.16349.

ANNEX 1
VRT – Smart News Assistant

This annex describes a VRT's use-case of a Smart News Assistant which is used to illustrate some of the concepts outlined in this Technical Report.

***Note: It is intended to update this annex as more information becomes available*.**

What is Smart News Assistant?

Smart News Assistant (SNA) is a tool designed to help journalists create news content more efficiently. Using generative AI, the SAM repackages existing news articles for different audiences, platforms and channels.

SNA utilises in-context learning, drawing on the article text and other attributes, such as the title or existing summary, to generate new content. In-context learning allows LLMs to adapt their responses based on the context provided within a single input prompt, without requiring additional fine-tuning or parameter updates. This capability enables the model to 'learn' from examples in the prompt, allowing it to produce responses that align with the specific task or content structure

Key Tasks for SNA

While SNA relies on guidance prompts, the interface is structured so that users do not need to interact directly with these prompts. Key tasks supported by SNA include:

- **Summary creation**: Condensing the main points of the original into a shorter format.
- **Title generation**: Create a catchy and relevant title for the repackaged content.
- **Bullet points**: Breaking down the content into easily digestible bullet points

Benchmarking strategy

The tasks performed by SNA are scored by experts who provide feedback to improve its performance. To further refine the scoring process, the team is also developing a user feedback mechanism.

For data collection, VRT is using Aimstack, a lightweight MLOps platform, to gather insights during prompt generation. This data will help identify areas for continuous improvement.

Data confidentiality

The issue of data confidentiality is a key concern, especially when it comes to unpublished texts and sensitive information. Currently VRT do not recommend inputting confidential information into LLMs. As the project progresses however, it may become necessary to handle sensitive data, which would require additional security measures.

While services such as OpenAI provide enterprise-level security, managing local models is also an option. While this approach can provide greater control, it often comes with higher costs and operational challenges. In addition, new solutions are being developed to strip sensitive information from text before it is sent to APIs, further enhancing privacy.

Future work on SNA

Although the data layer is not the primary focus of the project, VRT sees it as a key area for future development. Managing the massive flow of both external and internal information is essential. The goal is synthesising it in a way that enables journalists to create new stories efficiently.

VRT has not yet established KPIs to measure the impact of the technology, although there is significant demand for the tools and users have reported improvements in efficiency and productivity. Going forward, VRT intends to implement observability and monitoring systems to track the performance and impact of LLMs.

The tool is currently being used for Dutch-language content but has the potential to be extended to other languages. The tool, which is still under development, will eventually include additional monitoring and data collection features, as well as a user feedback mechanism to further improve its functionality.

Conclusion on the usage of SNA

The use and evaluation of LLMs in media organisations is still evolving. LLMs are seen as valuable tools to increase productivity, re-use existing content and manage information overload. Journalists are generally positive about LLMs, recognising their potential to improve efficiency and free up time for more creative work