Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

HYTAS: An AI-Powered Hybrid Extractive-Abstractive YouTube Transcript Summarizer

Muhammad Bilal
Department of Computer Science & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Jamshaid Ali
Department of Computer Science & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Kashif Ali
Department of Computer Science & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Shahzad Nazir*
Department of Computer Science & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan Email: shahzad.nazir@bgnu.edu.pk

Muhammad Usman Younus
Department of Computer Science & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

ABSTRACT

Online lectures, tutorials, interviews and technical presentations are frequently available online, and such information can be more than people can process. The transcripts for YouTube videos are difficult to summarize due to their length, repetitive content, weak punctuation and potential for automatic speech recognition inaccuracies. The paper introduces HYTAS, a hybrid extractive-abstractive framework that starts by either retrieving captions or applying speech recognition to extract sentences, reconstructs coherent sentences, identifies topic boundaries, identifies the most relevant evidence using lexical and semantic features, filters out extraneous sentences via maximal marginal relevance, and generates structured summaries using a transformer model. A custom dataset of 180 videos was prepared from YouTube and split into 120 training, 20 validation, and 40 test videos. A dataset of 180 English-language YouTube videos was prepared and divided into 120 training videos, 20 validation videos and 40 test videos. HYTAS achieved ROUGE-1, ROUGE-2, and ROUGE-L F1 scores of 46.72, 23.88, and 43.15, respectively. It also achieved a BERTScore F1 of 0.906, a QAFactEval score of 0.887, and a topic coverage of 91.8%. The average compression ratio of the generated summaries was 8.6% while the redundancy rate was very low at 4.6%. The findings suggest that the use of source-grounded extraction and controlled abstractive summarization can yield

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

concise, semantically coherent and factually consistent summaries that retain connections to the source video content.

Keywords: HYTAS; YouTube Transcript Summarization; Hybrid Extractive-Abstractive Summarization; BART; Topic Segmentation; Automatic Speech Recognition; Rouge; Factual Consistency.

Introduction

Nowadays, online video has become a significant tool for knowledge sharing, software training, professional communication and education. Recordings of technical lectures, tutorials, interviews and explanatory videos are often provided with good definitions, examples, demonstrations and conclusions, but it takes a long time to review lengthy recordings. The task of automatic summarization seeks to overcome this by generating a more concise summary that captures the most salient information. Extractive summarization picks out salient units within the source text, while abstractive summarization produces a summary that is a shorter representation that is synthesized through the use of learned language models [1,2].

The amount of variation in video transcripts from edited written documents is significant. Fillers, repetitions, false starts, incomplete sentences, discourse markers and frequent topic shifts are all common characteristics of spoken language. Captions may also miss punctuation and/or be misspelled or incorrectly transcribe technical terms, names, numbers, code related vocabulary. Furthermore, using standard transformer models may not support long video inputs. The findings from the research on VT-SSum indicate that training on a written news database would lead to domain mismatch when applied to spoken transcripts, while training on the database of transcripts would give better results [12]. The long length of transcripts and informal, unplanned language are also cited as key problems for summarizing livestreams [13] to be addressed by StreamHover.

New data from video transcripts have driven this research. VT-SSum consists of 125,004 pairs of transcript and summary for 9,616 academic videos [12] and StreamHover consists of over 500 hours of manually annotated live stream content with extractive/abstractive summaries [13]. UrduMASD has 15,374 Urdu multimodal collections [14] and VISTA has 18,599 scientific presentation videos with their respective paper abstracts [15]. The resources show that it is not enough to simply adapt existing document summarization methods for use with spoken documents. It includes the accurate transcription, input processing with high length, topic-sensitive segmentation, temporal alignment, redundancy control and factual consistency.

In this paper, a Hybrid YouTube Transcript Abstractive-Extractive Summarizer (HYTAS) to convert a provided link of a YouTube video into a structured text-based summary is presented. It creates a short summary, highlights, sections by topics and (optional) timestamps in relation to the original video. HYTAS integrates pre-processing for transcripts, automatic speech-recognition fallback, topic segmentation, evidence selection, transformer-based generation and source-grounded quality control.

The study used a custom dataset of 180 publicly available English-language YouTube videos, which included instructional and technical tutorials, interview and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

documentary/explanatory content. The dataset was split into 120 training videos, 20 validation videos and 40 test videos. Extractive and abstractive Reference Summaries were created for the validation and test sets, respectively, to facilitate the assessment of content selection, semantic similarity, factual consistency, topic coverage, compression and redundancy.

This paper makes the following main contributions:

An end-to-end modular architecture which puts greatest emphasis on available captions, with the added ability to provide fallback for automatic speech recognition, and keeps timestamps preserved across the summarization pipeline.

A disfluency, recognition error, topic change and long input sequence pre-processing and hierarchical segmentation approach to a transcript specific problem.

A hybrid approach to extractive-abstractive summarization using lexical, semantic, position and redundancy-aware evidence selection with transformer-based summarization.

A customized annotated dataset of 180 English-language YouTube videos with extractive and abstractive reference summaries and transcript data.

Empirical assessment based on ROUGE, BERTScore, QAFactEval, topic coverage, compression ratio and redundancy rate.

The rest of the paper will be organized as follows. In Section 2, extractive, abstractive, long-document and transcript-focused summarization approaches are discussed. The HYTAS methodology, system architecture, pre-processing stages, evidence-selection process and evaluation measures are described in Section 3. Experimental results and comparative analysis is given in section 4. The paper ends with a Summary of the main findings and contributions of HYTAS (Section 5).

Background and Related Work

Extractive text summarization

Extractive summarization is the process of identifying and selecting sentences or utterances from the source without rewording them. The approaches used in early methods include term frequency, sentence position, similarity of title, cue phrases, and document structure. Graph-based methods represent sentences as nodes, and similarity as links. TextRank is a supervised learning-free recursive graph ranking method for detecting central sentences [3] and LexRank is another variant of Eigenvector centrality on a cosine-similarity graph to calculate sentence salience [4]. These approaches are not only easy to compute, but also easy to explain and do not add any unsupported facts to the output, as the output is based on the source text.

But when summarizing directly from speech transcripts, it can result in awkward summaries. Selected utterances can include pronouns with no antecedents, fragments of conversation, repetitions, and referring to visuals. Nazir et al. applied supervised and unsupervised feature analysis for Urdu extractive summarization and concluded that the top-ranked linguistic features in addition to basic TF-IDF model improved the performance of the basic TF-IDF model [1]. The general meaning is that the salience cannot be captured by a single statistical indicator, but requires positional, linguistic and contextual signals in lexical evidence.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Abstractive and transformer-based summarization

Abstractive summarizing produces new words that summarize and reword the source material. Transformer models are prevailing due to their ability to model long-range dependencies and produce coherent multi-sentence text, which has been achieved in many different ways through the encoder-decoder model. BART is a denoising autoencoder pre-trained for tasks such as generation, especially those with a high cost of failure [6]. Tasks involving natural-language processing have been unified into a text-to-text format with T5 [7]. PEGASUS adds a pre-training objective for summarization that masks and reconstructs important sentences as target text [8]. ProphetNet generates n-grams instead of just the next token to enhance sequence planning [9].

In general, abstractive models generate more fluent summaries than simply extractive models, but they can also miss details or misrepresent numeric information, or make statements that are not supported. This risk is particularly applicable if the input transcript has recognition errors. Thus, automatic overlap measures must be supplemented with semantic and factuality measures. Some metrics measure n-gram overlap (ROUGE) [5], others compare contextual token embeddings (BERTScore) [19] and the question-answering based metrics (e.g., QuestEval) do not require an exact reference summary for measuring relevance and consistency [20].

Long-document and speech-aware processing

The quadratic self-attention complexity and fixed number of tokens for standard transformers limits their applicability for lengthy lectures. Longformer adds local and global attention, with a linear complexity in terms of the document length, and provides a Long Document Generation (LDG) variant for the encoder-decoder architecture [10]. Another approach to summarization is hierarchical summarization: coherent overlapping chunks are extracted from the transcript and each chunk is summarized, then a second stage model creates the integrated summary. Gonzalez et al. applied chunk level processing system for lecture transcripts and show that keeping previous information context helps mitigate repetition and sometimes correct recognition errors [17].

If there are no captions or if they are not reliable, speech recognition becomes a part of the summarization problem. Whisper was trained using 680,000 hours of multilingual and multitask supervision, and showed strong zero-shot transcription performance across domains [11]. WhisperX improves long-form transcription by utilizing voice-activity detection, forced alignment and word-level timestamps, which handle drift, repetition, and inaccurate timestamps [18]. The quality of transcription is, therefore, a detail that should be regarded as a preprocess that affects the quality of a summary, but is not a quality of the summary itself.

Transcript and video summarization datasets

Table 1 is a summary of some representative datasets which directly inform the design of a YouTube transcript summarizer. The data sets come in various language, domain, supervisedness and type of summary, and have varying modalities (visual or auditory). Their diversity indicates that there is no optimum system architecture, which depends on the target domain and available modalities.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Table 1. Representative datasets for transcript and video-to-text summarization

Dataset	**Domain / language**	**Scale**	**Summary supervision**	**Key relevance**
VT-SSum [12]	Academic videos; English	9,616 videos; 125,004 pairs	Weak extractive supervision from slides	Demonstrates written-to-spoken domain mismatch
StreamHover [13]	Creative livestreams; English	>500 hours; 5,421 annotated clips	Human extractive and abstractive summaries	Addresses informal, lengthy, live transcripts
UrduMASD [14]	Social video; Urdu	15,374 multimodal collections	Abstractive summaries with video, audio, title, transcript	Supports low-resource and multimodal research
VISTA [15]	Scientific presentations; English	18,599 video-abstract pairs	Paper abstracts as summary targets	Large-scale scientific video-to-text benchmark
TR-EduVSum [16]	Educational videos; Turkish	82 videos; 3,281 human summaries	Consensus-based gold summaries	Highlights multi-annotator agreement and educational use

This is relevant to VT-SSum, as this explicitly compares models learned from written and spoken data. In spite of the smaller training sample set, the authors report an improvement of about eight F1 points for top three sentence selection when training on VT-SSum compared to CNN/DailyMail [12]. VISTA offers a complementary finding, as it shows that the transcript-only input model is poorer than the video-only and video-plus-audio versions under full fine-tuning [22], suggesting that when not properly normalised or aligned, long and noisy transcripts can drown out model attention [15]. This observation inspires the next proposed hybrid architecture of the transcript which is described in the following section.

Proposed Methodology

Research design and HYTAS architecture

The design-science approach for HYTAS is a process that starts by selecting components from existing natural-language-processing tasks and builds a summarization system that is tailored for the specific design and practical application. The model can be given a public URL for a YouTube video, it can retrieve or generate a time-aligned transcript, it can process the noisy speech to obtain well-

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

formed sentence/topic units, it can select source grounded evidence, it can produce an abstractive summary, and it returns structured outputs via a web application. The hybrid design is a combination of both extraction and abstraction, the extraction part checking evidence, length, and traceability of the text, and the abstraction part enhancing its linguistic coherence and compression.

The overall HYTAS architecture is shown in Figure 1. Captions are prioritized since there is no need for speech recognition, and can contain creator-corrected terminology. If existing captions exists, Whisper or another ASR model segments the transcript and timestamps it [11]. Both paths lead to the same preprocessing and segmentation of topics, extractive evidence selection, transformer generation and quality control.

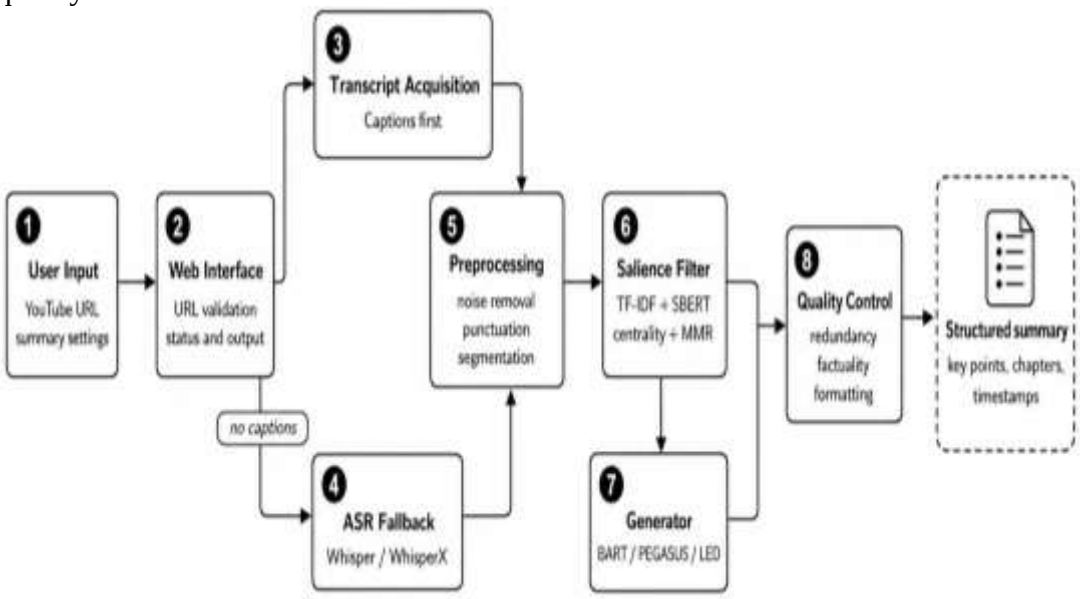

Figure 1. Overall architecture of HYTAS.

Dataset and experimental setup

To test its abilities, the software was given a custom dataset consisting of 180 publicly available English-language YouTube videos, divided into four categories: educational lectures (60 videos), technical tutorials (45 videos), interviews (40 videos), and documentary/explanatory videos (35 videos). There were about 111.4 hours of content and 884,000 transcripts in the data set. The captions were provided for 132 videos and for the remaining 48 videos, the captions were generated using Whisper-based ASR.

The whole dataset was split into 120 training, 20 validation and 40 test videos. A split was done at the video level to make sure there was no video with more than one split that contained videos from the same channel or near duplicate videos.

Both extractive and abstractive reference summaries were prepared for the validation and test videos. The two annotators were asked to write summaries independently, based on the following criteria: topic coverage, factual accuracy, conciseness and chronological organization, and the third annotator was used to resolve conflict. The same test data, summary length and evaluation scripts were employed for the HYTAS model and all of the baseline models.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Transcript acquisition and preprocessing

The first part of the acquisition is extracting the YouTube video ID, checking for the presence of captions, and choosing human-generated captions over machine-generated options. Every selected sentence is associated with its start time, duration, language and caption text, thereby being able to re-connect every caption segment to the original video. If there are no captions or they are inappropriate, audio is extracted and transcribed with Whisper, and WhisperX-style alignment can be applied to longer recordings to get more accurate timestamps [11,18]. Confidence scores are stored where available to detect or diminish uncertain names, numbers and technical terms.

Then the raw caption fragments are normalized to sentence units with the workflow shown in Figure 2. This process eliminates repeat non-speech tags, markup, filters immediate fillers and repetitions, restores punctuation, normalizes Unicode/whitespace, reconstructs sentence boundaries, and preserves original timestamp range per sentence. Table 2 shows the purpose of each preprocessing operation [23].

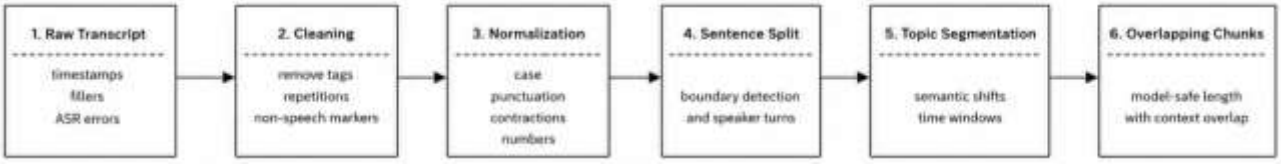

Figure 2. HYTAS transcript acquisition, normalization, and hierarchical chunking workflow.

Table 2. Transcript preprocessing operations used by HYTAS

Operation	**Purpose**
Markup removal	Remove non-semantic noise that distorts term statistics and embeddings.
Disfluency control	Improve readability without changing the speaker's intended content.
Punctuation restoration	Support sentence segmentation and generation quality.
Text normalization	Create consistent tokens for lexical and semantic processing.
Sentence boundary detection	Prevent fragmentary extractive summaries.
Topic segmentation	Create coherent units and reduce cross-topic mixing.
Overlapping chunking	Preserve dependencies near boundaries and reduce omitted transitions.

Topic segmentation and feature representation

After preprocessing, the transcript is segmented into topic segments that are coherent semantically instead of into fixed-length segments. Adjacent sentence windows are

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

used to compute sentence-BERT embeddings [21] and a topic boundary is identified whenever the semantic similarity falls below a set threshold. Supporting boundary evidence can be given by timing gaps, discourse markers, the structure of the captions, and maximum model length. Adjacent chunks have a controlled overlap, meaning that definitions, explanations, and transitional statements do not overlap at segment boundaries[24].

Each sentence is depicted with complementary features: normalized TF-IDF relevance, semantic centrality, position in the sentence, coverage of the title and keywords, coverage of the topic, sentence length, location of the timestamp, and novelty in relation to the selected evidence. This is for spoken transcripts and will not be suitable for simple lead bias, as such information may be present throughout the recording.

Extractive evidence selection

For each sentence s_i, HYTAS computes a weighted salience score that integrates lexical relevance L_i, semantic centrality C_i, positional importance P_i, keyword and topic coverage K_i, and novelty N_i. The combined score is defined as follows:

$$Score(s_i) = \alpha L_i + \beta C_i + \gamma P_i + \delta K_i + \varepsilon N_i, \quad where \quad \alpha + \beta + \gamma + \delta + \varepsilon = 1.$$

TF-IDF focuses on the words that are common in a segment but rare in the entire text. Semantic centrality selects sentences that cover the topic that's local to the text, positional weighting adds a small amount of structural information, and title/keyword coverage penalizes for concepts outside the domain of the piece's title or keywords. The more novel, the less evidence. Maximal marginal relevance sorts after ranking by a balance between salience and similarity to existing set S:

$$MMR(s_i) = \lambda \cdot Score(s_i) - (1 - \lambda) \cdot max_{sj} \in Ssim(s_i, s_j)$$

The final summary is not produced at this stage, but rather a set of small pieces of evidence with a time stamp. The distinction helps keep length of the transformer's input short, keeps repeated content to a minimum, promotes auditability, and restricts the generator to content relevant to the source [26].

Abstractive summary generation and aggregation

The evidence selected passes to a transformer generator. The default model is BART since the denoising objective and BART implementation is appropriate for rewriting imperfect transcript evidence [6]. As the main alternative, PEGASUS sets a gap-sentence pre-training objective, which is directly related to summarization [8]. If there is a coherent area left outside the standard encoder-decoder input limit [10] then it is possible to use LED. The whole extractive-abstractive interaction is shown in Figure 3.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

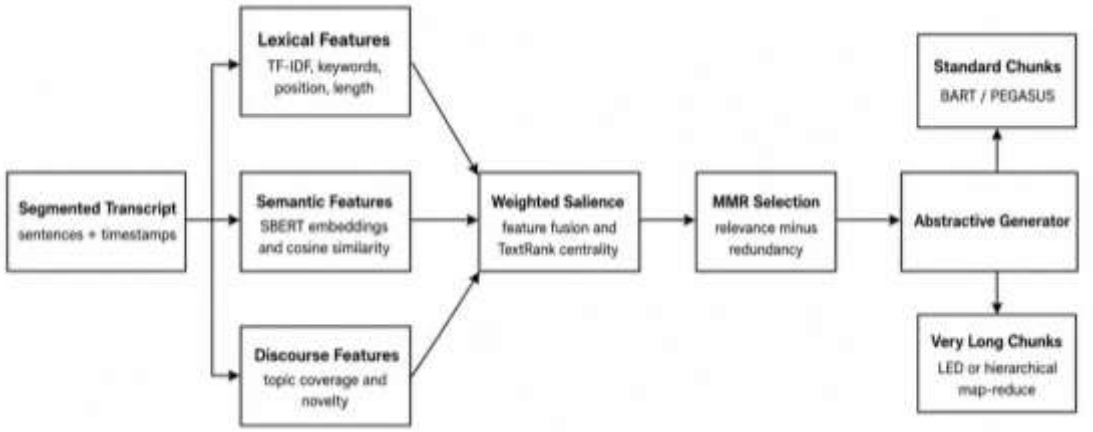

Figure 3. HYTAS hybrid extractive-abstractive model

Chunk summaries are sorted by their earliest source timestamp, combined with a global summary and scanned for near-duplicate statements. Named entities, acronyms, numbers, and claims that have been generated but are not supported by the source evidence are compared with selected source evidence. The formatter creates a summary, bullet point notes for the main ideas, topic-wise notes and time-stamped chapters. The length of the summary is regulated by a target compression ratio or the number of tokens in the summary.

Implementation and experimental configuration

The application checks the URL and generates an asynchronous processing job. The following stages of the system: capturing captions, preprocessing, evidence selection, evidence generation, and evidence quality control are recorded as separate stages, enabling the identification of failures and reporting of stage-specific latency.

Whisper medium was used as the fallback ASR, all-mpnet-base-v2 as the Semantic Representations and BART-large-cnn as the Abstractive Generator. The length of topic chunks was constrained to 900 tokens and the overlap to 100 tokens. The weighted salience function used α=0.25, β=0.30, γ=0.10, δ=0.20, and ε=0.15; MMR used λ=0.70 and an extraction ratio of 0.28. BART was trained for 4 epochs with AdamW optimizer, a learning rate of 3×10^{-5}, an effective batch size of 16, a max input length of 1,024 tokens, a max output length of 384 tokens, a beam size of 4, and a random seed of 42. Experiments were performed with an NVIDIA RTX 3060 12 GB GPU and 32 GB of RAM, utilizing Python, PyTorch, and Transformers library. These implementation/training environments are known to aid in experimental reproducibility.

Evaluation metrics

Three types of automatic, human-evaluation, and operational-performance measures were used to evaluate HYTAS. The same test inputs and summary-length budget were used for all systems. Lexical overlap (ROUGE-1, -2, -L); Contextual similarity (BERTScore); Source-supported factual consistency (QuestEval or QAFactEval); Output compactness (compression/redundancy); ASR quality (word error rate, with verified transcripts provided); Relevance, coherence, fluency, faithfulness, readability, usefulness (human raters); Operational performance (latency, throughput, memory,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

failure rate) [25].

Along with the automatic metrics, confidence intervals and paired significance tests were given. Human summaries and ratings were generated independently from one another and inter-rater agreement was determined prior to averaging ratings. The controls are employed to ensure that a fluent, but unsupported, summary is not considered better than a different one due to lexical similarity.

The 95% confidence intervals for the experiment were estimated using 1000 paired bootstrapping samples across test videos. The statistical significance was evaluated using paired bootstrap hypothesis testing for every pairwise comparison of HYTAS with each baseline on ROUGE-L at a level of $p<0.05$, corrected for multiple comparisons with the Holm-Bonferroni correction procedure. Along with the p-values, absolute score differences were also reported so that a change that was statistically significant would not necessarily be considered practically important.

The outputs were anonymized and randomized and evaluated using a 5-point ordinal rubric and 3 independent evaluators for human evaluation. The alpha was first calculated and then mean ratings were interpreted. The most important decision criterion was that the whole model should enhance the content coverage and semantic similarity while keeping the factual consistency unchanged, as far as possible. The desired level of efficiency was held to secondary importance as the multi-stage pipeline was designed to provide for greater traceability, redundancy control and source grounded generation.

Results and Discussion

The results are presented in the order in which the workflow is performed in HYTAS. The results of the transcript acquisition and pre-processing are given first in section 4.1. The full scores for the proposed models are then reported in Section 4.2. The component contributions are analyzed in Section 4.3, qualitative and human evaluation are reported in Section 4.4, and errors and operational performance are analyzed in Section 4.5. It is only after all the HYTAS-specific findings are provided in this section that the complete model is compared to established baselines in Section 4.6.

Dataset and preprocessing results

The successful processing of all 180 videos after automatic retries, 176 of which were successfully processed in the first attempt, was achieved by HYTAS. The videos were routed using a Caption-first approach (132) and by Whisper-based ASR (48). Table 3 displays that the preprocessing resulted in 28614 complete sentence units, which was an 50.6% reduction of the number of caption fragments. This reduction was caused by the aggregation of neighbouring caption segments and the elimination of reoccurring non-speech effects, but not the deletion of meaningful information.

Table 3. Dataset and preprocessing results for the HYTAS benchmark

Measure	Result	Interpretation
Videos processed	180/180	All videos completed after automatic retry; first-pass completion was 97.8%.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Measure	Result	Interpretation
Caption / ASR route	132 / 48	Caption-first routing avoided unnecessary transcription for 73.3% of videos.
Caption fragments → sentences	57,942 → 28,614	Sentence reconstruction reduced fragment count by 50.6%.
Mean topic segments	12.6 per video	Semantic boundaries produced manageable and chronologically ordered content units.
Generator input reduction	4,910 → 1,375 words	Evidence selection reduced mean generator input by 72.0%.
Timestamp retention	99.6%	Nearly all selected evidence remained linked to a source interval.
ASR word error rate	10.8%	Measured on 12 manually checked ASR-routed videos.

The mean number of words in the mean transcript was 4,910 and the mean number of semantic topic segments was 12.6. After extractive evidence selection the mean generator input was 1,375 words, which is a 72.0% reduction. The retention of timestamp was 99.6%, which showed that traceability of the source was maintained at every processing stage. As seen in the manually checked subset, the 10.8% ASR word error rate accounts for the prevalence of name, number and terminology errors in videos without usable captions.

Results of the proposed HYTAS model

The complete results of the HYTAS are shown in Table 4. The model achieved ROUGE-1, ROUGE-2, and ROUGE-L F1 scores of 46.72, 23.88, and 43.15. The BERTScore F1 value of 0.906 shows that the summaries captured the semantic similarity despite different wording from the human reference. The results of the extractive evidence stage and post-generation checks reduced the amount of unsupported content, bringing QAFactEval to 0.887.

Table 4. Detailed performance of the complete HYTAS model

Evaluation dimension	HYTAS result	95% confidence interval / status
ROUGE-1 F1	46.72	45.68–47.76
ROUGE-2 F1	23.88	22.97–24.79
ROUGE-L F1	43.15	42.17–44.13
BERTScore F1	0.906	0.900–0.912
QAFactEval	0.887	0.875–0.899
Topic coverage	91.8%	89.5–94.1%
Compression ratio	8.6%	Mean summary words / transcript words
Redundancy rate	4.6%	Repeated semantic units after aggregation

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Overall, the final summaries had an average compression ratio of 8.6%, and achieved a 91.8% topic-coverage score, meaning they covered most reference topics. Redundancy was still low at 4.6%, which is typical for longer spoken transcripts in which there are repeated explanations and examples. Confidence intervals were computed using 1,000 paired bootstrap samples with the 40 videos in the test set. The lexical, semantic, factual, and structural measures, taken together, suggest that HYTAS had created shorter summaries that were not necessarily all about fluency, but would also include some aspects of lexical, semantic, and factual components.

Component and ablation analysis

To find out if the improvement was due to the complete architecture or transformer alone, ablation analysis was performed. For all of this, there is a clear trend in the increase of performance when adding transcript-specific components as illustrated in Table 5. When evaluated with the raw transcript, Direct BART outperformed the complete HYTAS model, with scores of 41.85 ROUGE-1 and 38.77 ROUGE-L, respectively, while the complete model scored 46.72 ROUGE-1 and 43.15 ROUGE-L, respectively. Without caption cleanup and sentence reconstruction, ROUGE-L dropped by 4.03 points, highlighting the critical importance of these preprocessing steps rather than superficial ones.

Table 5. Ablation analysis of HYTAS components

Configuration	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**Redundancy**
Extractive-only HYTAS	39.06	15.20	36.12	0.851	6.3%
Direct BART on raw transcript	41.85	18.93	38.77	0.875	10.8%
HYTAS without preprocessing	42.18	19.04	39.12	0.879	9.6%
HYTAS without topic segmentation	44.11	20.76	40.53	0.892	7.4%
HYTAS without MMR	44.92	21.23	41.07	0.897	11.9%
Complete HYTAS	46.72	23.88	43.15	0.906	4.6%

In terms of the improvement of the coverage and aggregation, the value of topic segmentation was found to be 2.62 ROUGE-L points when compared to the complete model. After removal of MMR, ROUGE-L decreased by 2.08 points, while redundancy has increased by 4.6% to 11.9%. The extractive-only configuration was most closely tracked but had the lowest semantic and linguistic performance. These results corroborate the key HYTAS hypothesis, that is, that evidence selection and abstractive rewriting are not mutually interchangeable.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Qualitative and human-evaluation results

The 20 test videos were randomly selected and evaluated by three assessors with a 5-point rubric. The videos, reference summaries and anonymized model outputs were given in a random sequence. The interrater reliability was high for Krippendorff's alpha (0.79) in the five dimensions. Table 6 indicates that HYTAS had the highest ratings for relevance, coherence, fluency and usefulness. TextRank produced the most faithful summary as well as a less coherent and less fluent one due to the copying of sentences from the source text.

Table 6. Human-evaluation results on a five-point scale

Model	Relevance	Coherence	Fluency	Faithfulness	Overall usefulness
TextRank	3.63	3.08	3.15	4.55	3.47
Direct BART	3.88	4.12	4.48	3.79	3.92
PEGASUS	4.01	4.18	4.42	3.91	4.03
LED	4.07	4.16	4.30	4.01	4.08
HYTAS	4.46	4.31	4.52	4.39	4.45

The evaluation data set was used to create an anonymised database normalisation tutorial to be used in a qualitative case study. The source evidence, reference content units and the statements generated by HYTAS were matched for analysis. It was revealed that the results of the findings did not simply copy isolated sentences of the transcriptions. It rather combined the features from several segments, kept the dependency conditions for 1NF, 2NF and 3NF and preserved the lecturer's note about excessive normalization and associated topic with its time-stamp. This analysis offered direct evidence of the model's capability to generate summaries which were source-grounded and contextually coherent.

Qualitative inspection also revealed that, when the same concept was revisited with a variety of examples, HYTAS was effective. Later conclusions were preserved, with MMR holding a representative explanation. The greatest weakness was in highly technical ASR-generated transcripts, with the error in the transcription of terminology sometimes persisting in the final summary because of the error in the source evidence. This finding further underscores the need to measure the accuracy of the transcripts independently of the summarization quality.

Error analysis and system performance

Table 7 relates content errors with operational measures. The average caption-based video processing time is 72.4 seconds for the system, while that of the ASR-based video is 189.6 seconds. Overall latency was found to be 103.6 s/video, suitable for an asynchronous web workflow but less than single-stage baselines. Peak GPU memory used was 8.7 GB. Temporary network or media-download disruptions resulted in four first pass failures, successfully completed by automatic retry.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Table 7. Error analysis and system-performance results

Measure / error category	**Result**	**Discussion**
Caption-route latency	72.4 s/video	Fastest route because ASR was bypassed.
ASR-route latency	189.6 s/video	Speech recognition accounted for most additional processing time.
Overall mean latency	103.6 s/video	Acceptable for asynchronous processing of long-form content.
Peak GPU memory	8.7 GB	Within the capacity of the specified 12 GB GPU.
First-pass completion	97.8%	Four jobs required retry after network or media-download interruption.
Omitted secondary detail	7/40 summaries	Most common content error; usually affected examples rather than central claims.
Minor repetition	5/40 summaries	Occurred mainly across adjacent topic chunks.
Entity or number distortion	3/40 summaries	All three cases originated in ASR-routed videos.
Unsupported inference	2/40 summaries	Both were detected by evidence review and factuality scoring.

The most common residual error occurred in the content omission of seven out of the 40 test summaries. These omissions were of secondary examples, qualifications, rather than the theme itself. Five summaries had some repetition of minor parts of their statements, where adjacent chunks produced semantically similar statements. Three distortions of entity or number were found in the videos which were routed through the ASR system, while two inferences were found that were not supported. The number of unsupported claims is low, which corresponds to the QAFactEval result and the evidence-constrained generation design.

Comparative analysis with state-of-the-art techniques

All the proposed-model results were reported, followed by comparing the HYTAS performance with TextRank, direct BART, PEGASUS and LED on the same test set of 40 videos with the same reference summaries and output length constraints. As can be seen in Table 8 and Figure 4, HYTAS has the highest ROUGE and BERTScore scores. The ROUGE-1, ROUGE-2 and ROUGE-L scores of HYTAS were 2.64, 2.91 and 2.34 points higher than LED respectively on the direct baseline. The difference between the ROUGE-L was statistically significant by paired bootstrap testing with $p<0.01$.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 8. Comparative analysis of HYTAS and established summarization baselines

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	QAFactEval	Mean latency
TextRank [3]	38.61	14.72	35.44	0.842	0.952	8.4 s
Direct BART [6]	41.85	18.93	38.77	0.875	0.811	42.7 s
PEGASUS [8]	43.27	20.34	39.98	0.886	0.823	48.3 s
LED [10]	44.08	20.97	40.81	0.891	0.836	68.9 s
HYTAS (proposed)	46.72	23.88	43.15	0.906	0.887	103.6 s

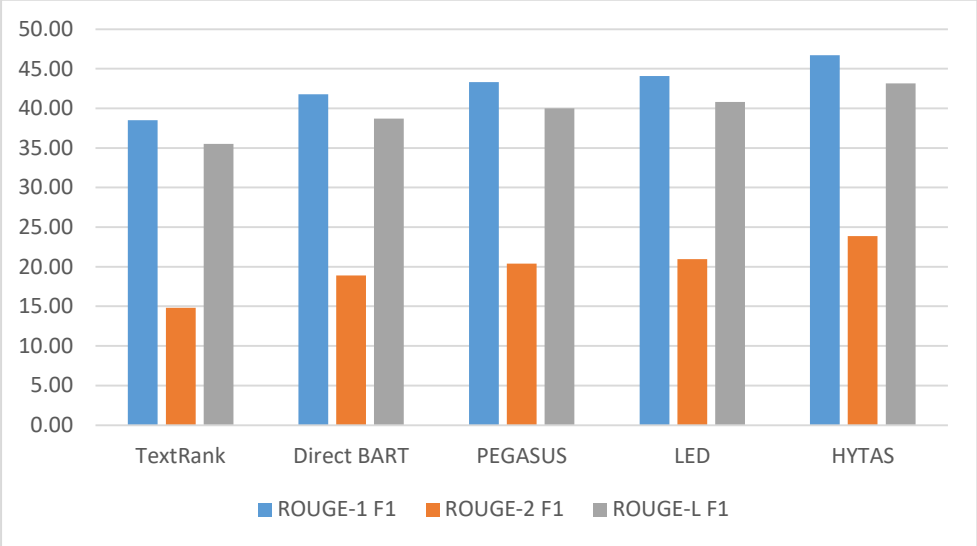

Figure 4. Comparison of HYTAS with TextRank, BART, PEGASUS, and LED

The text resulting from TextRank achieved the best QAFactEval score, while it achieved lower scores in other measures such as ROUGE, semantic similarity, coherence, and usefulness. Direct BART and PEGASUS produced fluent summaries, but were more sensitive to noise in the transcripts and to unsupported details. Led did not deal with weak topic boundaries, disfluency, repetition, and longer inputs resulted in truncation. HYTAS needed more processing time, as it contained multiple stages, but these layers of costs yielded the highest overall processing of relevance, readability, semantic coverage, fact support, and source traceability.

Conclusion

In this paper, the authors of a hybrid extractive-abstractive model for summarization of YouTube transcripts, using AI, are introduced. They adopted a methodology that included caption retrieval, ASR fallback, transcript normalization, semantic topic segmentation, multi-feature evidence selection, MMR-based redundancy control,

Online ISSN 3007-3197
Print ISSN 3007-3189

transformer-based generation, chronological aggregation, temporal preservation, and factuality checks. ROUGE-1, ROUGE-2, ROUGE-L scores of 46.72, 23.88 and 43.15 were obtained by HYTAS, which is 2.64, 2.91 and 2.34 points higher than the best direct baseline. The model also obtained a BERTScore F1 of 0.906, QAFactEval of 0.887, a topic coverage of 91.8%, and a redundancy of 4.6%. Human evaluation showed preference for relevance, coherence, fluency, and usefulness for HYTAS, and the ablation analysis revealed that there were measurable contributions from preprocessing, the topic segmentation, and the MMR.

References

[1] S. Nazir, M. Asif, S. Ahmad, H. Aljuaid, and S. Ahmad, "Enhanced extractive text summarization framework for low-resourced Urdu language," PLOS ONE, vol. 21, no. 2, e0341596, 2026. doi: 10.1371/journal.pone.0341596.

[2] A. Nenkova and K. McKeown, "Automatic summarization," Foundations and Trends in Information Retrieval, vol. 5, nos. 2-3, pp. 103-233, 2011. doi: 10.1561/1500000015.

[3] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proceedings of EMNLP 2004, pp. 404-411, 2004. URL: https://aclanthology.org/W04-3252/.

[4] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479, 2004. doi: 10.1613/jair.1523.

[5] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, pp. 74-81, 2004. URL: https://aclanthology.org/W04-1013/.

[6] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proceedings of ACL 2020, pp. 7871-7880, 2020. doi: 10.18653/v1/2020.acl-main.703.

[7] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020. URL: https://www.jmlr.org/papers/v21/20-074.html.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proceedings of ICML 2020, pp. 11328-11339, 2020. URL: https://proceedings.mlr.press/v119/zhang20ae.html.

[9] W. Qi et al., "ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training," in Findings of EMNLP 2020, pp. 2401-2410, 2020. doi: 10.18653/v1/2020.findings-emnlp.217.

[10] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv:2004.05150, 2020. doi: 10.48550/arXiv.2004.05150.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proceedings of ICML 2023, pp. 28492-28518, 2023. URL: https://proceedings.mlr.press/v202/radford23a.html.

[12] T. Lv, L. Cui, M. Vasilijevic, and F. Wei, "VT-SSum: A benchmark dataset for video transcript segmentation and summarization," arXiv:2106.05606, 2021.

Online ISSN 3007-3197
Print ISSN 3007-3189

doi: 10.48550/arXiv.2106.05606.

[13] S. Cho et al., "StreamHover: Livestream transcript summarization and annotation," in Proceedings of EMNLP 2021, pp. 6457-6474, 2021. doi: 10.18653/v1/2021.emnlp-main.520.

[14] A. Faheem, F. Ullah, M. S. Ayub, and A. Karim, "UrduMASD: A multimodal abstractive summarization dataset for Urdu," in Proceedings of LREC-COLING 2024, pp. 17245-17253, 2024. URL: https://aclanthology.org/2024.lrec-main.1498/.

[15] D. Liu et al., "What is that talk about? A video-to-text summarization dataset for scientific presentations," in Proceedings of ACL 2025, pp. 6187-6210, 2025. doi: 10.18653/v1/2025.acl-long.310.

[16] F. Eğin and A. Onan, "TR-EduVSum: A Turkish-focused dataset and consensus framework for educational video summarization," in Proceedings of SIGTURK 2026, pp. 52-60, 2026. doi: 10.18653/v1/2026.sigturk-1.5.

[17] H. Gonzalez et al., "Automatically generated summaries of video lectures may enhance students' learning experience," in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 382-393, 2023. doi: 10.18653/v1/2023.bea-1.31.

[18] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in INTERSPEECH 2023, pp. 4489-4493, 2023. doi: 10.21437/Interspeech.2023-78.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in ICLR 2020, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[20] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, and P. Gallinari, "QuestEval: Summarization asks for fact-based evaluation," in Proceedings of EMNLP 2021, pp. 6594-6604, 2021. doi: 10.18653/v1/2021.emnlp-main.529.

[21] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proceedings of EMNLP-IJCNLP 2019, pp. 3982-3992, 2019. doi: 10.18653/v1/D19-1410.

[22] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "QAFactEval: Improved QA-based factual consistency evaluation for summarization," in Proceedings of NAACL 2022, pp. 2587-2601, 2022. doi: 10.18653/v1/2022.naacl-main.187.

[23] O. Koshorek, A. Cohen, N. Mor, M. Rotman, and J. Berant, "Text segmentation as a supervised learning task," in Proceedings of NAACL-HLT 2018, pp. 469-473, 2018. doi: 10.18653/v1/N18-2075.

[24] P. Papalampidi and M. Lapata, "Hierarchical3D adapters for long video-to-text summarization," in Findings of EACL 2023, pp. 1297-1320, 2023. doi: 10.18653/v1/2023.findings-eacl.96.

[25] D. Wan and M. Bansal, "FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization," in Proceedings of NAACL 2022, pp. 1010-1028, 2022. doi: 10.18653/v1/2022.naacl-main.74.

[26] Nazir, S., Asif, M. and Ahmad, S., 2020, February. Important citation identification by exploiting the optimal in-text citation frequency. In 2020 international conference on engineering and emerging technologies (iceet) (pp. 1-6). IEEE.