

UNIVERSITÉ DE LORRAINE

MASTER'S THESIS

Off-Topic Detection in Automatic Spoken Language Assessment System

Author:

Pin-Xun HUANG

Supervisor:

Sylvain COULANGE

*A thesis submitted in fulfilment of the requirements
for the degree of Master of Science in Natural Language Processing
in the*

Laboratoire d'Informatique de Grenoble
Laboratoire de linguistique et didactique des langues étrangères et maternelles
Institut des sciences du Digital, Management & Cognition



October, 2025 - August, 2026

Declaration of Authorship

I, Pin-Xun HUANG, declare that this thesis titled, “Off-Topic Detection in Automatic Spoken Language Assessment System” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:



Date: 17. August. 2026

UNIVERSITÉ DE LORRAINE

Abstract

Laboratoire d'Informatique de Grenoble
Laboratoire de linguistique et didactique des langues étrangères et maternelles
Institut des sciences du Digital, Management & Cognition

Natural Language Processing

Off-Topic Detection in Automatic Spoken Language Assessment System

by Pin-Xun HUANG

Large language models and speech processing technologies have allowed a generalisation of spontaneous speaking automated assessment systems in second language acquisition field. Statistical models are now able to predict quite precise proficiency levels based on diverse speech features from acoustic to linguistic quality, yet it remains difficult to judge whether the produced speech actually answers the question asked to the student, as there are infinite ways of answering.

In the framework of the SELF placement test Speaking assessment module design, we have been testing three prompt-response relevance approaches for off-topic detection: 1) how likely the answer is given a large language model and a context prompt; 2) the semantic similarity between the prompt and the answer; and 3) an instruction following language model to evaluate prompt-response relevance (LLM-as-a-judge).

A dataset of 445 responses to one task from the SELF English placement test has been made, including 127 artificially augmented off-topic responses. Based on our dataset, semantic similarity and LLM-as-a-judge methods performed better than LLM likelihood to identify off-topic answers correctly, with the former being more deterministic and light-way.

Moreover, finding a consensus definition of off-topic remains an ongoing discussion, and regardless of which method adopted, the impact of prompt descriptions cannot be ignored.

Acknowledgements

First, no words could ever express my gratitude to my supervisor Sylvain. Thank you for welcoming me into the SELF Project, and for guiding me with such care throughout my whole internship. Under your mentorship, I came to believe in myself.

I would like to thank all my colleagues at GETALP and LIDILEM. You showed me the charm of academic research, and made me feel the warmth of Grenoble.

My thanks go to IDMC and to all the professors who introduced me to the world of NLP. I never imagined I could gain such a breadth of knowledge during my Master's studies.

Lastly, I want to thank everyone I love. Your presence by my side has been my greatest source of strength.

Contents

Declaration of Authorship	iii
Abstract	v
Acknowledgements	vii
Dedicated Acknowledgements	xv
1 Context of the Internship	1
1.1 Host Organisation: LIG-GETALP and LIDILEM	1
1.2 The SELF Project	1
1.3 The SELF PO Project	2
1.3.1 Speaking Assessment Module Development	3
1.3.2 Scoring Pipeline	4
1.3.3 Non-Scorable and Problematic Responses	5
2 Off-Topic in Spoken Test	7
2.1 Topic Relevance	7
2.1.1 Example Responses with Manual Scoring	7
2.1.2 How to Define Off-Topic Responses	8
2.2 Related Work	10
2.2.1 Off-Topic Detection in Written Assessment	10
2.2.2 Off-Topic Detection in Spoken Assessment	11
2.3 Goals of Off-Topic Detection in SELF PO	12
3 Methodology	13
3.1 Approaches to Off-Topic Detection	13
3.1.1 Decoder-Only Transformer	13
3.1.2 Sentence Transformer	14
3.1.3 LLM-as-a-Judge	15
3.2 Combining and Comparing the Approaches	15
3.2.1 Logistic Regression	16
3.2.2 Evaluation Metrics	17
3.3 Dataset Construction	18
3.3.1 Overview of the SELF PO Dataset	18
3.3.2 Tasks Presentation and Case Study on Storytelling	20
3.3.3 Prompt Design	22
3.3.4 Off-Topic Responses Creation	24
3.3.5 Final Dataset for Off-Topic Detection in SELF PO	25
4 Results	27
4.1 LLM Likelihood	27
4.2 Semantic Similarity	29
4.3 LLM-as-a-judge	31

4.4	Comparison of the Three Approaches	34
5	Discussion	39
5.1	Advantages and Limitations of Each Approach	39
5.2	Challenges and Limitations of the Present Work	40
5.3	Future Work	42
6	Conclusion	43
A	LLM-as-a-Judge	45
	Bibliography	47

List of Figures

1.1	The SELF PO scoring pipeline	4
3.1	The rating interface	19
3.2	The number of recordings per CEFR level	20
3.3	10 tasks piloted in SELF English across two parallel test forms	21
3.4	The images of task: Storytelling	23
3.5	Composition of the storytelling evaluation dataset	26
4.1	LLM likelihood by class, neutral prompt description.	28
4.2	LLM likelihood by class, interpreted prompt description.	29
4.3	Semantic Similarity by class, neutral prompt description.	30
4.4	Semantic Similarity by class, interpreted prompt description.	31
4.5	LLM-as-a-judge by class, neutral prompt description.	32
4.6	LLM-as-a-judge by class, interpreted prompt description	33

List of Tables

3.1	Three approaches are compared, each modelling relevance in a fundamentally different way	13
3.2	Each task presents an audio question, in some cases accompanied by an image or a text.	21
4.1	LLM likelihood: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.	27
4.2	Semantic Similarity: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.	29
4.3	LLM-as-a-judge: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.	31
4.4	Different scores and reasons for one single response–audio 1763323244	33
4.5	Different scores and reasons for one single response–audio 1758983136	34
4.6	Precision, recall, F1 and ROC-AUC across three single approaches. . .	34
4.7	F1 and ROC-AUC by prompt variant, across seven combinations. . . .	35
4.8	Detection rates by off-topic construction method, for the three best configurations in Neutral run	37
4.9	Detection rates by off-topic construction method, for the three best configurations in Interpreted run	37

Dedicated Acknowledgements

During this work, AI tools were used for formatting, translation, grammar checking, and test-case generation. I conducted this research independently.

Chapter 1

Context of the Internship

1.1 Host Organisation: LIG-GETALP and LIDILEM

The internship took place jointly in two laboratories of Université Grenoble Alpes.

GETALP (Groupe d'Étude en Traduction Automatique / Traitement Automatisé des Langues et de la Parole) is a team of the LIG (Laboratoire d'Informatique de Grenoble). It was created in 2007 at the founding of the LIG, descending from the earlier GEOD and GETA teams of the CLIPS laboratory. GETALP is a multidisciplinary team of computer scientists, linguists, phoneticians, translators, and signal processing specialists, whose objective is to address the theoretical, methodological, and practical aspects of multilingual communication and information processing, in both written and spoken form. Its work spans machine translation, automatic speech transcription, text processing and semantic representation, and multilingual and multimodal processing. The present work relates most directly to the last two of these: it processes transcribed speech and compares it relevantly against a task prompt.

LIDILEM (Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles) is the linguistics and language didactics laboratory of Université Grenoble Alpes, and the lead laboratory of the SELF project. It was founded in 1987 through the federation of five research centres, on the initiative of Louise Dabène, and is under the sole supervision of Université Grenoble Alpes. Its work spans the language sciences: linguistic description, sociolinguistics, language acquisition, corpus building and exploitation, language didactics, and natural language processing, unified by a shared focus on language in context and on language learning and teaching. Within this range, the present work belongs to second language testing and specifically to the assessment of L2 speaking.

Therefore, this internship sits at the intersection of the two laboratories: it applies GETALP's expertise in speech and natural language processing to the language assessment questions developed within LIDILEM, through the SELF project presented in the following section.

1.2 The SELF Project

SELF (Système d'Évaluation en Langues à visée Formative) is an online, formative oriented language placement test developed by Université Grenoble Alpes to position learners by level ¹. It was created within the IDEFI project Innovalangues (ANR-11-IDFI-0024, 2012–2020), an initiative for improving language teaching. Its objectives are to position students efficiently and to provide a consistent test design

¹<https://self.univ-grenoble-alpes.fr/>

across many languages, answering the institutional need for effective placement in a multilingual setting.

The test is not time constrained but is calibrated to take approximately one hour regardless of level, and reports levels on the CEFR scale from A1 to C1. It assesses three skills: oral comprehension, written comprehension, and short written expression. Afterwards, it returns a global target level together with a level for each skill immediately on completion.

SELF is currently available in seven languages: English, French, Spanish, Italian, Mandarin, Japanese, and German. It is used by approximately 40 universities in France and worldwide, administering around 40,000 tests per year. It was developed by an interdisciplinary team of several dozen researchers engineers, and teachers, with LIDILEM as the lead laboratory and Monica Masperi as director of the SELF project.

SELF is built on four principles of language assessment:

- **Validity** — the test measures the ability it claims to measure;
- **Reliability** — results are consistent across sessions and across raters;
- **Equity** — different learner groups receive comparable treatment;
- **Fairness** — the test design does not disadvantage any learner group.

The last two principles in particular return in Section 5.2, where an automatic speaking assessment is shown Challenges and Limitations of Present Work.

1.3 The SELF PO Project

SELF PO (Production Orale) is the spoken production strand of SELF, and it is where this internship is located. Université Grenoble Alpes is actively developing an automatic speaking assessment module for the test. Two modules are currently under development, one for English and one for French. SELF PO is a collaboration between LIDILEM and LIG, and benefits from substantial financial support from Université Grenoble Alpes.

Oral production is a central component of communicative language competence, and it is also the hardest skill to assess at scale. The reasons are partly technical: spoken responses must be recorded, transcribed and analysed before any linguistic judgment can be made, and each of these steps can occur error. They are also historical, in that oral production has often been left aside relative to the other competences, which are easier to deliver and to score automatically.

The difficulty is compounded by the nature of the speech itself. Students' spoken responses frequently contain hesitations, repetitions, self corrections and pronunciation variation, whether because of their proficiency level or because of nervousness during the exam. Responses may also be short or incomplete. Any automatic system must be robust to noisy, imperfect and heterogeneous data, and must distinguish between different dimensions of speaking ability: communicative fluency, linguistic accuracy, and content relevance, rather than treating them as one undifferentiated quality.

Manual scoring of such responses is costly. It requires well trained, certified raters, who are difficult to recruit; it requires those raters to apply consistent criteria and to agree with one another; and also it takes considerable time. An automatic

speaking assessment module addresses this cost, and can support large scale formative assessment by providing faster and more systematic analysis. This is the motivation for developing SELF PO, whose module complements the existing SELF assessments of listening comprehension, reading comprehension and short written expression, giving a more complete picture of a learner's abilities.

In practice, at the end of the test, students respond orally to a series of questions by recording themselves through their microphone. Then an oral production score is calculated automatically from these recordings and added to the SELF skills profile.

The speech processing builds on **PLSPP**, the Pauses and Lexical Stress Processing Pipeline (Coulange et al., 2024), originally designed to annotate pause positions and lexical stress patterns in spontaneous non-native English speech. PLSPP evolved for a broader annotation of speech, based both on acoustic and transcription.

Project Team

- Sylvain Coulange (LIDILEM, LIG-GETALP) — Technical and Scientific Coordinator
- Anne-Cécile Perret (CUEF) — SELF French Coordinator
- Alex Carr (LIDILEM) — Instructional Design
- Alireza Pournouri (LIDILEM) — Instructional Design
- Estelle Méléze (Service des Langues) — Instructional Design
- Eli Stafford (LIG-GETALP) — Speech Processing
- Pinxun Huang (LIDILEM, LIG-GETALP) — Speech Processing

The English module follows a timeline running from task design and implementation in May 2025, through data collection, speech processing, human evaluation, model training, and SELF implementation, to experimental deployment in 2026/2027. The French module follows the same stages on a later schedule. SELF PO English comprises ten tasks eliciting spontaneous spoken responses; these tasks and the data collected are presented in Section 3.3.

1.3.1 Speaking Assessment Module Development

The module does not score speech by applying a fixed set of rules. It is a trained system, and understanding how it was built is necessary to understand what its outputs mean and where they can fail. Development proceeded in five stages.

Task creation and piloting. A set of tasks was designed to elicit spontaneous spoken responses across the proficiency range, and piloted before being used for collection.

Data collection. The tasks were administered through SELF, and recordings were collected without any automated assessment being applied to them. At this stage the system produced no scores; it gathered speech.

Human expert rating. Trained raters listened to a subset of the collected recordings and scored them on a range of dimensions, providing the reference judgments the system is trained to reproduce. This rating campaign and the interface used for it are described in Section 3.3.1.

Automatic feature extraction. From each recording, features were extracted covering pronunciation, fluency, syntax, vocabulary, and semantics. That is how the response was spoken and what it contained.

Score prediction model training. A model was trained to map these features onto the human ratings, learning the relationship between measurable properties of a response and the level an expert would assign to it.

Once trained, the model predicts a proficiency level for a new recording without a human rater being involved. Two consequences follow, and both matter for this thesis. First, the quality of the predicted scores depends heavily on the training dataset, its coverage of tasks and levels, its size, and how representative it is of the population being tested. Second, it depends on the quality of the human ratings themselves, since the model can only learn to reproduce the judgments it is given. A system built this way inherits both the strengths and the gaps of the data it was trained on.

1.3.2 Scoring Pipeline

The scoring pipeline is the sequence of processing steps that turns a student’s audio recording into a CEFR level for oral production. Figure 1.1 shows the full sequence.

The pipeline receives two inputs: the raw audio recording, and the name of the task the response answers. It receives the task name and not the text of the prompt. This is a detail with consequences: off-topic detection, the subject of this thesis, is the one component that needs the prompt itself, since it must compare what was asked with what was said.

Before a score is produced, the pipeline applies two distinct mechanisms to responses that may not be suitable for scoring. Filtering is a hard rejection: the recording is excluded and downstream processing stops. It operates on the audio signal during preprocessing, using checks on duration, speech activity and audio quality. Flagging is a soft warning: the recording is kept and scored, but marked as potentially unreliable. Flags are raised at two points: some from the audio signal at preprocessing, others only after transcription, since they depend on what was said. Off-topic detection is one of the latter.

Recordings that pass the filters are transcribed by the speech recognition module of PLSPP, which also produces a time-aligned annotation. Transcript-level warning flags are then computed. The response is analysed both syntactically and acoustically, and features are extracted covering both how it was spoken and what it contained. Finally, the prediction model maps these features onto a CEFR level from A1 to C1+.

The two mechanisms are the subject of the next section, where the responses they are designed to catch are described.

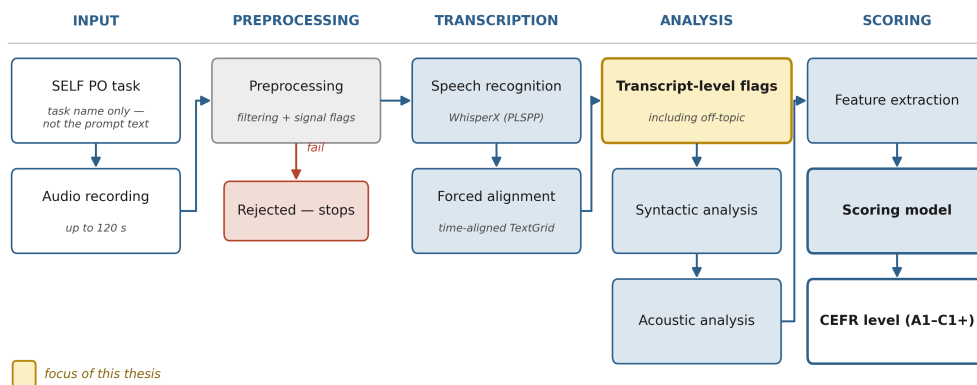


FIGURE 1.1: The SELF PO scoring pipeline.

1.3.3 Non-Scorable and Problematic Responses

Not every recording is suitable for standard scoring, and such recordings fall into the two mechanisms described above.

Unusable audio is filtered out. This covers recordings that are too short, entirely silent, too noisy for the speech to be recovered, or in a file that cannot be processed. In these cases, feature extraction is judged too unreliable, and processing stops.

Problematic content is flagged and scored with a warning. This covers responses that can be processed but whose content is questionable: speech with a high probability of being an other language than the target language, responses containing very few words, responses that merely repeat the question, overly generic responses such as *I don't know* or *I don't understand*, these are typical cases of filtering out (number of words), but in case of longer generic answers it is only flagged, and responses that do not answer the task will be seen as off-topic responses.

Among these, off-topic responses are the most challenging case. The others announce themselves: a silent recording or a two-word answer is recognisable from the surface of the response. An off-topic response is not. It may be fluent, well pronounced, grammatically sound and lexically rich, and look in every respect like a good answer, except that it answers a different question. A pipeline attending to fluency, vocabulary and grammar may consequently assign it a level it does not deserve.

This thesis focuses on testing different approaches to detect off-topic responses in automatic spoken English assessment system as a warning flag used.

Chapter 2

Off-Topic in Spoken Test

2.1 Topic Relevance

Our automatic scoring system computes a CEFR score based on acoustic, linguistic and semantic features. A response that displays high quality speech but does not address the task risks receiving a score it does not deserve: its acoustic, linguistic, and semantic quality are all high, yet relevance what the response is talking about is not among the features the system measures. Nebhi and Szaszák (2023) state an automated system may assign a high score to a response that is well formed but off-topic, however, a human rater would only give a high score to on-topic responses and penalise those that are not.

Judging a response as on-topic or off-topic is a relevance judgment. It compares what the task asked for with what the student actually said, which is a different question from how well the student answered it.

The judgment is not easy to make. Some responses are clearly unrelated to the task and some clearly address it, but between these false responses that stay within the right domain and use the expected vocabulary while not answering the question. These are the hardest cases, for human experts and automatic systems alike, and they recur throughout this thesis.

2.1.1 Example Responses with Manual Scoring

To provide reference judgments for the SELF PO module, 14 human experts were recruited to manually score more than 800 recordings on a range of dimensions, including topic relevance, recorded as a binary off-topic checkbox. Most recordings were evaluated by two or three raters, which allows their judgments to be compared. The rating interface and the full set of dimensions are described in Section 3.3.1.

Several recordings from the storytelling task illustrate why relevance is difficult to judge, and why raters do not always agree with each other.

Short answers are hard to place. A response may be too brief to show whether the student addressed the task or simply produced very little:

audio 1761217684 —*i read three books*

The expert who heard this recording judged it on-topic. On review, my supervisor and I judged it as off-topic response, and it is labelled off-topic in the dataset used in this thesis in Section 3.3.5. The response mentions books, which appear in the task images, but it narrates nothing and does not tell the story for the given task. Whether it is a minimal attempt at the task or a remark that fails to perform, it is a judgment call, and two judgments were made differently for this response.

Proficiency responses may be far from the expected, yet still addressing the task. A beginner learner student with limited vocabulary and simple grammar may produce something that does not resemble an expected answer at all:

audio 1760984350 —*i need books okay honey here robinson it's a great book i need more books okay what is she doing*

This response is labelled on-topic, the student mentions the books and the book title that appear in the picture sequence. But it is fragmentary and does not narrate the story, we had some debates within this response, but there's no correct judgement for it.

audio 1758917040 —*one day a little girl asked the librarian for her book first we have her robinson crusoe and she read it quickly then she came back and asked for more books this time he gave her many books by goethe the girl also finished them very fast finally the librarian was surprised because she still wanted to read more she even stood on the books to reach a new one on the top of the shelf*

The expert believed that the student did tell a story, but the ending of the story did not match the task of the pictures, also the content deviated from the narrative line of the task we asked for, yet it was still rated as on-topic.

audio 1758744184 —*a little girl comes into the room asking the man for a book the man hands her a book she leaves and asks for another book the man gives her multiple books and wonders what she's going to do with it as he walks outside he sees her standing on a pile of books attempting to grab fruit from the tree*

audio 1763135546 —*the little girl asked the librarian for some big books the man gave her one book at first but the little girl asked for more big books the little girl needed the books to reach the fruit in the tree*

These two recordings were judged on-topic by the same expert, and by listening these two answers, the expert suspected the speech was not natural enough, sounded more like a prepared script.

What these instances show is that the difficulty is not only technical. The same recording can receive different relevance judgments from different judges, boundary setting to say what is a topic shifting, unrecognisable and any automatic system is evaluated against labels that carry this uncertainty.

2.1.2 How to Define Off-Topic Responses

There is no single definition of off-topic in the literature. This subsection reviews the framings that recur across written and spoken assessment, then states the definition adopted in this thesis.

Response to an unexpected prompt. An off-topic response is a well formed response that addresses a different topic than the one asked about. Higgins, Burstein, and Attali (2006) introduce this framing for essays, defining off-topic essays as those not written to the topic of the test-question. Also distinguish unexpected-topic essays which are well formed but on the wrong topic, from the bad-faith responses where the student is not attempting the task at all.

Response not relevant to the assigned prompt, possibly within the right domain. The relevance view is a primary framework used in spoken assessment Zechner and Evanini (2019). Malinin et al. (2016) frame the task as judging whether a

response is on topic with respect to the question, and Raina, Gales, and Knill (2020) define it as assessing whether the response is relevant to the prompt.

Persing and Ng (2014) go further and treat relevance not as a binary property but as a graded prompt-adherence score. This is effective because it matches what the instances in Section 2.1.1. Relevance is better understood as a continuum, ranging from clearly unrelated responses, through partially relevant or topic-shifting, to responses that completely address the task. The difficulty is the middle region, responses that remain within the right domain and use the expected vocabulary without answering the question.

Deliberate behaviour: memorised, copied or canned responses. Yoon and Xie (2014) and Wang et al. (2019) emphasise that test-takers may attempt to inflate their scores by reciting prepared answers to familiar questions, speaking in their first language, or repeating the question back. This is a type of off-topic behaviour, and it is distinguished by intent Yoon et al. (2018) instead of the relationship between prompt and response.

Off-topic as one category within a non-scorable group. Yoon and Xie (2014) defined off-topic as not related to the prompt, as one labelled type alongside no response, generic, question copy and non English responses. On this framing, off-topic is one member of a family of responses that cannot be scored in the standard way.

Off-topic responses in SELF PO

In the SELF PO data, off-topic responses take several forms:

- the student speaks about a completely different domain
- the student ignores the question and recites a prememorised answer in order to display vocabulary and grammar
- the student only repeats the question
- the student speaks within the right domain and uses the target words, but has not understood the question and consequently does not answer it.

The last of these is the case that motivates this work. It is the middle region of the relevance continuum, it cannot be identified from just the surface of the response, and it is the one that other warning flags do not cover.

Definition adopted in this thesis. This work follows the relevance framing, and takes the four forms listed above as the definition of what it detects. An off-topic response is a spoken response falling into any of them: addresses a different domain, recites a prememorised answer, repeats the question, or stays within the right domain without answering the question. All of them fail to address the task it was given, judged by comparing the task prompt with the content of the spoken response.

According to our definition, any response does not require any judgment about the student's intent, it is only about the relationship between prompt and response. Although we compute the probability of being off-topic which is a continuous value, it is operationalised as a binary decision, since the warning flag of the module raises is binary and The system was evaluated against human ratings which were collected as a binary checkbox.

Off-topic detection is one warning flag among several, and it is the only one that uses the task prompt, since it is the only one that must compare what was asked

with what was said. And it raises a warning rather than rejecting the recording: a flagged response is still scored, but marked as potentially unreliable Section 1.3.2.

2.2 Related Work

Automatic off-topic detection has been studied for around years, first in written assessment and subsequently in spoken assessment. This section reviews both parts. The written exam literature starts first because the spoken test builds on it. Both share the idea of comparing a content of the response with the prompt. The research works from lexical overlap towards deep semantic modelling, but one more concern for the spoken literature: the response is not given as text, need to first be recognised from speech.

2.2.1 Off-Topic Detection in Written Assessment

In written assessment, the core operation is to compute the similarity between a context of the essay and the question. In the early relied on comparing words directly which performed poorly when the prompt was short and contain few words against.

Higgins et al. (2004) measured relatedness between each sentence of an essay and the prompt by using Random Indexing, a vector based semantic representation, as one of four dimensions of essay coherence. When the relationship did not exist, the authors treated it as possible evidence that the student had written an off-topic essay. Higgins, Burstein, and Attali (2006) applied content vector analysis to off-topic detection, scoring the cosine between the content vectors of an essay and a prompt without requiring topic specific training data. The vector space representation underlying these methods originates outside educational assessment with Salton, Wong, and Yang (1975). In Attali and Burstein (2006)'s account, e-rater V.2 is described as employing content vector analysis among a small set of features covering grammar, mechanics, style, development, lexical complexity and prompt specific vocabulary. Even though the paper is not an off-topic detection study, it is worth knowing for a design decision is that e-rater flags anomalous and bad-faith essays and withholds a score from them.

A recurring difficulty in the early work is the sparseness of the prompt itself. Louis and Higgins (2010) reduced the limitations of vocabulary matching by augmenting brief prompt texts and rectifying spelling errors within student submissions, thus enhancing detection precision where standard methods failed.

Unlike the work relying on word matching, Persing and Ng (2014) adopted a feature-rich approach to evaluate prompt adherence. The model incorporates n-grams and LDA-based semantic generalizations together with error-type features to assess performance on a fine-grained graded scale.

Rei (2017) extends the problem to visual prompts, mapping images and sentences into a shared distributed vector space and evaluating their semantic similarity, a setting close to our case study: Storytelling where the content to be produced is carried by images rather than by text.

Recently, Das, Vadi, and Yadav (2024) developed the Automated Open Essay Scoring framework. This transformer-based approach uses a topic regularization module alongside Mahalanobis distance scoring to rate essays and detect off-topic in an unsupervised environment.

Higgins et al. (2004) noted responses maintain a tangential thematic link to the prompt but fail to resolve the actual task. This work observed that vector-based

techniques like Random Indexing are often insensitive to such failures because the existence of relevant words does not guarantee task fulfillment. This limitation is a property of measuring relevance as vector similarity, and Section 3.1.2 shows it persisting in a modern sentence encoder.

2.2.2 Off-Topic Detection in Spoken Assessment

In spoken assessment, the response must be transcribed first so that it can be compared with the prompt, but recognition errors propagate into the relevance judgment, and the system has to balance the content relevance against the acoustic features which speaking scores also depend on, these add more challenges for scoring.

Malinin et al. (2016) introduced a recurrent neural network language model to evaluate the conditional probability of responses relative to test questions. By accounting for the sequential of speech, this method offers superior scalability. However, it depends on precollected cases which prevents the system from assessing unseen prompts, that were not represented in the initial dataset.

Yoon et al. (2017) measure the distance between the spoken response and the prompt using average word embeddings, weighted word embeddings and word mover's distance. Lee et al. (2017) adopt a Siamese convolutional network with shared weights, processing prompt and response simultaneously and computing the semantic distance between them in the feature space that the approach is extended by Wang et al. (2019).

Wang et al. (2019)'s work also introduce the similarity grid model, which changes semantic matching as an image problem. Cosine distances between content words in the response and in the prompt form a grid then classified by an image recognition network. Raina, Gales, and Knill (2020) apply the same technique, and note that it helps to distinguish topic specific responses from generic memorised templates.

Malinin, Knill, and Gales (2017) solved the cold-start problem through the Hierarchical Attention Topic Model. Instead of only training the system on seen prompts, the second layer learns how those training prompts relate to one another in vector space. When a new prompt appears, it can estimate where this new prompt belongs relative to the ones it already knows. Later on Raina, Gales, and Knill (2020) showed that combining this model with similarity grid models works well together, each captures different aspects of the detection problem, and their combination exceeds any single model in unsupervised settings.

Nebhi and Szaszák (2023) integrate audio features like acoustic, prosodic and spectral, also with text features from a multitask BERT model, treating scoring, coherence modelling, and prompt relevancy scoring as three tasks learned together. This work to relevance is different from the similarity methods above, instead of measuring the distance between a response and its prompt, they shuffle responses across prompts and train the model to predict which prompt a given response was written for. Then relevance is recast as a classification problem over the prompt set.

All in all, the evolution of off-topic detection in both written and spoken assessment has moved from simple word frequency matching towards deep neural semantic modelling, furthermore, spoken exam requiring robustness to transcription error.

A work exploits multimodal consistency, to detect anomalous responses Metallinou and Cheng (2014) compare syllable locations derived from the audio signal with those derived from the ASR output, inconsistencies between these two indicate non scorable or off-topic responses.

2.3 Goals of Off-Topic Detection in SELF PO

Five requirements were set for an off-topic detection component in SELF PO.

It should work for the current tasks. The ten SELF PO English tasks are actively used, and any component must be able to process the responses they elicit.

It should work for future unseen tasks. As the assessment development continues and new tasks are added, choosing a method that can adapt every prompt is preferable.

Keep low computationally cost. Off-topic detection is one of the warning flags in the whole scoring pipeline. Its cost should stay proportional to the role, so a lightweight model that do not add significant load to the pipeline is what we wish for.

It should be open source. The SELF PO scoring module is intended to be open source, if there was a closed component would make the module from being inspected, redistributed and reused. Therefore, restrict to open licensed models that can be run locally.

Transferable to other languages. SELF is available in seven languages, and SELF PO is intended to expand beyond English. A method that can be applied to French, German, Italian, Japanese or Mandarin without replacement is better to only one language.

It should align with the teacher's view. The purpose of the flag is to raise a warning that human raters would agree with. If a system disagrees with the experts about which responses are off-topic, would not be ideal in a formative assessment.

Section 5.2 returns to these requirements and states how far each was addressed.

Chapter 3

Methodology

3.1 Approaches to Off-Topic Detection

Off-topic detection is formulated here as a **prompt-response relevance estimation task**. Each instance consists of a task prompt, constructed as described in Section 3.3.3, and a spoken response represented as its transcription. The objective is to determine whether the response sufficiently addresses the prompt.

The choice of models follows from the requirements set out in Section 2.3. All three are open source and openly licensed, which the SELF PO scoring module requires of every component it depends on: a closed component would prevent the module as a whole from being inspected, redistributed and reused. GPT-2 and the Sentence-Bert are lightweight, which suits a component whose role in the pipeline is a single warning flag. Mistral is considerably larger, and is included precisely to test whether an instruction-following model estimates relevance accurately enough to justify that additional cost.

Each approach produces a single numeric relevance score for a prompt-response pair. The three scores: one is negative log-likelihood, one is cosine similarity, one is rating on a fixed integer scale. The purpose of comparing multiple approaches is to identify which type of relevance signal is most useful for off-topic detection in SELF PO English. The results are reported in Section 4.

Approach	Model	Relevance signal
Decoder-only transformer	openai-community/gpt2	Likelihood-based predictability
Sentence transformer	distiluse-base-multilingual-cased-v1	Embedding-based semantic similarity
LLM-as-a-judge	Mistral-7B-Instruct-v0.2	Explicit instruction-based judgment

TABLE 3.1: Three approaches are compared, each modelling relevance in a fundamentally different way

3.1.1 Decoder-Only Transformer

A decoder-only transformer is a language model that predicts the next token from the preceding context. GPT-2 is used here as a representative model of this type, in the openai-community/gpt2 (Radford et al., 2019) the smallest version with 124M parameters, without any fine-tuning. In this case, when we adopt GPT-2 for evaluating prompt-response relevance, the model will first read the question prompt then the response, and it will calculate the probability of the response content similar to the question.

How the score is computed

The prompt and the response are tokenised separately and their token sequences concatenated directly, prompt first, with no separator token between them. The loss is computed over the response tokens only, with the prompt tokens masked out, so the model is scored on how well it predicts the response given the prompt rather than on the prompt itself. The relevance score is the average negative log-likelihood (NLL) per response token, which makes it independent of response length, since SELF PO responses range from five seconds to two minutes of speech. Where a prompt–response pair exceeds the model’s 1,024-token context window, the response is truncated to fit.

A lower NLL means the response is more predictable given the prompt; a higher NLL means it is less compatible with it. The hypothesis is that an on-topic response continues the prompt more predictably than an off-topic one, therefore produces a lower NLL.

Why this model

GPT-2 is open source and openly licensed, unlike closed models such as GPT-4 (OpenAI, 2023) that are available only through a commercial API and could not be included in an open-source scoring module. It is also small enough to run on CPU, which suits a component that contributes one warning flag among several.

3.1.2 Sentence Transformer

A sentence transformer is an encoder model that maps an entire sentence or passage into a single fixed-length vector representing its meaning, such that texts with similar meaning lie close together in the vector space. The model used here is *distiluse-base-multilingual-cased-v1* (Reimers and Gurevych, 2019), a member of the Sentence-BERT family. By applying this model, it will see how the two sentences’ meaning close to each other, one is question and the other is answer.

How the score is computed

The prompt and the response are encoded separately into dense semantic vectors, and their cosine similarity is taken as the relevance score. Cosine similarity ranges from -1 to 1 , with a higher value indicating stronger semantic relatedness between prompt and response. A small number of responses receive slightly negative scores, indicating no semantic relation to the prompt at all.

Unlike the LLM likelihood approach, this measures relevance as semantic closeness rather than as predictive fit. LLM likelihood asks whether the response is a plausible continuation of the prompt, while Semantic Similarity asks whether the two texts are about the same thing.

Why this model

Because the encoder maps texts into a shared semantic space, this approach captures relatedness beyond exact word overlap: two responses expressing the same content in different vocabulary receive similar embeddings. This is particularly important for the storytelling task, where the same story can be validly retold in many different ways, and a lexical overlap measure would penalise the more inventive retellings.

A multilingual encoder was chosen deliberately. SELF is currently available in seven languages and SELF PO is intended to expand beyond English, so a model that already covers multiple languages transfers to other language modules without replacement. It is also open source, openly licensed and lightweight enough to run without a GPU.

3.1.3 LLM-as-a-Judge

The third approach uses an instruction following language model to evaluate prompt–response relevance directly, rather than deriving a score from model internals. Mistral-7B-Instruct-v0.2 is used as a representative model of this type (Jiang et al., 2023). We let the model become an examiner, it score the responses’ relevance with the question, and comment why giving this score, under a rating instruction.

How the score is computed

The model is given the task prompt, the response, and scoring criteria, and is instructed to rate relevance on a 1–10 scale, returning the score together with a short phrase reason in a fixed JSON format. The numeric score is used directly as the relevance measure; the accompanying reason is not used in scoring but is retained for Section 4.

The instruction explicitly directs the model to disregard grammar, pronunciation and fluency, and to judge only whether the meaning of the response relates to the prompt.

The model is prompted in a zero-shot setting using instructions exclusively. To ensure deterministic and reproducible judgments, we employ greedy decoding during inference.

Refinement of the judging instruction

The instruction was not fixed in advance. An initial minimal version, little more than a request to rate relevance from 1 to 10, produced judgments that were both biased and unstable. This led to a detailed version that specifies the rating scale, definition of off-topic behaviour. Both versions are reproduced in the Appendix A, and Section 4.3 compares their output on individual responses.

Why this model

Mistral-7B-Instruct-v0.2 is open source, openly licensed, and can be run locally, unlike proprietary LLMs which would be costly at scale and incompatible with an open-source pipeline. It also belongs to a different model family from GPT-5.4 Thinking and Gemini 3.1 Pro, used to generate part of the off-topic data in Section 3.3.4, so the judge is never asked to evaluate text produced by a model of its own family.

3.2 Combining and Comparing the Approaches

The three approaches of Section 3.1 produce different kinds of relevance features. **LLM likelihood** captures the predictability of a response given the prompt; **Semantic Similarity** captures semantic closeness between the two texts; **LLM-as-a-judge**

produces a relevance judgment. Whether these features are complementary or redundant cannot be determined from the methods alone. Therefore, the approaches are evaluated both individually and in combination.

Labels

Each response in the dataset carries a binary gold label: 1 for off-topic, 0 for on-topic. These labels are treated as ground truth for the purposes of this evaluation. What the evaluation measures is how well each approach separates the off-topic responses in this dataset from the on-topic ones, not how well it reproduces expert human judgment, which the data does not permit, since the expert raters flagged no storytelling response as off-topic.

Experimental conditions

The evaluation is carried out on the case study: storytelling dataset described in Section 3.3.5. The entire pipeline ran twice, once with the neutral prompt description and once with the interpreted one Section 3.3.3, this isolates the effect of prompt design while keeping all other conditions constant.

3.2.1 Logistic Regression

Logistic regression is used as a lightweight binary classifier that predicts whether a response is off-topic, taking the relevance scores as its input features. The model outputs a probability that a response belongs to the off-topic class, which is converted into a binary prediction by a decision threshold. The threshold was left at its default value of 0.5 and was not tuned, a decision whose consequences are in the results of Section 4.4.

Feature handling

The three scores are on incommensurable scales, an average negative log-likelihood bounded from 0 to infinity, a cosine similarity bounded between -1 and 1, and an integer rating from 1 to 10, so features are standardised before fitting. Z-score standardisation is performed inside the cross-validation pipeline rather than on the full dataset, so the scaling parameters are estimated from the training folds alone, avoiding leakage from the test fold.

Class weights are balanced to compensate for the imbalance between the 318 on-topic and 127 off-topic responses; without this, the classifier would default to predicting the majority class.

Only the three relevance scores are used as input features. The classifier has no access to response source or speaker identity, so any separation it achieves is attributable to the relevance features themselves.

Feature combinations

All **seven** combinations of the three features are evaluated: each of the three features alone, each of the three possible pairs, and all three together. Evaluating the full set rather than only the individual features and the full combination is what makes it possible to see whether a given feature contributes information the others lack, or merely adds noise.

Why this classifier

Logistic regression is appropriate here because the dataset is small. With 445 instances and three features, a higher capacity model would risk fitting the idiosyncrasies of a constructed dataset rather than the relevance feature. It is also computationally efficient, consistent with the requirement that off-topic detection remain proportionate to its role as one warning flag. Plus, it is interpretable, its coefficients show the direction and relative weight of each relevance feature's contribution.

3.2.2 Evaluation Metrics

Cross-validation

Performance is estimated by **stratified 10-fold cross-validation**. The dataset is divided into ten folds, nine are used for training and one for testing, and the procedure is repeated ten times so that each fold serves once as the test set. Given that the minority class accounts for less than a third of the instances, stratification based on topic label is essential to preserve the class ratio in every fold. Averaging performance across the ten folds also gives a more stable estimate than a single held-out split of this size would allow. As well for SELF PO score prediction model evaluation, but the difference is we did a stratification on proficiency and task.

Note on reproducibility

The fold assignment was randomised without a fixed random seed. In order to read Section 4 correctly. First, the neutral and interpreted runs used independently randomised partitions, so the comparison between prompt variants is an observation under differing fold assignments rather than a controlled measurement; differences falling within the range of fold-to-fold variation are accordingly not interpreted as genuine performance differences. Second, because repeating the procedure would produce slightly different values, the numbers reported are one realisation rather than fixed quantities. Reporting the standard deviation across folds alongside every mean is intended to make the scale of this variation visible.

Choice of metrics

Accuracy alone is not sufficient for this task. Because the dataset is imbalanced, a model that predicted almost every response as on-topic would still achieve high accuracy while detecting no off-topic responses at all, precisely the failure the flag exists to prevent. Therefore, four metrics are reported, each as a mean and standard deviation across the ten folds, with the off-topic class treated as the positive class:

- **Precision** — the proportion of responses predicted off-topic that are truly off-topic; high precision means few valid responses are wrongly flagged.
- **Recall** — the proportion of actually off-topic responses that are correctly detected; high recall means few off-topic responses go undetected.
- **F1** — the harmonic mean of precision and recall, summarising the two in a single figure.
- **ROC-AUC** — how well the model ranks off-topic responses above on-topic ones, independently of where the decision threshold is placed. A value of

0.5 corresponds to chance; 1.0 to perfect separation. Because it is threshold-independent, ROC-AUC and F1 can disagree about which configuration is best, and where they do, the disagreement is itself informative about threshold choice rather than about ranking quality.

3.3 Dataset Construction

This section describes the data used in this study and how the evaluation dataset for off-topic detection was built from it. It first presents the SELF PO original dataset and how it was manually rated by experts (3.3.1). We then present the ten elicitation tasks and the storytelling task selected as a case study (3.3.2), the design of the prompts against which relevance is measured (3.3.3), the construction of off-topic responses (3.3.4), and the resulting final dataset used for evaluation (3.3.5).

Because genuine off-topic responses proved extremely rare in the SELF PO data, for the straightforward reason that most test takers attempt to answer the question they are given. Therefore, an evaluation set large enough to compare detection approaches had to be constructed consciously, and the way it was constructed constrains what the results in Section 4. The following subsections set out that construction process and report the composition of the dataset in full.

3.3.1 Overview of the SELF PO Dataset

The data used in this study comes from SELF PO English, the oral production component of SELF. Responses were elicited through ten tasks designed to lead to spontaneous speech, presented in Section 3.3.2.

Two sets must be distinguished throughout: the full collection of recordings, and the smaller subset that was manually rated by experts. Only the second one carries the human judgments against which automatic scoring is developed and validated.

Collected data

Across the ten SELF PO English tasks, the collection comprises approximately 5,500 recordings, amounting to around 45 hours of speech from about 800 students.

Manually rated data

A subset of these recordings was rated by experts through a dedicated rating interface. As of April 2026, the human-rated set comprises:

- 14 raters, each rating 150 recordings
- 2,100 individual ratings in total
- 873 distinct recordings, approximately 10 hours of speech
- 339 speakers
- between one and three ratings per recording

The rating interface was configured to collect three ratings per recording, but not every recording received the full three; multiple ratings of the same recording allow inter-rater comparison, which matters for a judgment as subjective as topic relevance.

The rating interface

Before rating, the expert can listen to the task question exactly as the student heard it, displayed on the left of the interface, together with any image or text the student saw. Figure 3.1 shows the interface as the raters used it. After listening to the recording, the expert records the following on the right:

The screenshot displays the 'SELF PO Rating Interface'. At the top, it shows 'Progress: 108 / 150 recordings' and user options 'IGOR' and 'LOGOUT'. The interface is divided into two main sections: a left panel for task context and a right panel for the rating form.

Left Panel (Task Context):

- Task name:** Street_Interview
- Task code:** 10494
- Context:** A video player showing a woman holding a microphone, with a progress bar at 0:00 / 0:03.
- Question:** A video player showing a question, with a progress bar at 0:00 / 0:04.

Right Panel (Rating Form):

- Recording ID:** 1761079945
- Student ID:** 119446
- Recording to Rate:** A video player showing a recording, with a progress bar at 0:00 / 0:29 and a volume slider at 100%.
- Rating Form:**
 - Non-scorable:** ☐ (with a red circle icon)
 - Off-topic:** ☐ (with a yellow triangle icon)
 - Task Completed:** Buttons for NO, PARTIALLY, and YES.
 - Pragmatics:** Buttons for A1, A2, B1, B2, C1, and C2.
 - Vocabulary and Grammar:** Buttons for A1, A2, B1, B2, C1, and C2.
 - Pronunciation / Prosody / Fluency:** Buttons for A1, A2, B1, B2, C1, and C2.
 - Recording's Global Level:** A slider from A1 to C2 with a value of 4 (B2).
 - Rating Confidence Score:** Buttons for LOW, MEDIUM, and HIGH.
 - Comments (optional):** A text input field with the placeholder 'Add any additional comments...'.
 - Submit Button:** A blue button labeled 'SUBMIT RATING AND LOAD NEXT'.

FIGURE 3.1: The rating interface.

- a **Non-scorable** flag and an **Off-topic** flag, both binary
- **Task Completed:** No, Partially, or Yes
- three skill scales: Pragmatics; Vocabulary and Grammar; Pronunciation, Prosody and Fluency
- a **CEFR level** from A1 to C2
- a **Global Level** from A1 to C2, together with a continuum score from 1 to 7
- a **Rating Confidence** score: Low, Medium or High
- an optional text comment

The binary off-topic flag is the human judgment against which the approaches in this thesis are evaluated.

Known limitations of the metadata

The speakers' actual speaking proficiency is unknown: only the SELF global score (A1–C1+, computed from written expression, oral comprehension and written comprehension) is available, and it is a proxy for speaking ability rather than a measurement of it. The speakers' first language is also unrecorded, although the large majority are presumed to be French L1 students, since the recordings were collected during SELF tests administered at French universities.

The distribution of levels across the rated recordings is uneven, with B2 the most frequent level and A1 the least frequent, computed from each student's mean score on the listening, reading and writing components. Figure 3.2 gives the full distribution.

Moreover, the biggest limitation in our dataset is barely has off-topic label, and the disagreement between human raters about what is off-topic, so creating more off-topic responses that similar to human like answer is essential.

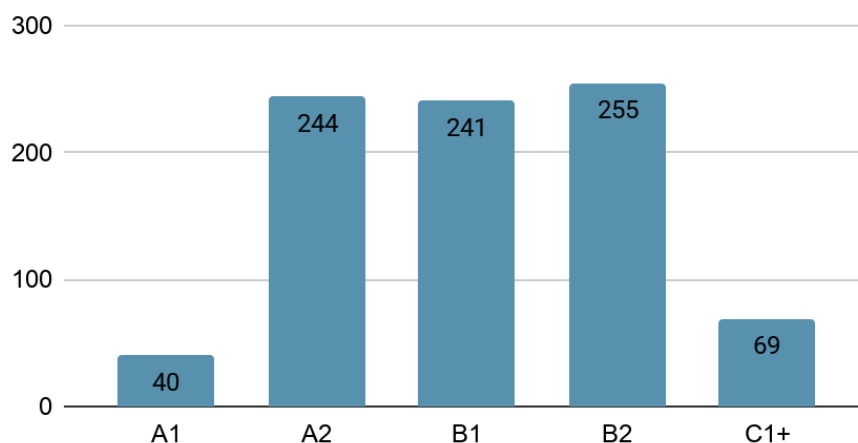


FIGURE 3.2: The number of recordings per CEFR level.

Recording conditions

Two features of the recording procedure affect the data. First, students are given two attempts at each task and choose which recording to submit as their final answer, so the collection consists of responses the students themselves considered acceptable. Second, students may record for up to 120 seconds; recordings shorter than 5 seconds were excluded from the collection used here, as they almost always contain no speech at all or only background noise.

3.3.2 Tasks Presentation and Case Study on Storytelling

The ten tasks

Ten tasks eliciting spontaneous speech have been used across two parallel tests, named AMBER and JADE (Figure 3.3). Each student answers to one set or the other. Two tasks are shared across both tests.

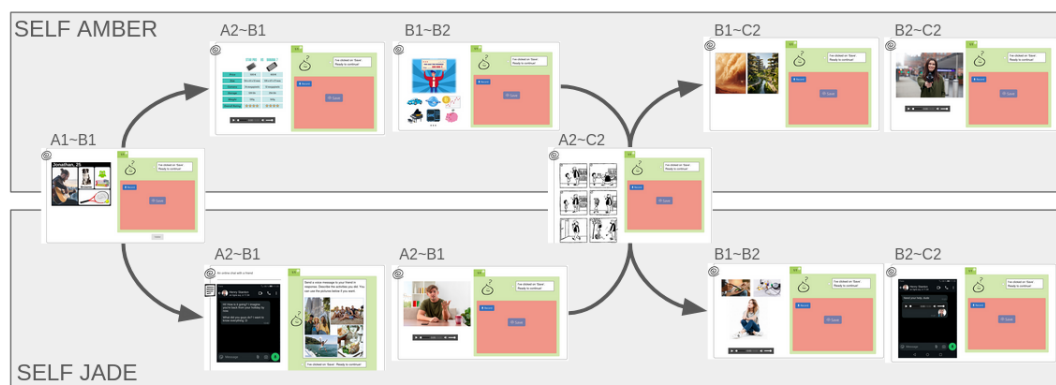


FIGURE 3.3: 10 tasks piloted in SELF English across two parallel test forms.

Students answer every task in their form regardless of their own proficiency, so each task receives responses spanning the full range of levels. The intent here is rather to see how each proficiency level behave on each task whatever its target level.

Task	Description	Visual support
1	Describing a person the student has just met	Image giving the person's details
2	Talking about activities done during the summer holiday	Optional images offered as ideas
3	Comparing two phones and giving a recommendation	Table giving the phones' specifications
4	Apologising to a friend and giving an explanation	None
5	Imagining what the student would do after winning a prize	Optional images offered as ideas
6	Describing two pictures and saying which resembles the planet's future	Two images
7	Advising on the pros and cons of university path choices	None
8	Giving an opinion on climate change in a street interview	None
9	Giving and defending advice about an unexpected sum of money	None
10	Telling a story from six ordered pictures	Six sequential images

TABLE 3.2: Each task presents an audio question, in some cases accompanied by an image or a text.

Case study: Storytelling

Task 10: Storytelling, was selected as the case study for this thesis. The task presents six pictures in a fixed order that together tell a story about two people, and asks the student to narrate it.

Two properties of the task motivated the choice. First, it is accessible across the whole proficiency range: a beginner can produce a minimal sequence of events and an advanced learner can produce a structured narrative, so the task yields usable responses at every level rather than only at the top. Second, it elicits comparatively

long responses, since narrating a six-picture sequence requires more speech than answering a single question.

3.3.3 Prompt Design

Off-topic detection in this study is formulated as a comparison between what the task asked for and what the student said. The task side of that comparison has to exist as text. In this thesis, the **prompt** is the transcribed text of the spoken task question, together with a written description of the accompanying image. The storytelling task is used here as the worked example.

The central design concern

The prompt is not a neutral object. Because every approach in Section 3.1 measures a response against this text, whatever the prompt says implicitly defines what counts as an acceptable answer. The central concern in writing the prompts was to avoid biasing the relevance judgment against responses that address the task in unexpected ways.

Three principles were adopted in response. A textual description of the question and task context, leaving them open enough to admit varied on-topic responses. We prefer plain, generic wording over specific vocabulary, which reduces the pull toward surface lexical overlap with any one expected answer. And each image was described only to the extent needed to convey its task-relevant content, so the level of detail is not standardised across the ten tasks.

How the prompt is built depends on task type

The ten tasks differ in how they use images, and this difference determines how the prompt is constructed. Three axes arise:

- **Audio-only tasks** (4, 7, 8, 9). There is a contextual image, which only put the task in a given communicative situation with a specific interlocutor, and the question transcription alone forms the prompt.
- **Audio-plus-image tasks** (1, 3, 6, 10). The image carries information the response must address, and the question is incomplete without it. The prompt combines the transcription with a textual description of the image.
- **Audio-plus-optional-image tasks** (2, 5). Images are offered only as suggestions the student may use or ignore, so they do not define what an on-topic answer contains.

The design concern above applies most sharply to the second group, and mostly to storytelling, where the images are not supporting material but the content to be produced.

Authorship of the descriptions

I subjectively wrote the descriptions, following as strictly as possible the principles I've given above. However, what if a different person wrote the description, probably it will form differently. So in order to test the impact of different descriptions of the same task, I've tested 2 different prompts in the following part; a large language model was used only to correct the grammar of the description text, never to generate its content; the model used for this correction was Gemini.

Case study: Storytelling prompt

Storytelling Figure 3.4 presents six ordered pictures telling a story about two people. The spoken question carries almost no content of its own, transcription reads: *Telling a story. Take a look at these pictures and tell the story.* Everything a response must address is in the images, so the prompt necessarily combines this transcription with a description of the six-picture sequence.

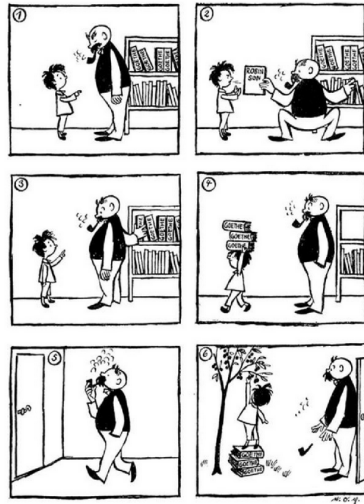


FIGURE 3.4: The images of task: Storytelling.

Because this task depends heavily on the description, it was used to test the design concern directly. During initial testing, the way I described the storytelling's pictures was like I was the test-taker. But the output was not ideal because each student has their own way to respond, if the prompt was completely following my perspective, then I will penalise the creativeness of each answer. Therefore, **Two description variants were written: a neutral one and an interpreted one.** The audio-question transcription is identical in both, only the image description differs.

The **neutral** variant restricts itself to observable actions and objects:

A child and a man stand facing each other near a bookshelf containing many books. The man holds a book out toward the child, and the child gestures with both hands raised. The child extends a finger toward the bookshelf while facing the man. The child walks away carrying three books stacked on top of the head while the man stands behind. The man is walking toward a door while holding a pipe. The child stands on the three books under a tree and reaches upward toward fruit, while the man stands by the door as a pipe falls from the hand.

The **interpreted** variant includes inferred intentions and emotions:

A child and an old man are standing near a bookshelf filled with books and talking. The man offers a book from the bookshelf to the child, but the child refuses it. They continue talking as the child points toward the bookshelf. The man stands behind and watches the child walk away while carrying three books stacked on the head. The man walks toward the door while smoking a pipe. The man stands near the door and drops the pipe in surprise upon seeing the child standing on the books and trying to grab fruit from a tree.

The two differ in their level of interpretation rather than in the events they describe. Where the neutral variant records that *the child gestures with both hands raised*, the interpreted variant states that *the child refuses the book*; where the neutral variant records that *a pipe falls from the hand*, the interpreted variant states that *the man drops the pipe in surprise*. The interpreted variant supplies the mental states a narrator would supply, however, the neutral variant leaves them to be inferred.

The neutral variant is more neutral, not free of assumption. No verbal description of an image is entirely free of it. Both variants were therefore carried through the entire evaluation, and every result in Section 4 is reported for both.

3.3.4 Off-Topic Responses Creation

Human evaluation of SELF PO recordings ran from December 2025 to early April 2026. In the latest snapshot, 2,100 ratings covering 873 recordings contained 14; restricted to the storytelling task, we have 271 ratings for 121 recordings. My work is based on 82 recordings which had been rated 160 times (from the first snapshot of February) and none of them carried an off-topic flag, we would like to expand our dataset but there is no more off-topic response during the collecting time.

This rarity is expected rather than anomalous. But it makes a purely observational evaluation impossible, no approach can be meaningfully compared against another on a handful of positive cases.

Consequence: we decided to create artificial off-topic answers to test our detection models. Constructing off-topic examples is established practice in the literature for precisely this reason (Yoon and Xie, 2014; Malinin et al., 2016; Yoon et al., 2017; Wang et al., 2019). Four construction methods were used here, chosen to cover different kinds of off-topic behaviour rather than to maximise the number. Each contributes a distinct type of case, and each carries limitations.

Method 1 — Manual creation

31 Off-topic responses were written by hand against the storytelling prompt, deliberately spanning a continuum from obviously unrelated to genuinely ambiguous, including responses that stay within the right domain and use the target vocabulary while not answering the question. For example: *a teacher gave three books to his student he wanted the student to work harder*.

Method 2 — SELF PO task-answer mismatch

40 Genuine on-topic responses to the other nine SELF PO tasks were re-paired with the storytelling prompt, which makes them off-topic with respect to it while leaving the speech itself untouched. Each response was drawn from one of the other nine tasks, randomly assigned, with the tasks roughly evenly represented. For example, from Task 1 (describing a person): *jonathan is twenty five years old his hobbies are tennis and guitar his dog's name is brownie he speaks three languages french italian and spanish*.

Method 3 — External corpus

30 Responses were drawn from *Spontaneous dialogues in L2 English by Japanese-L1 speakers* (Sylvain Coulange, 2026), a corpus of Japanese university students engaged in ten-minute role-plays involving argumentative discussion of contentious topics, a task format inspired by the CLES English certification. The corpus is segmented

into speech turns, and isolated speech turns were used as off-topic responses, monologic stretches of speech of duration comparable to SELF PO spoken answers. For example: *well I'm not sure if you're going to take that line of reasoning I'm not sure where it stops because all you need is a computer really or a phone and an internet connection and thats true for a lot of things now i mean do students who have an internet connection have a computer at home have a leg up with research yes but are we going to ban that i dont think so i mean the library is still a resource the computer lab i definitely think we should make sure that students have access to technology but i dont think that that means that we have to ban everything because of that.*

Method 4 — LLM generation

20 Off-topic responses were generated by GPT-5.4 Thinking (OpenAI, 2026) and Gemini 3.1 Pro (Google, 2026). Two types were specified, with five of each produced per model:

Story drift. The response begins on the theme of the images but moves to an unrelated subject partway through. For example: *At first glance, the story looks like it will focus on a kid, a book, and something happening in the room. However, it makes me think about how my daily routine has changed recently. I've been trying to reduce screen time, organize my desk, and focus better. Small changes, like cleaning regularly, really affect my mood and concentration.*

Thematic overlap. The response stays in the semantic neighbourhood of the images but never performs the task. For example: *I recently visited a public library that had a massive bookshelf stretching all the way to the ceiling, and it made me think about how difficult it is to reach the items at the very top. If you aren't tall enough, you usually have to find a sturdy ladder or ask someone else for help. It's always a surprise to see what kind of interesting things people keep in high places, whether they are rare collections or just everyday household items they want to keep out of sight.*

3.3.5 Final Dataset for Off-Topic Detection in SELF PO

The final dataset for the storytelling case study comprises **445 responses: 318 on-topic and 127 off-topic**.

Provenance of the on-topic responses

All 318 on-topic responses are authentic SELF PO storytelling recordings. Of these, 80 come from the 82 recordings that had been rated by experts through the interface described in Section 3.3.1. The remaining 238 had not been reached by the human rating campaign, and I annotated them myself for topic relevance, applying the same binary criterion the raters applied.

Provenance of the off-topic responses

6 out of the 127 off-topic responses are real recordings, manually annotated as off-topic; the remaining 121 were constructed by the four methods described in Section 3.3.4.

The 6 real cases require an clear account, because they were not produced by the expert rating campaign. 4 recordings I annotated as off-topic among the 238 that carried no expert rating. The other 2 are recordings that were expert-rated and that the raters did not flag: on review, my supervisor and I judged them as off-topic and relabelled them.

First, the gold labels for this dataset are effectively single-annotator: I reviewed all 324 authentic recordings, and the expert ratings function as a partial cross-check on 82 of them rather than as the label source. Second, the labels include two cases in which that annotation overrode an expert judgment. Agreement between my annotation and the expert off-topic flag was therefore 80 out of 82 on the overlapping subset.

Composition

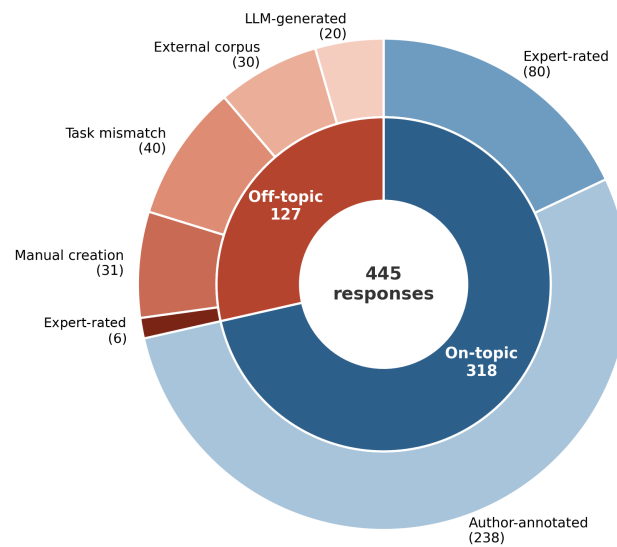


FIGURE 3.5: Composition of the storytelling evaluation dataset. The inner ring gives the on-topic/off-topic split; the outer ring gives the provenance of each class.

The resulting class balance is approximately 71% on-topic to 29% off-topic. This is far denser in off-topic responses than the collected data, where the observed rate is below 1%, and it is a deliberate consequence of the construction.

Chapter 4

Results

The off-topic detection study is carried out on the storytelling final dataset described in Section 3.3.5, containing 445 responses: 318 on-topic and 127 off-topic. The three relevance scores are evaluated individually and in all seven combinations, each feature alone, each of the three pairs, and all three together, using logistic regression with stratified 10-fold cross-validation. Precision, recall, F1 and ROC-AUC are reported as the mean and standard deviation across the ten folds (Section 3.2).

As set out in Section 3.2.2, the two runs used independently randomised fold assignments. Differences between the two prompt variants are accordingly reported as observations rather than as controlled measurements, and differences falling within the range of fold-to-fold variation are not treated as evidence of a genuine performance gap.

This section reports each approach separately (Section 4.1–4.3), then compares them and analyses their errors (Section 4.4).

4.1 LLM Likelihood

Prompt variant	Precision	Recall	F1	ROC-AUC
Neutral	0.518 ± 0.099	0.551 ± 0.130	0.527 ± 0.088	0.722 ± 0.070
Interpreted	0.533 ± 0.099	0.584 ± 0.130	0.553 ± 0.103	0.752 ± 0.082

TABLE 4.1: LLM likelihood: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.

This score corresponds to the average negative log-likelihood (NLL) of the response, conditioned on the prompt. A lower NLL means the response is a more predictable continuation of the prompt. A higher NLL means it fits the prompt less well.

Based on our dataset, this LLM likelihood approach identified about half of the off-topic responses correctly: the F1 score is $0.53 \pm .09$ for the neutral prompt, and $0.55 \pm .1$ for the interpreted one. The interpreted prompt seems to be better suited to the task here, however the overall score is low for both of them.

The ROC-AUC of 0.72–0.75 shows there is still some signal. The model gets this comparison right about three times out of four. So LLM likelihood carries some information, but the two classes overlap heavily.

Figures 4.1 and 4.2 show this overlap between the two categories of answer, for the neutral and interpreted prompts. Off-topic responses do sit slightly higher in NLL than on-topic ones, in the direction the hypothesis predicts, a median of 4.42 against 3.95 in the neutral run. But the two distributions largely overlap: the

interquartile ranges are 3.93–5.41 for off-topic and 3.62–4.27 for on-topic. In the neutral run, 35% of off-topic responses fall inside the on-topic interquartile range; in the interpreted run, 34%. A third of the off-topic class is indistinguishable from most on-topic responses on this feature alone. No suitable threshold can separate them.

A few off-topic responses reach much higher values, an NLL of 22.6 against a maximum of 7.0 for on-topic responses. A small number of off-topic responses are very unlikely under the prompt, and this is what keeps the ROC-AUC above chance. But these responses are a minority. Most of the off-topic class sits inside the on-topic range.

This result does not support the hypothesis from Section 3.1.1 that on-topic responses are more predictable continuations of the prompt than off-topic. What it shows is even narrower, one possible reason is that LLM likelihood reflects general fluency and grammaticality more than topical fit. This cannot be confirmed from the present results, since no independent measure of fluency or grammaticality was compared against it here.

The two prompt variants shows the interpreted description scores marginally better on every metric, but the difference is within the fold-to-fold standard deviation and is not treated as meaningful.

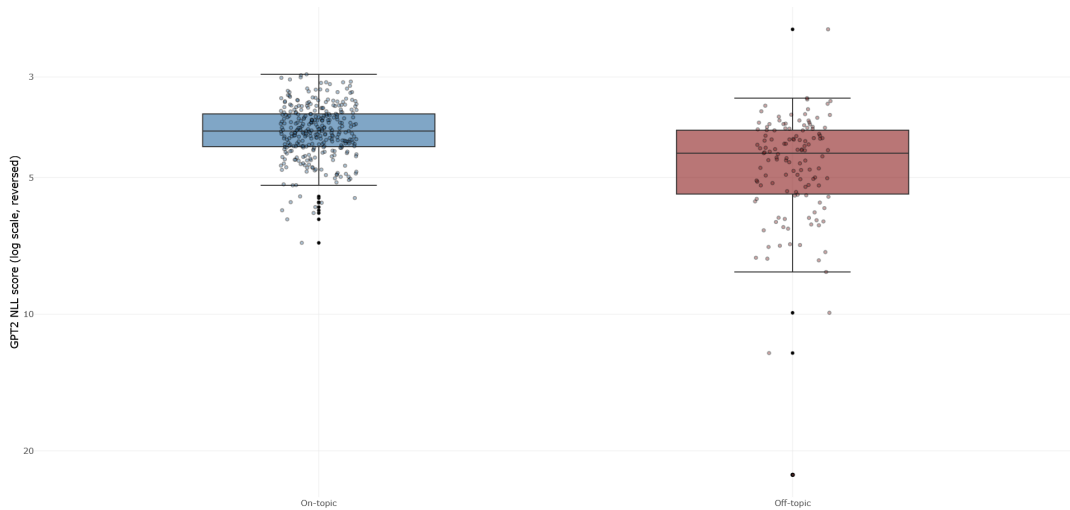


FIGURE 4.1: LLM likelihood by class, neutral prompt description.

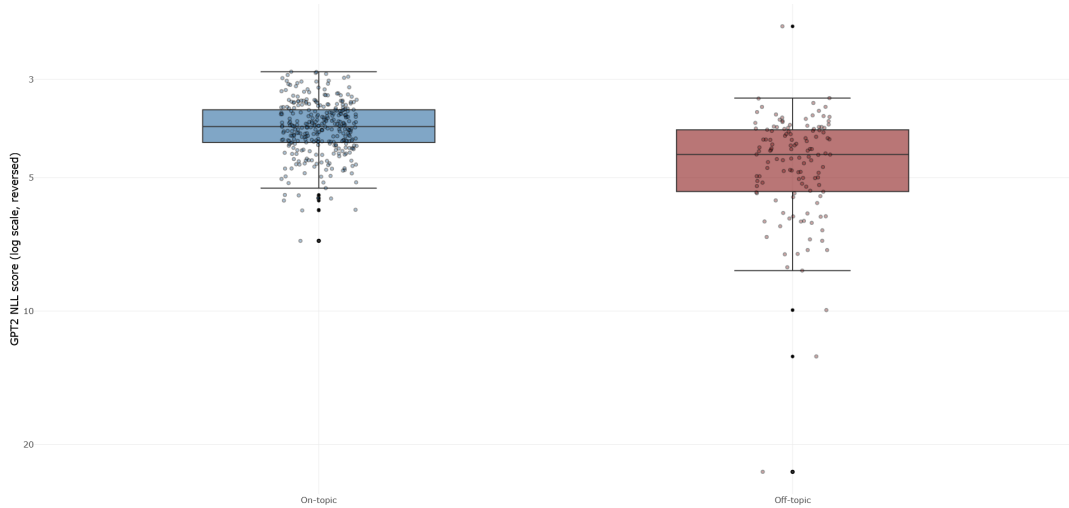


FIGURE 4.2: LLM likelihood by class, interpreted prompt description.

4.2 Semantic Similarity

Prompt variant	Precision	Recall	F1	ROC-AUC
Neutral	0.915 ± 0.071	0.906 ± 0.075	0.908 ± 0.091	0.980 ± 0.022
Interpreted	0.892 ± 0.071	0.896 ± 0.075	0.893 ± 0.065	0.968 ± 0.032

TABLE 4.2: Semantic Similarity: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.

This score corresponds to the cosine similarity between the prompt and response embeddings. A higher similarity means the response is more semantically related to the prompt, which we expect for on-topic responses. From a given semantic distance, we can consider the response off-topic.

Semantic similarity score separates the two classes well, under both prompt variants, with an F1 near 0.90 and a ROC-AUC near 0.97–0.98. The neutral description scores slightly better than the interpreted one on every metric. This matches the reasoning in Section 3.2.2: similarity is computed directly against the prompt text, so a description that sticks to observable actions and objects probably gives a broader target. It fits more valid retellings, while a description that adds intentions and emotions narrows what counts as close. The difference is small relative to fold variation and is reported as an observation, not a strong result.

Figures 4.3 and 4.4 show the score distributions for the neutral and interpreted prompts. The two classes are almost fully separated. In the neutral run, on-topic responses have a median similarity of 0.52 (IQR 0.45–0.57), against 0.07 for off-topic responses (IQR 0.02–0.18). Only 2.4% of off-topic responses fall inside the on-topic interquartile range — far below the 35% seen for the likelihood feature in Section 4.1. The interpreted run shows the same pattern: medians of 0.48 and 0.08, with 4.7% overlap. The remaining errors come from a small number of specific cases, not from general confusion between the classes.

Clearly unrelated responses are detected reliably. Short generic responses such as *i don't understand the question* or *thank you very much*, and responses drawn

from a different task or from the external corpus, get low similarity scores and are correctly flagged.

Three weaknesses show up in the error analysis:

- **Partially relevant responses are not distinctly separated.** Responses that shift topic partway through, or that come from lower proficiency test-takers answered only part of the story, get similarity scores in between. These don't form a clear gap against fully on-topic responses. This is the difficult middle region described in Section 2.1.1, and Semantic Similarity does not resolve it.
- **Responses that reuse the prompt's words without doing the task get high similarity score.** This is the Thematic Overlap case built in Section 3.3.4, and it matches the failure predicted in Section 3.1.2. A response that mentions books, a child, an old man and a tree, without telling the true story, is close to the prompt by exactly the measure this method uses.
- **Question copy responses are not caught.** A response that just repeats the task instruction reproduces the prompt's wording directly, so it gets a near maximal similarity score. It ends up at the opposite extreme from an off-topic response, even though it is just as invalid. Semantic Similarity treats it as the most on-topic response in the whole dataset.

This last weakness points to a fix that was not tested here. A question copy fails by being too similar, not enough different. So a ceiling threshold, alongside the existing floor threshold, could catch it: a response with Semantic Similarity above some upper bound would be flagged as a likely question copy. One feature could then cover three problems at once, unrelated responses, generic responses, and question copies. This idea is proposed as future work in Section 5.3, yet not tested in this thesis.

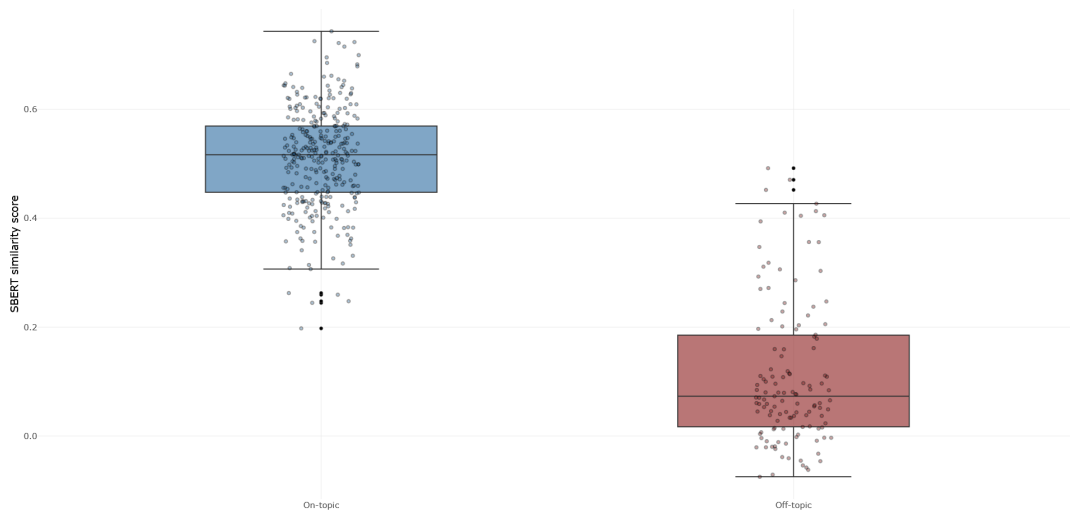


FIGURE 4.3: Semantic Similarity by class, neutral prompt description.

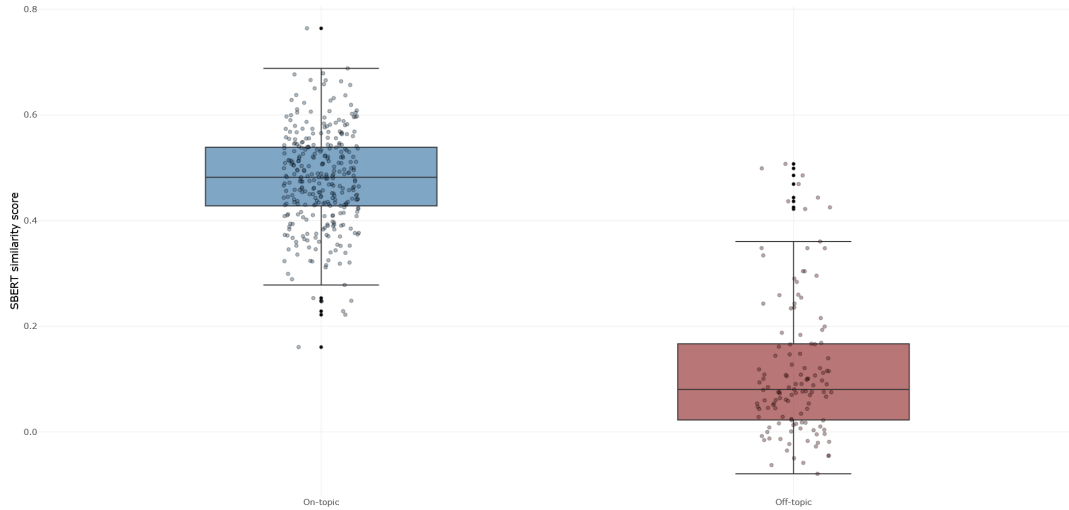


FIGURE 4.4: Semantic Similarity by class, interpreted prompt description.

4.3 LLM-as-a-judge

Prompt variant	Precision	Recall	F1	ROC-AUC
Neutral	0.803 ± 0.069	0.929 ± 0.093	0.852 ± 0.069	0.979 ± 0.015
Interpreted	0.917 ± 0.065	0.905 ± 0.093	0.908 ± 0.064	0.968 ± 0.027

TABLE 4.3: LLM-as-a-judge: precision, recall, F1 and ROC-AUC by prompt variant. Values are mean $\pm$ standard deviation across the ten cross-validation folds.

This score corresponds to the relevance rating from 1 to 10 provided by Mistral-7B-Instruct model. A higher score means the response is judged more relevant to the task.

This approach performs close to the Semantic Similarity one, but seems to prefer interpreted prompt (F1 of 0.85 ± 0.07 and 0.91 ± 0.06). It gets high recall under both prompt variants (0.905 — 0.929), but its precision differs a lot between them, 0.803 with the neutral description, against 0.917 with the interpreted one. In real terms, the neutral description leads it to flag 31 on-topic responses as off-topic, compared with 11 under the interpreted description.

Nevertheless, the ROC-AUC is almost the same across both variants, 0.979 and 0.968 . What differs is where the fixed threshold of 0.5 falls on the ranking (Section 3.2.1). The lower precision under the neutral description comes from where the threshold sits.

Figures 4.5 and 4.6 show how decisive LLM-as-a-judge is. On-topic responses get a median rating of 8 out of 10. Off-topic responses get a median of 1, more than three quarters of the off-topic class gets the lowest rating. Only 3.9% of off-topic responses fall inside the on-topic interquartile range in the neutral run, and 5.5% in the interpreted run.

Unlike the Semantic Similarity approach, this method catches responses that repeat the task question. By applying LLM-as-a-judge to rate whether a response answers the prompt, it has no problem recognizing that a response only repeating the prompt is not valid.

Three limitations are observed:

- **Short off-topic responses are sometimes rated highly.** Responses such as *this is an interesting story* or *thank you very much* get high relevance scores despite not addressing the task. A brief remark that is topically close can be scored as relevant, because LLM-as-a-judge treats it as a comment on the prompt.
- **LLM-as-a-judge is sometimes stricter than the task requires,** assigning some genuinely on-topic responses lower scores than it assigns to some off-topic ones.
- **LLM-as-a-judge reflect biases introduced by the prompt description,** being strongly influenced by how the task and the image description are worded.

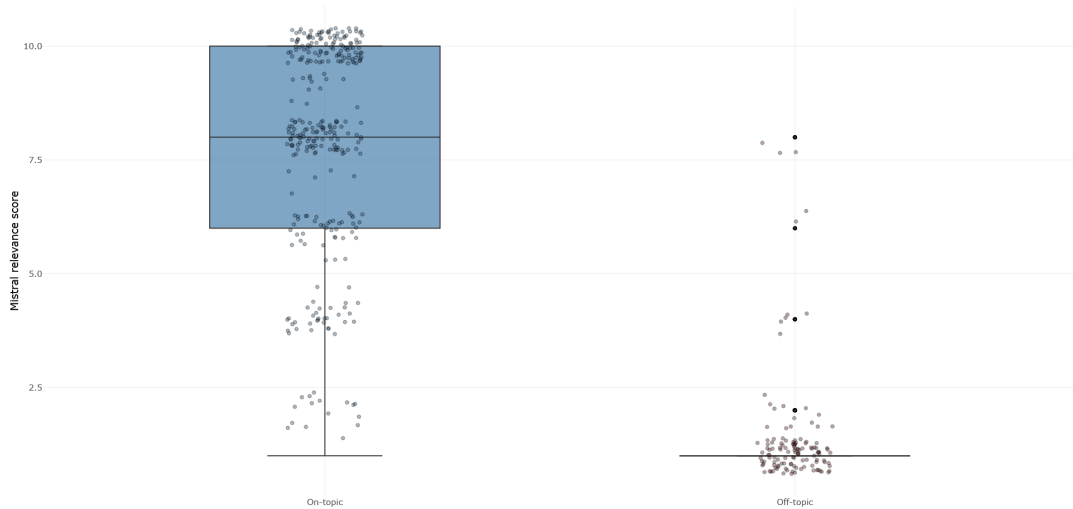


FIGURE 4.5: LLM-as-a-judge by class, neutral prompt description.

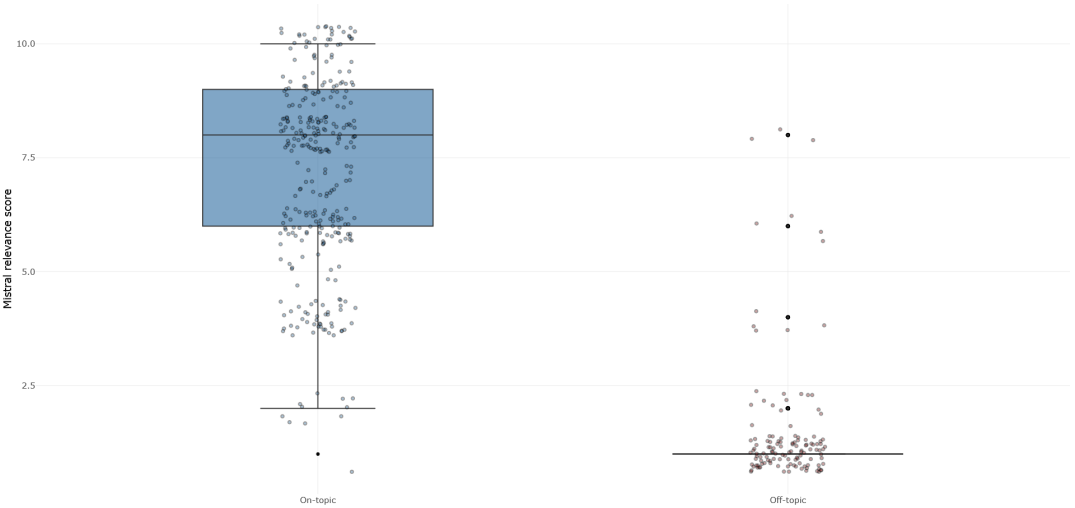


FIGURE 4.6: LLM-as-a-judge by class, interpreted prompt description.

The strictness and bias

Two responses illustrate both effects and both of them recount the story correctly.

audio 1763323244 —so it’s a story of a little girl who just asked a man maybe her father for help he gives her a book named rubinson then he goes out to smoke and as he leaves the house he sees the little girl using the books to climb and reach something in a tree

Instruction	Score	Reason given
Minimal	5	The student answer mentions some elements of the story but not in the correct order and some details are incorrect, such as the gender of the child and the reason for the old man leaving the room.
Detailed, interpreted	6	The response mentions a little girl instead of a child, and the man is described as possibly being the father. The story of the man offering a book and the child refusing it is present, but the response lacks detail about the conversation and the man’s surprise upon seeing the child on the books.
Detailed, neutral	8	The response tells a story involving a child and a man, and includes some elements from the pictures, such as the man giving a book to the child and the child climbing a tree. However, it is not an exact match to the given prompt as the child is described as a little girl and the man is referred to as possibly being her father, while the pictures show a boy and an unidentified man.

TABLE 4.4: Different scores and reasons for one single response—audio 1763323244

audio 1758983136 —so this is the story of a girl that is asking to borrow some books to her father but she doesn’t read these books she uses them to be able to pick some cherries in the tree in her garden

Instruction	Score	Reason given
Minimal	5	The student answer is partially correct, but it incorrectly identifies the gender of the child and focuses too much on the father. The story does not mention anything about the child’s father.
Detailed, interpreted	2	The response is mostly unrelated to the given prompt as it tells a story about a girl borrowing books to pick cherries, which does not involve the child and old man from the pictures or their interaction with the bookshelf.
Detailed, neutral	2	The response is mostly unrelated to the given prompt as it tells a story of a girl borrowing books to pick cherries, which does not match the pictures showing a child and a man interacting with books and gesturing towards each other.

TABLE 4.5: Different scores and reasons for one single response–audio 1758983136

Both responses narrate the sequence the images depict. The second is a compressed but accurate summary of the whole story, and it receives only 2.

Two distinct forms of bias are visible in these judgments:

- **An assumption never present in the prompt.** Neither description specifies the child’s gender, and neither gives the man a title; the neutral description says only a child and a man. LLM-as-a-judge nonetheless asserts that the pictures show a boy, and penalises both responses for describing a girl. This is an inference the model supplied and then evaluated against, and it is invisible in the score alone, also it is recoverable only because the approach returns a reason.
- **An expectation of wording match.** LLM-as-a-judge expects the response to reproduce the vocabulary and the ordering of events given in the description, and assigns lower scores where a student narrates the same content differently. The second response is penalised for compression, not for inaccuracy.

These observations indicate that LLM-as-a-judge is sensitive not only to whether the response addresses the task, but also to how closely it matches the particular wording of the prompt description and of the judging instruction. On the other hand, LLM-as-a-judge produces the same score and almost the same output short phrase for multiple runs on the same responses with greedy decoding during initial testing, it shows a reliable consistency for it to apply on an automatic assessment.

4.4 Comparison of the Three Approaches

Approach	Variant	Precision	Recall	F1	ROC-AUC
LLM likelihood	Neutral	0.518	0.551	0.527	0.722
	Interpreted	0.533	0.584	0.553	0.752
Semantic					
Similarity	Neutral	0.915	0.906	0.908	0.980
	Interpreted	0.892	0.896	0.893	0.968
LLM-as-a-judge	Neutral	0.803	0.929	0.852	0.979
	Interpreted	0.917	0.905	0.908	0.968

TABLE 4.6: Precision, recall, F1 and ROC-AUC across three single approaches.

LLM likelihood is clearly off then Semantic Similarity and LLM-as-a-judge get similar performance but prefer different degrees of interpretation in the prompt. The highest precision, recall and ROC-AUC each belong to a different configuration, and even the two best F1 scores are almost identical while being produced by different approaches under different prompt descriptions. The choice between Semantic Similarity and LLM-as-a-judge cannot simply be ranked one above the other.

Configuration	Neutral F1	Neutral ROC-AUC	Interpreted F1	Interpreted ROC-AUC
LLM likelihood	0.527 ± 0.088	0.722	0.553 ± 0.103	0.752
LLM likelihood + LLM-as-a-judge	0.853 ± 0.066	0.981	0.859 ± 0.075	0.971
LLM-as-a-judge	0.852 ± 0.069	0.979	0.908 ± 0.064	0.969
LLM likelihood + Semantic Similarity	0.903 ± 0.059	0.985	0.894 ± 0.067	0.969
Semantic Similarity	0.908 ± 0.091	0.980	0.893 ± 0.065	0.968
LLM likelihood + Semantic Similarity + LLM-as-a-judge	0.913 ± 0.058	0.985	0.897 ± 0.050	0.978
Semantic Similarity + LLM-as-a-judge	0.930 ± 0.029	0.987	0.901 ± 0.045	0.976

TABLE 4.7: F1 and ROC-AUC by prompt variant, across seven combinations.

Configurations ordered by mean F1 across the two runs, from weakest to strongest. F1 and ROC-AUC are averaged across the ten cross-validation folds; F1 shows the standard deviation across folds.

Combining LLM likelihood with LLM-as-a-judge produces almost exactly LLM-as-a-judge’s own performance under the neutral description, but under the interpreted description it performs worse. Adding LLM likelihood to the Semantic Similarity and LLM-as-a-judge combination lowers F1 under both prompt variants. The three features combination gets the best interpreted ROC-AUC at 0.978.

The best configuration is Semantic Similarity combined with LLM-as-a-judge under the neutral prompt, with an F1 of 0.930 and a ROC-AUC of 0.987. It is also the most stable configuration, a fold standard deviation of 0.029. Combining the two features does not merely raise average performance, it makes performance consistent across folds.

audio 7676767 simply repeats the audio task instruction: Semantic Similarity assigns it a similarity score of 0.41 (neutral) and 0.44 (interpreted), placing it among the on-topic responses, because repeating the prompt reproduces the prompt’s own wording. LLM-as-a-judge assigns it 1 out of 10 under both prompt variants.

audio 7676767 —*telling a story take a look at these pictures and tell the story*

audio 1919191 that uses the right setting and characters, but tells the wrong story. Here Semantic Similarity gives 0.50 under the interpreted description, above the on-topic median, while LLM-as-a-judge scores 2. In both cases LLM-as-a-judge recovers what Semantic Similarity misses.

audio 1919191 —*i saw one little boy was trying to buy cigarette from an old man in a book store but the old man refused him*

When LLM-as-a-judge gets something wrong, Semantic Similarity does not fix it: LLM-generated response, **audio 2026030001**, receives an LLM-as-a-judge rating of 8 out of 10, and its Semantic Similarity score is 0.42 (interpreted) and 0.40 (neutral), low relative to the on-topic median of 0.48 (interpreted) and 0.52 (neutral), but still well inside the on-topic interquartile range, so Semantic Similarity does not flag it either. Both approaches fail on the same response.

audio 2026030001 —*when i see pictures like this i usually think about how stories are told differently depending on age people can have very different perspectives i like stories that involve books because they often suggest learning or imagination but sometimes the meaning is more symbolic than literal and that's what makes every detail interesting*

Semantic Similarity gets its best F1 with the neutral description. However, LLM-as-a-judge obtains its better F1 with the interpreted description. This shows how each method uses the prompt.

Semantic Similarity measures relatedness directly against the prompt text. A description limited to observable actions and objects gives it a wider target, so it fits more valid retelling story, and adding inferred intentions narrows what counts as similar. LLM-as-a-judge works differently, it reasons about content against a stated standard, therefore, it benefits from a description that spells out the narrative it is checking against.

Under the interpreted prompt, LLM-as-a-judge alone gets the highest F1, but the three-feature combination obtains the highest ROC-AUC. Only F1 depend on the decision threshold, so this disagreement means some configurations rank responses well, but are poorly calibrated at the fixed threshold of 0.5, which was never tuned in this study (Section 3.2.1).

The same pattern appeared in Section 4.3, LLM-as-a-judge's precision changed a lot between prompt variants, while its ROC-AUC stayed almost the same. In both cases, tuning the threshold would solve part of the gap.

Error analysis

Errors were examined for the three best-performing configurations in each run, selected by mean cross-validation F1. The three best performing configurations are not the same in the two runs:

- **Neutral:** Semantic Similarity + LLM-as-a-judge (0.930), LLM likelihood + Semantic Similarity + LLM-as-a-judge (0.913), Semantic Similarity (0.908).
- **Interpreted:** LLM-as-a-judge (0.908), Semantic Similarity + LLM-as-a-judge (0.901), LLM likelihood + Semantic Similarity + LLM-as-a-judge (0.897).

LLM-as-a-judge alone and Semantic Similarity alone each appear in only one of the two rankings. The other two configurations appear in both lists, but in different positions.

Construction method	n	Semantic Similarit + LLM-as-a-judge	LLM likelihood + Semantic Similarity + LLM-as-a-judge	Semantic Similarity
Real, manually annotated	6	1.00	1.00	0.83
Manual creation	31	1.00	1.00	0.97
Task mismatch	40	1.00	1.00	1.00
External corpus	30	1.00	1.00	0.97
LLM generation	20	0.65	0.65	0.55
All off-topic	127	0.94	0.94	0.91

TABLE 4.8: Detection rates by off-topic construction method, for the three best configurations in Neutral run

Construction method	n	LLM-as-a- judge	Semantic Similarity + LLM-as-a-judge	LLM likelihood + Semantic Similarity + LLM-as-a-judge
Real, manually annotated	6	0.67	0.83	0.83
Manual creation	31	1.00	0.97	0.94
Task mismatch	40	1.00	1.00	1.00
External corpus	30	1.00	1.00	1.00
LLM generation	20	0.50	0.65	0.65
All off-topic	127	0.91	0.93	0.92

TABLE 4.9: Detection rates by off-topic construction method, for the three best configurations in Interpreted run

Four out of the five construction methods are detected well, task mismatch and external corpus responses are caught without surprise, same for manually created responses, except for LLM-as-a-judge under interpreted run.

However, LLM-generated responses are exception, detected at 0.50 to 0.65, depending on configuration and run. They account for 20 of the 127 off-topic responses, but they drive the large majority of all missed.

They have similar pattern, answering on the picture sequence, a child, an old man, books, a room, fruit, without narrating anything truly happens in it:

audio 2026030005 —for me telling a story from pictures is about connecting ideas rather than explaining everything clearly a kid an old man and books can represent different stages of life or learning sometimes fruit or other background details make the scene feel more natural even if they're not the main focus of the story

audio 2026030015 —walking through an old study filled with the smell of paper and wood is one of my favorite experiences there is something so peaceful about a room where a little boy can sit and focus on a story without any interruptions from the outside world i remember my own grandfather used to spend hours organizing his collection and he was always very protective of his favorite volumes making sure they were never misplaced or used for anything other than reading

These responses are topically saturated but content empty, this is the failure predicted for the Semantic Similarity in Section 3.1.2, and LLM-as-a-judge does not fix this either, it rates several of them 8 out of 10. Since this weakness shows up in both prompt variants, all three leading configurations, and both individual approaches, it looks systematic rather than random.

First, these responses were built to target exactly this weakness and their difficulty is by design Section 3.3.4. Second, the authentic off-topic responses are detected less than most constructed ones in the interpreted run, but not in the neutral run, where two of three configurations catch all six. However, with only six such responses is too small samples to say anything.

Chapter 5

Discussion

In this chapter, the pros and cons of the three approaches will be discussed (Section 5.1). Then we will talk about the current achievement in SELF PO, and what issues we are facing (Section 5.2). Lastly, how and what we can resolve in the future (Section 5.3).

5.1 Advantages and Limitations of Each Approach

LLM likelihood

Its advantage requires no training data, no annotation and no fine-tuning, and we can use a pretrained model directly and runs on CPU.

However, this is the only approach failed on every metric and under both prompt variants, the reason is obviously in the score distributions, about a third of off-topic responses fall inside the interquartile range of the on-topic responses Section 4.1, no suitable decision threshold can separate them.

LLM likelihood measures how predictable a string of words is following a context prompt, and seems heavily impacted by the probability of the text in general like word frequency and grammaticality. So this approach does not seem appropriate for our off-topic detection task (Section 3.1.1).

Semantic Similarity

The advantages of Semantic Similarity performs well overall, low computational cost, and the ability to capture meaning beyond exact word overlap. This is important for a task where students retell the same story in diverse wording, especially for dealing with completely unrelated responses (Section 4.4).

On the other hand, its limitations follow from measuring relevance as closeness to the prompt text. Partially relevant responses are not clearly separated from on-topic ones. Responses that reuse the same prompt vocabulary without answering the task asked for receive high similarity score.

LLM-as-a-Judge

LLM-as-a-Judge made the model's assumption visible, and a warning flag that can explain its own reasoning is easier to justify to teachers and easier to verify when wrong, and the reproducibility of output the same result by employing greedy decoding.

Its limitations are cost and very sensitive. It is substantially larger than the other two approaches and requires GPU inference. sometimes it is stricter than the task expect. And its sensitivity to the wording of the judging instruction and the image description, like an assumption about the child's gender that appears in neither.

The combination

Semantic Similarity and LLM-as-a-Judge’s combination was the best and most stable configuration tested. But this complementarity only runs one way. LLM-as-a-Judge catches question copy, also responses that use the target key words telling a different story. But when LLM-as-a-Judge wrongly rated an off-topic response, Semantic Similarity scored it inside the on-topic range too, so Semantic Similarity missed it as well (Section 4.4). Both approaches fail on the same responses, which limits this combination can achieve.

Adding LLM likelihood feature did not help at all. In other combinations it made no real difference and lowered F1 in the others, this feature is weak and noisy.

5.2 Challenges and Limitations of the Present Work

The goals set out in in SELF PO

Regarding the five requirements made for off-topic detection in SELF PO, it is worth for examining how each was satisfied:

- Working for the current tasks was addressed only in part. The approaches were evaluated on one task, storytelling. Nothing in this work establishes that the same performance holds for the other nine.
- Working for future unseen tasks was not tested at all. The design is in principle task-independent, since all three approaches take an arbitrary prompt and response and require no per-task training.
- Remaining lightweight was achieved by Semantic Similarity and not by the best configuration. Semantic Similarity alone runs on CPU and reaches an F1 of 0.908; adding LLM-as-a-Judge raises this to 0.930 but requires GPU inference for a component that is one warning flag among several. There is a genuine tension here between accuracy and proportionality, and the results do not resolve it. Semantic Similarity alone remains a defensible choice at a cost of roughly two F1 points.
- Being open source was met in full. All three models are openly licensed, and no component of the evaluation depends on a commercial API.
- To measure agreement with human raters we need responses that teachers labelled responses as off-topic. For the storytelling task, the experts rated 82 recordings, but none of them was off-topic.
The gold labels used in the final dataset are our own annotations, applied to 324 authentic recordings, plus 121 constructed responses. In the two instances where experts rating and our judgement disagree, we decided to keep our own label.

Most of the off-topic responses were created rather than found, 121 of 127 were constructed, and only 6 are authentic recordings annotated as off-topic. The evaluation shows how each approach separates the off-topic instances we built.

The composition of the constructed each method shapes the result. It contains more clearly off-topic responses than ambiguous ones, and the harder cases come mainly from the manually created and from the LLM-generated thematic overlap responses. Section 4.4 presents the LLM-generated responses are detected only half

to two-thirds of the time. Therefore, An evaluation set with different proportions of methods would produce a visibly different aggregate score for the same system.

The study focused on one task with 445 responses. Results may not transfer to the other nine SELF PO English tasks which prompts are structured differently like four are audio-only and two offer optional images.

The decision threshold was fixed at the default of 0.5. Configurations with high ROC-AUC but lower F1 would be expected to improve under a better threshold.

This is a speaking assessment, but every approach operates on the transcript. Any transcription error becomes a serious detection error. WhisperX Bain et al. (2023) provides high transcription quality, but task-taker accents still cause individual words to be mistranscribed, and in a short response a single wrong word can be decisive.

Three responses illustrate this issue, and they divide into two distinct mechanisms.

The first two instances, the mistranscription caused a false alarm:

audio 1758821202 —*the child really wants some spooks so his father gives to him robinson but the child wants bigger spooks so his father is really surprised but the child uses them just to catch some fruits*

The student said *books*, WhisperX produced *spooks*, and rated on-topic by the expert.

audio 1758744672 —*this story exposed two characters the first one is a tree drone and the second one is our dad the tree drone asks for books to her dad to his two her daddies and at the end we learn that the books have served to take high to grab fruits on the tree*

The student said *children*, WhisperX produced *tree drone*, and rated on-topic by me.

The error from third instance caused more issues:

audio 1762783021 —*the dad gave a box in your shield but you use*

The student said *books*; WhisperX transcribed *box*, and rated on-topic by the expert.

In the rating interface, the expert listened to the audio recording and judged as on-topic. However, we annotated this response as off-topic, because the transcription is too short and retains almost no information connecting to the story. The ASR error changed the meaning of response, for this reason what a judgment made from the transcript and a judgment made from the speech no longer the same.

This is a fairness concern problem, not only an accuracy concern. Because the mistranscriptions are driven by accent, the errors fall on students whose pronunciation the ASR model handles less well. Flagging a valid response as off-topic penalises a student for how the system transcribed their speech rather than for what they said, which conflicts with the equity and fairness principles is designed to reach Section 1.2.

5.3 Future Work

- Collect more authentic off-topic data. As the human rating campaign grows, real off-topic responses can supplement or replace the constructed dataset, this is the change that would most improve the validity of the results.
- Agree a definition of off-topic with human raters. The experts flagged no storytelling response as off-topic, while our own annotation of the same material identified a couple. Before the rating process, the teachers need a shared and explicit criterion. A short annotation instance with the rating team by using the ambiguous cases in Section 2.1.1, would establish how much they agree with each other and produce the definition of SELF PO should be.
- Extend beyond the storytelling task. Applying the approaches to the other nine SELF PO tasks would test how the performance holds, moreover, extending to other languages would test whether the multilingual encoder transfers as intended.
- Tune the decision threshold. ROC-AUC exceeds F1 for some configurations, the threshold should be calibrated and optimised for recall, since a missed off-topic response is more serious.
- Add a threshold for question copies. A question copy fails by being too similar to the prompt, so a two-sided threshold on cosine similarity should capture it, a response whose similarity exceeds an upper bound would be flagged as a probable question copy. A single Semantic Similarity feature would then cover unrelated responses, generic responses and question copies, potentially removing the need for the heavier model.
- Reduce sensitivity to prompt wording. LLM-as-a-Judge's dependence on the image description and its assumption about gender, so testing more descriptions per task, or refining the judging instruction to clarify which details are required and which are not.
- Analyse errors across all configurations. We retained only for the three best configurations. If keeping them for all seven would show more about which approach detects which kind of off-topic response, revealing each method's weaknesses more precisely.
- Test stronger combinations. A non-linear model such as XGBoost or a small neural network, or a model that down weights the unhelpful LLM likelihood feature, might exploit the complementarity between Semantic Similarity and LLM-as-a-Judge better than a linear combiner but any of these models would require a larger dataset to remain reliable, since higher capacity models overfit on limited data.

Chapter 6

Conclusion

This thesis addressed off-topic detection for the SELF Speaking assessment module of English. It aimed to identify spoken responses that do not answer the task, so that they can be flagged before scoring.

Off-topic detection was framed as a prompt–response relevance estimation task and implemented as warning flag within the existing pipeline Section 1.3.

Since authentic off-topic responses were extremely rare in our expert-rated dataset: only 14 out of 2,100 ratings, and none for our storytelling task, a balanced evaluation set of 445 storytelling responses was constructed using four augmentation methods: manual creation, SELF PO task-answer mismatch, an external corpus of L2 English speech by Japanese L1 speakers, and LLM generation Section 3.3.

Three off-topic detection approaches were compared: LLM likelihood, Semantic Similarity, and LLM-as-a-judge. Each was evaluated individually and in combination by using logistic regression with stratified 10-fold cross-validation (Section 3.2).

LLM likelihood did not separate the two classes usefully. Semantic Similarity and LLM-as-a-judge performed comparably to each other and substantially better than LLM likelihood. The strongest and most stable configuration combined Semantic Similarity with LLM-as-a-judge, reaching an F1 of 0.930 ± 0.029 , with the lowest fold-to-fold variation of any configuration tested.

In addition, this thesis gave us more ideas to think about different realistic questions:

- What do we do concretely with answers identified as off-topic?
- Should an off-topic response be scored or not?
- How an off-topic response impact the global score given to the student and to the teacher?
- How much off-topic answers derive from comprehension issues rather than actual production problems?

These questions are still ongoing discussions in the SELF PO Project.

Appendix A

LLM-as-a-Judge

Both versions of the instruction given to LLM-as-a-judge are below. The minimal version was used only during initial testing; all results reported in this thesis use the detailed version.

Minimal Instruction

Prompt:

{prompt}

Response:

{response}

Is the Response relevant to the Prompt?

Respond ONLY in the following JSON format:

{"score": <integer 1-10>, "reason": "<short phrase>"}

Do not include any extra text.

Detailed Instruction

You are an English teacher evaluating responses in a second-language English speaking assessment.

Task:

Evaluate how relevant the student's response is to the prompt.

Important:

Ignore grammar mistakes, pronunciation errors, and fluency issues.

Focus only on whether the meaning of the response is related to the prompt.

Scoring (1-10):

10 = fully answers the prompt and directly addresses the topic

8-9 = clearly related to the prompt but partially incomplete

6-7 = somewhat related but lacks clear connection or detail

4-5 = weakly related or only marginally connected

2-3 = mostly unrelated with minimal topical overlap

1 = completely off-topic or meaningless

Examples of OFF-TOPIC responses:

- discussing a completely different topic
- giving irrelevant information unrelated to the prompt

- repeating the prompt without answering
- producing filler or meaningless content

Prompt:
{prompt}

Response:
{response}

Respond ONLY in the following JSON format:

```
{"score": <integer 1-10>, "reason": "<short phrase>"}
```

Do not include any extra text.

Bibliography

- Attali, Yigal and Jill Burstein (2006). "Automated essay scoring with e-rater® V. 2". In: *The Journal of Technology, Learning and Assessment* 4.3.
- Bain, Max et al. (2023). "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *Proc. Interspeech 2023*, pp. 4489–4493. DOI: [10.21437/Interspeech.2023-78](https://doi.org/10.21437/Interspeech.2023-78). URL: <http://doi.org>.
- Coulange, Sylvain et al. (2024). "An automated pipeline for preprocessing spontaneous L2 English prosody". In: *The 13th International Seminar on Speech Production (ISSP 2024)*.
- Das, Sourya Dipta, Yash A Vadi, and Kuldeep Yadav (2024). "Transformer-based joint modelling for automatic essay scoring and off-topic detection". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16751–16761.
- Google (2026). *Gemini 3.1 Pro: Large Language Model*. Computer software. URL: <https://ai.google.dev/>.
- Higgins, Derrick, Jill Burstein, and Yigal Attali (2006). "Identifying off-topic student essays without topic-specific training data". In: *Natural Language Engineering* 12.2, pp. 145–159.
- Higgins, Derrick et al. (2004). "Evaluating multiple aspects of coherence in student essays". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 185–192.
- Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: [2310.06825](https://arxiv.org/abs/2310.06825) [cs.CL]. URL: <https://arxiv.org/abs/2310.06825>.
- Lee, Chong Min et al. (2017). "Off-topic spoken response detection using siamese convolutional neural networks". In: *Proc. Interspeech 2017*, pp. 1427–1431.
- Louis, Annie and Derrick Higgins (2010). "Off-topic essay detection using short prompt texts". In: *proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pp. 92–95.
- Malinin, Andrey, Kate Knill, and Mark JF Gales (2017). "A hierarchical attention based model for off-topic spontaneous spoken response detection". In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 397–403.
- Malinin, Andrey et al. (2016). "Off-topic response detection for spontaneous spoken english assessment". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1075–1084.
- Metallinou, Angeliki and Jian Cheng (2014). "Syllable and language model based features for detecting non-scorable tests in spoken language proficiency assessment applications". In: *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 89–98.
- Nebhi, Kamel and György Szaszák (2023). "Automatic assessment of spoken English proficiency based on multimodal and multitask transformers". In: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 769–776.

- OpenAI (2023). "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774*.
- OpenAI (2026). *GPT-5.4 Thinking (Large Language Model)*. Accessed: 2026-08-17. URL: <https://openai.com/index/introducing-gpt-5-4/>.
- Persing, Isaac and Vincent Ng (2014). "Modeling prompt adherence in student essays". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1534–1543.
- Radford, Alec et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.
- Raina, Vatsal, Mark Gales, and Kate Knill (2020). "Complementary systems for off-topic spoken response detection". In: *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pp. 41–51.
- Rei, Marek (2017). "Detecting off-topic responses to visual prompts". In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 188–197.
- Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: <http://arxiv.org/abs/1908.10084>.
- Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A vector space model for automatic indexing". In: *Communications of the ACM* 18.11, pp. 613–620.
- Sylvain Coulange Tsuneo Kato, Takayuki Konishi Mariko Sugahara (2026). *Spontaneous Dialogues in L2 English by Japanese-L1 Speakers*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. URL: https://doi.org/10.82270/cles-jp_corpus/v2.
- Wang, Xinhao et al. (2019). "Automatic detection of off-topic spoken responses using very deep convolutional neural networks". In: *Proc. Interspeech 2019*, pp. 4200–4204.
- Yoon, Su-Youn and Shasha Xie (2014). "Similarity-based non-scorable response detection for automated speech scoring". In: *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 116–123.
- Yoon, Su-Youn et al. (2017). "Off-topic spoken response detection with word embeddings". In: *Proc. Interspeech 2017*, pp. 2754–2758.
- Yoon, Su-Youn et al. (2018). "Atypical inputs in educational applications". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 60–67.
- Zechner, Klaus and Keelan Evanini (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.