Confabulation in a Caption-then-Generate Pipeline: Evaluating Vision–Language Generation of Adolescent Health Information

Abigail Naa Amankwaa Abeo[1,3] **Steven Smith**[1,3] **Hannah Goss**[1,3] **Michael Scriney**[2,3]

[1]School of Health and Human Performance, [2]School of Computing
[3]Dublin City University, Ireland
abigail.abeo2@mail.dcu.ie, {steven.smith, hannah.goss, michael.scriney}@dcu.ie

Abstract

Vision-language models make it newly possible to turn a photograph into written health information, but producing text an adolescent can act on is harder than producing text that is merely fluent. We evaluate an off-the-shelf pipeline for this task, in which Florence-2-Large captions an image and Llama 3 $8B$ Instruct generates a health message from that caption. Applying it to images taken by adolescents during a school Photovoice session, a participatory method in which young people photograph their own surroundings, we then collect ratings from six health-literacy evaluators on clarity, relevance, and helpfulness. Comparing each narrative with its caption, we find that many assert content the caption never contained, a behaviour we call confabulation, and that its severity tracks how unhelpful evaluators judged the output to be. Judging only the outputs, the evaluators independently identified three recurring gaps, which map onto three recognised challenges in natural language generation. Our findings show why fluent output is not the same as usable output, and why domain-expert evaluation is needed to tell them apart.

1 Introduction

Many of the health habits people maintain as adults are formed during adolescence, and a young person's ability to find, understand, and act on health information shapes how those habits develop (Sørensen et al., 2012; Sawyer et al., 2012). Low health literacy is common at this age (Sarhan et al., 2023; Caldwell and Melton, 2020; Fleary et al., 2018), and the technology being built to deliver health information is seldom designed with adolescents in mind, with reviews reporting usability problems and poor alignment with their developmental needs (Reen et al., 2019; Xie et al., 2021).

We ask a narrow version of a broad question: *can an off-the-shelf vision–language pipeline take a photograph an adolescent captured and produce a short, usable piece of health information from it?* Producing text on which an adolescent can act is a stricter requirement than producing fluent text (Xie et al., 2021; Reen et al., 2019). An output can be well-formed and factually correct yet still not serve its audience by being too generic to use, pitched for an adult, or detached from what the photograph shows. The task is therefore one of pragmatic, audience-specific generation, not description alone.

The system we evaluate runs in two stages. Florence-2-Large model (Xiao et al., 2024) generates a caption from the image, and Llama 3 8B Instruct model (Grattafiori et al., 2024) turns that caption into a health message for an adolescent. One property of this setup shapes everything that follows: the language model never sees the photograph. It works from the caption, so anything the caption omits is gone before generation begins. As we show, this constraint matters. Given a thin caption, the model does not hold back but supplies plausible health content the caption never contained.

We ran the pipeline on images from a school Photovoice (Wang and Burris, 1997) session with students aged 12 to 16, and had six health-literacy evaluators rate every narrative/output for clarity, relevance, and helpfulness. The outputs were rated clear far more often than relevant or helpful.

To understand where the weaker outputs came from, we then compared each narrative with the caption it was derived from. We observed that many narratives introduced content not present in the original captions, ranging from specific objects to entire health themes. In line with the NLP literature, this behaviour is commonly referred to as hallucination; however, we adopt the term confabulation (Smith et al., 2023; Sui et al., 2024), describing the generation of plausible yet unsupported content that extends beyond the information available in the input. This confabulation was not incidental: the outputs that departed most from

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1407–1419
October 17-21, 2026 ©2026 Association for Computational Linguistics

their captions were also the ones the evaluators rated least helpful. Based solely on the generated outputs, the evaluators identified three recurring gaps, which map onto three established challenges in natural language generation: error propagation, content selection, and pragmatic generation.

Our contributions are:

1. While prior work has explored AI-generated health content (Li et al., 2025; Huang et al., 2024; Stamatakis et al., 2025) and adolescent health literacy separately (Fleary et al., 2018; Caldwell and Melton, 2020), the generation of health information for adolescents from self-captured photographs remains, to our knowledge, unexplored. We apply a caption-then-generate pipeline (Florence-2-Large followed by Llama 3 8B Instruct) to this task setting.

2. We evaluate the outputs through expert human assessment using three functional criteria, complemented by analysis of inter-rater agreement.

3. We introduce a coding framework for confabulation in caption–narrative pairs and show that confabulation severity is associated with evaluators' ratings of the same outputs.

2 Related Work

2.1 Health literacy and adolescents

Health literacy is commonly defined as the ability to access, understand, appraise, and apply health information to make informed decisions (Sørensen et al., 2012). Adolescence is a formative period for developing health literacy: physical, cognitive, and social changes during these years shape health behaviours that can persist into adulthood (Sawyer et al., 2012; Bröder et al., 2017). Yet adolescent health literacy levels remain suboptimal across different settings (Fleary et al., 2018; Caldwell and Melton, 2020; Sarhan et al., 2023), which motivates work on communicating health information in ways that are accessible and age-appropriate. A recent systematic review of artificial intelligence in health literacy found a small but growing body of work applying AI-driven approaches, including language technologies to this problem, while noting that rigorous evaluation of their effectiveness remains limited (Abeo et al., 2025). Our study builds on this distinction, which recurs throughout the literature: health information can be readable without being *usable*, that is, without supporting the specific decisions and actions a person needs to take (Abeo et al., 2025). This functional requirement, rather than readability alone, motivates our evaluation framework.

2.2 Photovoice and the need to scale interpretation

Photovoice is a participatory method in which people photograph aspects of their own environment and experience, producing visual material grounded in their lived perspective (Wang and Burris, 1997; Wang, 2006). It has been used widely in health and education research to foreground participant voice (Baum et al., 2006; Wang and Pies, 2004), and is well suited to adolescent health, where young people's own framing of what matters is often missed by adults (Goss et al., 2021). Its strength, however, is also its bottleneck: photographs are easy to collect but labour-intensive to interpret, requiring facilitation and manual analysis that are difficult to scale (Wang and Burris, 1997).

A systematic review of youth photovoice bears this out: projects are typically small (a mean of ten participants), interpretation is carried out by researchers rather than the young people themselves, and when researchers interpret photographs, they may risk misreading the image and narrative (Fountain et al., 2021), precisely the interpretive step our system targets. A vision-language pipeline might ease this bottleneck by surfacing interpretable, health-relevant text from images that would otherwise require expert reading, complementing rather than replacing participant-led interpretation; this is the setting our system is applied to.

2.3 Vision-language and language models for health and educational content

Vision-language models (VLMs) and large language models (LLMs) have been applied to generating educational and health-related content (Stamatakis et al., 2025; Huang et al., 2024; Lin et al., 2025). Most such models are trained on general-purpose data rather than tuned to a specific audience, and prior work on health-related generation has documented recurring problems: overly complex language, weak adaptation to context, and susceptibility to misinformation (Abd-Alrazaq et al., 2023; Han et al., 2025; Nikandrou et al., 2025). A particularly relevant cautionary finding comes from Li et al. (2025), who evaluated LLMs on a nutrition

task and found that outputs which appear adequate under intrinsic measures can still fail in real-world use - a gap between apparent and actual usefulness that motivates evaluation by domain experts rather than automatic metrics alone. Our work adopts that stance.

2.4 Three generation challenges

A recurring risk when generating from language models is the production of fluent, confident text that the input does not support. This is commonly termed hallucination, but we follow the argument that confabulation is the more accurate description: "hallucination" implies a perceptual error, the model seeing something that is not there, whereas confabulation denotes the generation of plausible narrative detail to fill gaps in the available information (Smith et al., 2023; Sui et al., 2024). This fits our setting, where the language model, given a sparse caption, supplies health content the caption never contained rather than leaving the gap unfilled.

Rather than treat our system's gaps as incidental bugs, we frame them as instances of three established challenges in natural language generation. This framing is what our expert findings (Section 4.5) later map onto.

Error propagation in cascaded pipelines. Modular pipelines, in which each stage consumes the output of the previous one, are prone to error propagation: an inaccuracy introduced early is inherited and amplified downstream (Finkel et al., 2006; Caselli et al., 2015; Vinyals et al., 2015). A related setting is cascaded speech recognition and language understanding, where transcription errors propagate into downstream tasks and can degrade performance even when the errors themselves appear minor (Shapira et al., 2025). Our caption stage plays a similar role: the language model does not see the image, and inherits whatever the caption omits or distorts.

Content selection. Deciding *what* information to convey is a core task of the classical NLG pipeline, known as content determination or content selection (Gatt and Krahmer, 2018). In a multimodal setting this amounts to identifying the salient elements of an image that a faithful, purpose-relevant text should be built around. Recent work makes this concrete: Hashem et al. (2024) generate text from multimodal data with explicit attention to faithfulness and saliency, using a critic that flags salient features the output has missed, and Song et al. (2024) address visual grounding in a modular vision-to-text model. Faithfulness has also become an evaluation target in its own right in the vision-language setting: Jing et al. (2024) propose FaithScore, which decomposes a generated description into atomic claims and verifies each against the image, giving a fine-grained measure of hallucination in large vision-language models. When the system overlooks the health-relevant element of a photograph, this becomes a content-selection problem.

Pragmatic, audience-aware generation. Finally, text that is fluent and on-topic can still fail to perform its communicative function for a particular audience. Generating text that is appropriate and useful for a specific audience is a pragmatic problem, not merely a fluency one, and audience-tailored objectives are difficult to define without audience feedback (Moorjani et al., 2022). For health information aimed at adolescents, this is the difference between generic advice and guidance a young person can act on, the functional requirement introduced in Section 2.1. In practice, an off-the-shelf model is steered towards such an audience and register through in-context prompting: a system instruction together with a small number of demonstration examples, a paradigm introduced by Brown et al. (2020) and refined through instruction tuning (Ouyang et al., 2022).

2.5 Human and expert evaluation of generated text

Because automatic metrics can miss exactly these functional failures, human evaluation is central to our approach. Best-practice guidance for human evaluation of generated text stresses clearly defined criteria and transparent reporting of agreement (Van Der Lee et al., 2019; Van der Lee et al., 2021). We evaluate along three such criteria; clarity, relevance, and helpfulness; drawing on prior operationalisations of these dimensions (Fu et al., 2024; Zhu et al., 2024). Our analysis, which identifies where and how each output departs from its caption, follows a growing body of work on error annotation for generated text, which catalogues inaccuracies and unsupported content rather than reducing quality to a single score (Huidrom and Belz, 2022). This also connects to a broader shift towards graded rather than binary faithfulness eval-

uation: Jing et al. (2025) quantify hallucination on a five-point scale instead of a binary present/absent judgement, in the same spirit in which our four-level severity scheme treats confabulation as a matter of degree. Where prior human-evaluation work typically recruits crowdworkers, our evaluators are domain experts whose judgements reflect the intended application, an emphasis that connects to the real-world evaluation argument of Li et al. (2025).

3 Methodology

3.1 Task and System

We study the task of generating a short, audience-appropriate piece of health information from a single photograph taken by an adolescent. This is a challenge of pragmatic, audience-specific generation (Section 2.4): the output must not only be fluent and accurate but also function for an adolescent, connecting what the photograph shows to health-relevant guidance they can act on.

The system is a two-stage pipeline (Figure 1) assembled from off-the-shelf components, used as released with no fine-tuning; only the prompts are customised. In the first stage, Florence-2-Large (Xiao et al., 2024) produces a caption for the image, using its `DETAILED_CAPTION` mode. In the second stage, Llama 3 8B Instruct (Grattafiori et al., 2024) generates the health narrative from that caption. The language model receives a fixed input comprising a system instruction that defines its role as a youth health-literacy educator, a single few-shot exemplar demonstrating a desired output, and the caption produced from stage one (shown in Appendix A); outputs are constrained to 100 words. The system instruction and exemplar are identical for every image, so the caption is the only image-specific input that varies.

One property of this design governs the analysis that follows: the language model does not receive the image, only the caption. Any information the caption omits is therefore unavailable to the generation stage and cannot be recovered downstream. This makes the caption a strict bottleneck, and because the instruction and exemplar are held constant, it is also the sole locus of image-specific variation in the system's behaviour. Appendix B shows two examples of the full pipeline applied to researcher-captured images, illustrating the flow from image to generated narrative.

3.2 Dataset

The images were collected during a school-based Photovoice session (Wang and Burris, 1997) held between February and March 2025. Photovoice is a participatory method in which participants photograph their own environments, producing visual material grounded in lived experience rather than in a researcher's assumptions about what matters to them. This makes it particularly well suited to adolescent health, where the concerns and contexts that shape health behaviour are often best surfaced by young people themselves. In our session, adolescents aged 12–16 photographed aspects of their daily lives relevant to health, spanning food choices, physical activity, mental health and wellbeing, and sleep. The resulting images capture health as adolescents encounter it in practice, and it is this material that our pipeline is applied to.

48 images were processed through the pipeline. During analysis, we found that two images produced an identical narrative/output; thus, both were excluded from all analyses. This leaves 46 images, each with one caption and one generated narrative. Due to our ethical approval regarding participant consent and privacy protections for minors, the study data cannot be publicly shared (see Data Availability 6).

3.3 Evaluation Design

Six evaluators (hereafter also raters) assessed the narratives, with expertise spanning health-literacy research, education, and direct work with adolescents, including qualified teachers - the expertise relevant to the system's intended application. We do not use reference-based metrics such as BLEU and ROUGE, as there is no ground-truth reference to compare against for this task. Human evaluation is therefore central, and prior work shows that even where such metrics apply, apparent adequacy can diverge from real-world usefulness (Li et al., 2025).

Each evaluator independently assessed every narrative on three criteria: clarity (*is the text easy to understand?*), relevance (*does it meaningfully relate to the image's content?*), and helpfulness (*does it support health-related understanding?*), drawing on prior operationalisations of these dimensions (Fu et al., 2024; Zhu et al., 2024). Each criterion was rated on a three-level ordinal scale: *Very*, *Somewhat*, or *Not at all*. The design is fully crossed and complete: all six raters rated every narrative on all three criteria, giving 828 ratings and enabling

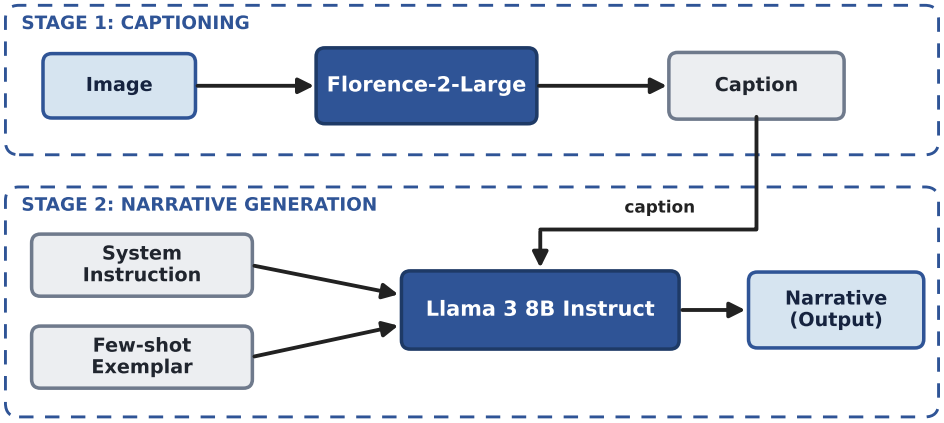


Figure 1: The two-stage pipeline. Florence-2-Large captions the image (Stage 1); the caption, with a fixed system instruction and few-shot exemplar, is passed to Llama 3 8B Instruct, which generates the narrative (Stage 2). The language model does not receive the image, only the caption.

the agreement analysis below. In addition to the structured ratings, four of the six evaluators provided open-ended written comments on how outputs could better serve the adolescents; these comments are the basis for the qualitative analysis in Section 4.5.

3.4 Analysis Methods

Confabulation coding. To locate where and how each narrative departs from its input (caption), we coded the content it introduced that was absent from the caption. Coding output content that is absent from the input, and grading it by severity, follows established error-annotation practice in NLG (Huidrom and Belz, 2022); our four-level scheme adapts this to caption–narrative departure. We use the term *confabulation* rather than hallucination for the reasons set out in Section 2.4: this is gap-filling invention to cover an underspecified input, not perceptual error (Smith et al., 2023; Sui et al., 2024). Confabulation more precisely names an output that supplies plausible detail to fill an underspecified input without any indication that the detail is invented.

We assigned each narrative one of four severity levels according to how far it departed from the caption: *None* (the narrative stays within the caption), *Minor* (small drift or generic additions), *Moderate* (a clear departure, adding content the caption did not contain), and *Major* (the narrative is substantially built on content absent from the caption). The Invented Content took two forms, concrete objects or settings, and inferred health themes, which we treated as equivalent when assigning severity. A theme was counted as invented only where the caption gave no reasonable basis for it; themes that could be reasonably inferred from the caption's content, such as physical activity from a description of people playing sport, were not treated as confabulation. Coding was performed by the first author. Critically, it was carried out on the caption–narrative pairs alone, blind to the evaluators' ratings, so that any observed relationship between confabulation and rating is not an artefact of the coding procedure. We did not assess inter-annotator agreement.

Inter-rater agreement. We quantified agreement among the six raters using Krippendorff's alpha with an ordinal difference function, computed per criterion. The ordinal form is appropriate because the rating scale is ordered: it treats a Very/Not at all disagreement as more severe than a Very/Somewhat one, information that a nominal measure discards.

Association between confabulation and ratings. We then asked whether narratives with more severe confabulation received lower ratings. Because the rating scale is ordered but not evenly spaced, we used only methods that rely on the order of the ratings, not their spacing. First, for each criterion and severity level, we report how the ratings are distributed across Very, Somewhat, and Not at all. Second, we report Kendall's τ_b between severity level and rating for each criterion, a measure of

how consistently the two move together, designed for ordered categories with many tied values. Because this test treats the six ratings of each narrative as independent observations, it does not model the clustering of ratings within images; we therefore treat the per-image rating distributions as the primary descriptive evidence and τ_b as a supporting measure.

4 Results

4.1 Rating Outcomes

Across the 46 narratives, the system's outputs were rated most highly for clarity and less so for relevance and helpfulness. The proportion of ratings of Very was 68.8% for clarity, 58.7% for relevance, and 56.9% for helpfulness (Figure 2). The gap is consistent: the system produces narratives that raters find easy to understand more clearly than it produces narratives they find relevant to the image or helpful for health understanding.

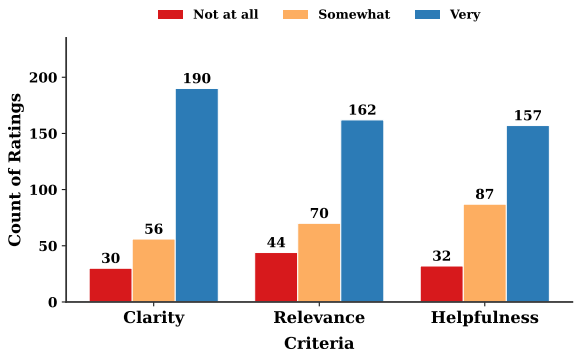


Figure 2: Distribution of ratings across the three criteria over the 46 narratives ($n = 276$ ratings per criterion).

4.2 Inter-rater agreement

We measured agreement among the six raters using Krippendorff's α (ordinal): 0.26 for clarity, 0.64 for relevance, and 0.39 for helpfulness. The low clarity figure reflects the skewed distribution of clarity ratings rather than genuine disorder: with 190 of 276 ratings *Very*, there is little variance for a chance-corrected measure to capture, so α is deflated even though raw agreement is high..

4.3 Confabulation analysis

Comparing each narrative with the caption it was generated from shows that outputs frequently asserted content the caption did not contain. The captions themselves were brief, with a median of 29 words, and 45 of the 46 did not begin as a complete sentence, indicating that the caption text had been clipped. Severity asks a single question of each narrative: did it stay within what the caption gave it, or did it add things the caption never mentioned? On this basis we assigned each narrative a severity level; the distribution is shown in Table 1.

Severity	Description	Img.	%
None	The narrative stays within the caption; it adds nothing not already there	17	37
Minor	The narrative mostly stays within the caption, with small generic additions	11	24
Moderate	The narrative clearly adds content the caption did not contain	11	24
Major	The narrative is substantially built on content absent from the caption	7	15

Table 1: Confabulation severity across the 46 narratives.

Severity reflects the extent of the departure from the caption, not the type of content invented (see Section 3.4). A representative case is an image of a mural. The caption described "a variety of shapes and designs," a building, benches, plants, and trees, with no mention of any appliance. The narrative nonetheless stated that "having an air conditioner nearby adds a practical touch," and framed the mural as "promoting mental wellness and relaxation," a concrete object and a health theme, neither of which the caption contained. An evaluator, seeing the photograph and the narrative together, described this as a misidentification. The intermediate caption shows the mechanism more precisely: it recorded no appliance at all, so the caption stage omitted whatever was in the scene, and the generation stage then invented the air conditioner to fill the gap. What appears from the endpoints as a single misreading is in fact two errors, one at each stage, a distinction visible only because the caption sits between the image and the output.

4.4 Confabulation predicts expert ratings

We coded confabulation severity from the caption–narrative pairs alone, without looking at the ratings. Even so, it lines up closely with how the evaluators rated the same outputs: the more a narrative departed from its caption, the worse it was rated. As severity rises, *Very* ratings become less common and *Not at all* ratings more common (Figure 3). For helpfulness, the share of *Very* ratings

fell from 75% for narratives coded *None* to 14% for those coded *Major*, while *Not at all* ratings rose from 3% to 48%. This downward trend was supported by Kendall's τ_b, a rank-based measure of association: -0.39 for helpfulness, -0.51 for relevance and -0.32 for clarity (all $p < .001$), where a negative value means ratings fall as severity increases.

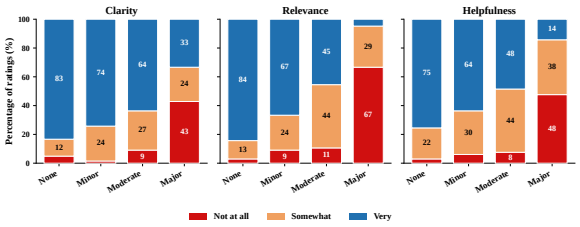


Figure 3: Rating distribution across confabulation severity levels, for each criterion. As severity rises, *Very* ratings fall and *Not at all* ratings rise.

First, the coding was done blind to the ratings, so the agreement between the two is not built in; it is two separate judgements pointing the same way. Second, the link is not merely a matter of definitions. Relevance overlaps somewhat with confabulation, since both concern whether the output matches the image, and it shows the strongest association ($\tau_b = -0.51$). Helpfulness, however, does not measure image-match, yet it still falls steadily as confabulation worsens. The outputs we coded as confabulating are, independently, the ones evaluators found least helpful.

4.5 Expert feedback: three recurring gaps

Beyond the numerical ratings, four of the six evaluators provided open-ended comments on how the outputs could better serve adolescents. This qualitative feedback shows how the system's choices of what to describe, and how to frame it, shape the experience for a young person. The comments cluster into three recurring gaps, each of which corresponds to one of the three generation challenges set out in Section 2.4.

Misinterpretation of visual elements. The most consistent issue was the output describing things incorrectly, which undermined both the accuracy and the usefulness of the narrative. In some cases the system mislabelled a scene: a school hallway with an extracurricular display board was rendered as a "quiet study space," shifting the output away from the activity the image actually depicted. In others, it overlooked elements that would have anchored the health message: in a park scene containing two goalposts, the description centred on the natural setting and missed the goalposts entirely, losing a clear cue for physical activity that the evaluators themselves picked out. In pipeline terms, these are instances of error propagation - a misreading introduced at the caption stage, or an invention at the generation stage, that the final output then presents to the adolescent as fact, with no opportunity for correction.

Missed contextual focus. Even when the individual elements were read correctly, the system sometimes built the narrative around the wrong one. In an image of five posters about healthy foods mounted on a glass frame, the output dwelt on the windows and the leisurely outdoor ambience, when - as the evaluators noted - the posters themselves were the health-relevant content and should have been the focus. The result is a narrative that is not inaccurate so much as beside the point: fluent, but aimed at the wrong part of the image. This is a content-selection challenge, the system's inability to identify which element of a scene a useful health message should be built around, and it is the gap most directly tied to the lower relevance ratings reported above.

Insufficient actionability. Finally, even when the output was accurate and on topic, it often stopped at broad encouragement rather than offering something a young person could act on. Responding to an image of a sports and fitness centre, for instance, the system produced only general remarks about the value of exercise. The evaluators expected more concrete narratives - for example, "How do you like to take part in physical activity?" or "Can you think of ways to make exercise more accessible for others in your community?" - which not only focus an adolescent's thinking but nudge them towards a health-literate mindset and, ultimately, action. This is a pragmatic challenge: the text is fluent, and yet it does not fully perform its intended function for its audience, which is precisely the difference between readable health information and usable health information as discussed in Section 2.

5 Discussion

Why the outputs confabulate. Because the language model never sees the image, the captions

(short and clipped, as noted in Section 4.3) are the whole of what it has to work with, and confabulation is what fills the gap: given a thin description, the model does not abstain or hedge but supplies plausible health content to match the confident, detailed style it has been asked for. The captioning configuration compounds this. Florence-2's DETAILED_CAPTION mode describes a scene but does not read text within it, and many of our images carried the health message as text - on posters, labels, and signs. This is exactly the content the captions tended to drop, so the element most relevant to health literacy was frequently the element the generation stage never received.

The role of the prompt. Our system was steered by a fixed prompt: a single, richly worded exemplar demonstrating a confident, detailed health message (Appendix A), applied unchanged to every image regardless of how much its caption contained. We cannot isolate the prompt's contribution without an ablation, which our study does not support, and we therefore offer this as a hypothesis rather than a result. A prompt optimised for rich, assured output, applied uniformly across inputs of varying informativeness, may itself encourage confabulation, because it demonstrates only one mode of response and never how to behave when the input is thin. This connects confabulation to a design choice rather than to model capability alone, and it points to conditioning generation on the informativeness of the input as a direction worth testing.

Implications for building such systems. The three gaps (Section 4.5) each point to a different lever. The misinterpretation gap is one of grounding and error propagation, calling for captioning suited to the task rather than general scene description that misreads or omits what matters. The missed-context gap is one of content selection (Gatt and Krahmer, 2018): recent multimodal work attending explicitly to faithfulness and to salient image regions (Hashem et al., 2024) suggests the effort should go into selecting the health-relevant element, not just describing the scene alone. The actionability gap is one of pragmatics: text that works for an adolescent, rather than merely fluent text, is not guaranteed by a capable language model and may require audience feedback even to specify (Moorjani et al., 2022). The common thread is that fluency is not where the difficulty lies; relevance and usefulness are, and the pipeline delivers neither reliably. This matters beyond our system: off-the-shelf vision–language pipelines are increasingly within reach of evaluators without machine-learning expertise, and in health and accessibility settings the gap between readable and usable information is the point. That the outputs departing most from their inputs were also the least helpful echoes findings that intrinsic adequacy can diverge from real-world usefulness (Li et al., 2025), and is why surfacing these challenges needs evaluation by domain experts rather than fluency or automatic metrics alone (Van Der Lee et al., 2019).

6 Conclusion

We evaluated an off-the-shelf vision–language pipeline for generating health information from photographs taken by adolescents, pairing Florence-2-Large with a Llama 3 8B Instruct model and assessing the outputs with six health-literacy evaluators. The outputs were often clear but less often relevant or helpful, and our confabulation analysis showed that those departing most from their captions were the ones evaluators rated least helpful. Judging only the final narratives, the evaluators independently surfaced three recurring gaps, misinterpretation of visual elements, missed contextual focus, and insufficient actionability, which correspond to three recognised problems in natural language generation: error propagation, content selection, and pragmatic generation. For an audience such as adolescents, the central lesson is that fluent output is not the same as usable output, and that evaluation by domain experts can reveal the difference.

Limitations

Our study has several limitations. First, the confabulation coding was done by a single annotator (the first author), so inter-annotator agreement could not be assessed. It was, however, blind to the evaluators' ratings, guarding against circularity. Because the boundary between a reasonable inference and an invented one involves judgement a second coder might draw differently, future work should use multiple annotators and report the reliability of the severity scheme. Second, our findings rest on 46 images from one school-based Photovoice session, a single setting and group; the pipeline should be evaluated across more images, schools, and pop-

ulations. Third, because the system used a single fixed prompt, we cannot separate the contribution of the prompt from that of the model, so our account of why confabulation occurs is a hypothesis consistent with the evidence rather than a demonstrated mechanism; testing it directly, for instance by varying the prompt with the informativeness of the caption, is an important next step. Fourth, the evaluation reflects six domain evaluators rather than adolescents themselves. Expert assessment is a reasonable proxy, but the ultimate audience may weigh these criteria differently, and involving them is necessary to confirm the outputs succeed. Relatedly, our tests treat the six ratings of each output as independent; because they are clustered within images, the reported significance levels are likely optimistic, which is why the per-image distributions are our primary evidence. Finally, our results describe a single configuration, one vision-language model paired with one instruction-tuned language model, and different captioning methods, prompts, or pairings may distribute these results differently.

Data Availability

The dataset comprises photographs captured by adolescent participants during a school-based Photovoice session, together with the captions and narratives derived from them, and the evaluators' ratings of those narratives. These images were captured by minors, and the terms of our ethics approval and participant consent do not permit their release. As such, the images and the derived data (captions, narratives, and ratings) are not made publicly available. The example images shown in Appendix B were captured by the researchers and are not participant data. Our code is available at `https://github.com/abigailabeo/INLG_Paper_Code`.

Ethics Statement

Ethical approval for this study was granted by the Dublin City University research ethics committee (DCUREC/2023/128, approval date 15 June 2023).

Acknowledgments

The work has been supported by Taighde Eireann Research Ireland ML Labs CRT under grant number 18/CRT/6183, the Insight Research Ireland Centre for Data Analytics under grant number 12/RC/2289 P2 and financial support of the EU Commission Recovery and Resilience Facility under the Science Foundation Ireland OurTech Challenge Grant Number 22/NCF/OT/11312. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR medical education*, 9(1):e48291.

Abigail Naa Amankwaa Abeo, Sophie Armstrong, Michael Scriney, and Hannah Goss. 2025. Artificial intelligence techniques and health literacy: a systematic review. *Mayo Clinic Proceedings: Digital Health*, page 100269.

Fran Baum, Colin MacDougall, and Danielle Smith. 2006. Participatory action research. *Journal of epidemiology and community health*, 60(10):854.

Janine Bröder, Orkan Okan, Ullrich Bauer, Dirk Bruland, Sandra Schlupp, Torsten M Bollweg, Luis Saboga-Nunes, Emma Bond, Kristine Sørensen, Eva-Maria Bitzer, et al. 2017. Health literacy in childhood and youth: a systematic review of definitions and models. *BMC public health*, 17(1):361.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elizabeth Perry Caldwell and Karen Melton. 2020. Health literacy of adolescents. *Journal of Pediatric Nursing*, 55:116–119.

Tommaso Caselli, Piek Vossen, Marieke van Erp, Antske Fokkens, Filip Ilievski, Rubén Izquierdo, Minh Le, Roser Morante, and Marten Postma. 2015. When it's all piling up: investigating error propagation in an nlp pipeline. In *WNACP@ NLDB*.

Jenny Rose Finkel, Christopher D Manning, and Andrew Y Ng. 2006. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626.

Sasha A Fleary, Patrece Joseph, and Jessica E Pappagianopoulos. 2018. Adolescent health literacy and health behaviors: A systematic review. *Journal of adolescence*, 62:116–127.

Sarah Fountain, Rachel Hale, Nicola Spencer, Jinger Morgan, Laura James, and M Kathryn Stewart. 2021. A 10-year systematic review of photovoice projects with youth in the united states. *Health Promotion Practice*, 22(6):767–777.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Hannah R Goss, Clare McDermott, Laura Hickey, Johann Issartel, Sarah Meegan, Janis Morrissey, Celine Murrin, Cameron Peers, Craig Smith, Ailbhe Spillane, et al. 2021. Understanding disadvantaged adolescents' perception of health literacy through a systematic development of peer vignettes. *BMC Public Health*, 21(1):593.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, and Harksoo Kim. 2025. Exploring the impact of instruction-tuning on llm's susceptibility to misinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26711–26731.

Tahsina Hashem, Weiqing Wang, Derry Tanti Wijaya, Mohammed Eunus Ali, and Yuan-Fang Li. 2024. Generating faithful and salient text from multimodal data. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 646–662.

Chieh-Yang Huang, Jing Wei, and Ting-Hao Kenneth Huang. 2024. Generating educational materials with different levels of readability using llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22.

Rudali Huidrom and Anya Belz. 2022. A survey of recent error annotation schemes for automatically generated text. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 383–398, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Xiaonan Jing, Srinivas Billa, and Danny Godbout. 2025. On a scale from 1 to 5: Quantifying hallucination in faithfulness evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7780–7795.

Karen Jia-Hui Li, Simone Balloccu, Ondřej Dušek, and Ehud Reiter. 2025. When llms can't help: Real-world evaluation of llms in nutrition. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 753–779.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song Xiaohui, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37975–37995. PMLR.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2022. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1919–1932, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. Crope: Evaluating in-context adaptation of vision and language models to culture-specific concepts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7917–7936.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Gurpreet Kaur Reen, Linden Muirhead, and Dawn Wendy Langdon. 2019. Usability of health information websites designed for adolescents: systematic review, neurodevelopmental model, and design brief. *Journal of medical Internet research*, 21(4):e11584.

Mohammed BA Sarhan, Rika Fujiya, Junko Kiriya, Zin Wai Htay, Kayono Nakajima, Rie Fuse, Nao Wakabayashi, and Masamine Jimba. 2023. Health literacy among adolescents and young adults in the eastern mediterranean region: a scoping review. *BMJ open*, 13(6):e072787.

Susan M Sawyer, Rima A Afifi, Linda H Bearinger, Sarah-Jayne Blakemore, Bruce Dick, Alex C Ezeh,

and George C Patton. 2012. Adolescence: a foundation for future health. *The lancet*, 379(9826):1630–1640.

Ori Shapira, Shlomo Chazan, and Amir David Nissan Cohen. 2025. Measuring the effect of transcription noise on downstream language understanding tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29978–30004.

Andrew L Smith, Felix Greaves, and Trishan Panch. 2023. Hallucination or confabulation? neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11):e0000388.

Yingjin Song, Denis Paperno, and Albert Gatt. 2024. Context-aware visual storytelling with visual prefix tuning and contrastive learning. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 384–401, Tokyo, Japan. Association for Computational Linguistics.

Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, and (HLS-EU) Consortium Health Literacy Project European. 2012. Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, 12(1):80.

Markos Stamatakis, Joshua Berger, Christian Wartena, Ralph Ewerth, and Anett Hoppe. 2025. Enhancing the learning experience: Using vision-language models to generate questions for educational videos. In *International Conference on Artificial Intelligence in Education*, pages 305–319. Springer.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th international conference on natural language generation*, pages 355–368.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Caroline Wang and Mary Ann Burris. 1997. Photovoice: Concept, methodology, and use for participatory needs assessment. *Health education & behavior*, 24(3):369–387.

Caroline C Wang. 2006. Youth participation in photovoice as a strategy for community change. *Journal of community practice*, 14(1-2):147–161.

Caroline C Wang and Cheri A Pies. 2004. Family, maternal, and child health through photovoice. *Maternal and child health journal*, 8(2):95–102.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Wenxiu Xie, Meng Ji, Yanmeng Liu, Tianyong Hao, and Chi-Yin Chow. 2021. Predicting writing styles of web-based materials for children's health education using the selection of semantic features: machine learning approach. *JMIR medical informatics*, 9(7):e30115.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.

A Full Prompt

The complete system instruction and the few-shot exemplar passed to Llama 3 8B Instruct are shown below.

System Instruction

Instruction:
You are an expert in educating youths on healthy lifestyles and health literacy. Your task is to discuss photographs taken by students, relating them to healthy living as defined by the following health literacy definition:
"*Health literacy is the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions.*"
You will not see the actual photographs but will receive detailed descriptions. Use these descriptions to provide insightful and detailed feedback as if you have seen the photos.
Guidelines:

1. Carefully read and analyze the description before responding.
2. Connect the content to healthy living principles and health literacy.
3. Focus only on health literacy (not aesthetics or fashion).
4. Suggest relevant health-related attire/equipment if applicable.
5. Do not repeat the student's description.
6. Use inclusive, supportive, and encouraging language.
7. Structure feedback clearly and offer actionable advice.
8. Add relevant examples or tips about healthy living.
9. Keep the response self-contained (no follow-ups).
10. Use provided examples for response style.

Constraint: Limit answers to **100 words maximum**.

Few-shot Example Template

Input: Photo Description: A young person reaching for their alarm after oversleeping on a school day.

Output:
Sleep is vital to your wellbeing. It is as important as the air we breathe, the water we drink, and the food we eat. Sleep is necessary for healthy body function and brain activity. Teens need about 8–10 hours of sleep each night to function best, however, most teens do not get enough sleep. Can you guess the percentage of teens that obtain the recommended hours of sleep? Only 15%! While most people are aware that lack of sleep isn't good for you, it's difficult to change.

B Illustrative Pipeline Examples

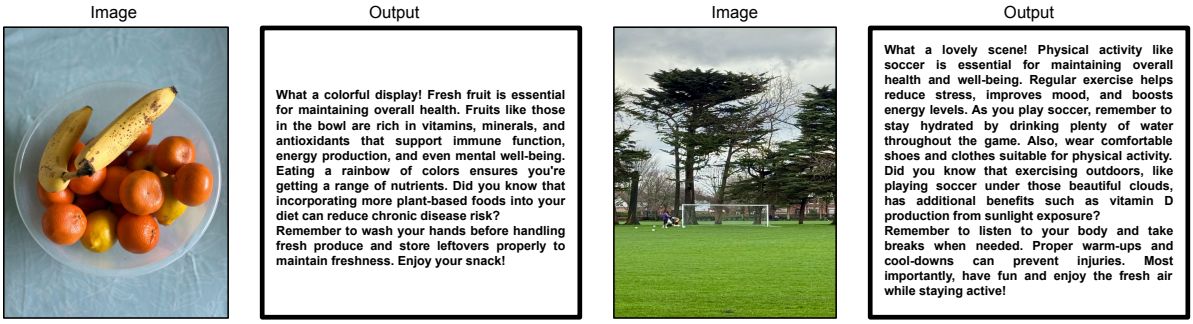


Figure 4: Two researcher-captured, non-identifying example images (food choice; physical activity) processed through the pipeline, with the resulting generated narratives. Included for illustration only and not part of the dataset.