Generating Hotel Highlights from Unstructured Text using LLMs

Srinivas Ramesh Kamath and **Fahime Same** and **Saad Mahamood**
trivago N.V., Düsseldorf, Germany
{srinivas.kamath, fahime.same, saad.mahamood}@trivago.com

Abstract

We describe our implementation and evaluation of the Hotel Highlights system which has been deployed live by trivago. This system leverages a large language model (LLM) to generate a set of highlights from accommodation descriptions and reviews, enabling travellers to quickly understand its unique aspects. In this paper, we discuss our motivation for building this system and the human evaluation we conducted, comparing the generated highlights against the source input to assess the degree of hallucinations and/or contradictions present. Finally, we outline the lessons learned and the improvements needed.

1 Introduction

It is crucial to provide updated and accurate content so that travellers can make informed choices about which accommodation to book. Content such as images, descriptions, reviews, facility and amenity information, and maps helps travellers compare different accommodations to determine their suitability. Given the diversity of content, it is not immediately apparent why a traveller should choose one accommodation over another. While images, descriptions, and reviews can help, they require travellers to extensively analyse and then come up with an assessment before making decisions. This can be challenging for travellers as content styles between accommodations are not uniform. Reviews, for example, can often be terse and written in various styles, with travellers only selectively mentioning aspects from their own perspective. Descriptions, on the other hand, while more objective, can be quite verbose and may also selectively mention aspects from the perspective of the hotelier. Past systems such as the SuRE (Tien et al., 2015) and Hotel Scribe (Mahamood and Zembrzuski, 2019) have focused more on either summarising opinions or describing an accommodation instead of surfacing unique aspects.

To streamline information access for travellers, we developed the Hotel Highlights project. These highlights are concise, one to two sentences summarising an accommodation's unique selling points, derived from traveller reviews and descriptions, allowing travellers to quickly grasp a property's distinctiveness.

To accomplish this, we will discuss the challenges of using LLMs for summarisation (§2). Afterwards, we will explain our system implementation for generating Hotel Highlights (§3). We then describe our human evaluation (§4) and the results obtained (§5). Finally, we will discuss our conclusions from the findings obtained and potential future work (§6).

2 LLMs and Summarisation

Until very recently, fine-tuning pre-trained models, such as BART (Lewis et al., 2020), on domain-specific datasets has been seen as the leading paradigm for text summarisation (Goyal et al., 2022). However, the rise of very large language models (LLMs) and the success of prompting these models have shown an alternative approach with these models being able with only a few demonstrative examples to generate convincing summaries without the need for updating model parameters (Goyal et al., 2022). When evaluated with human evaluators, there seems to be a strong preference for summaries generated by LLMs like GPT-3 (Goyal et al., 2022; Pu et al., 2023). This has led some to declare that the task of summarisation is "almost dead" due to the ability of LLMs to consistently outperform summaries generated by fine-tuned models (Pu et al., 2023) or, in other cases, be on-par with human summarisation (Zhang et al., 2024). However, the reasons for their success is not well understood.

Another area of focus has been trying to understand how faithful a model is to the input it has

Proceedings of the 17th International Natural Language Generation Conference, pages 280–288
September 23–27, 2024.

summarised. A model that hallucinates cannot be considered faithful. Maynez et al. (2020) define two types of hallucinations: *intrinsic* hallucinations, where the model misrepresents facts from the input, and *extrinsic* hallucinations, where the input is ignored and the extraneous text has no relation to the input. For the remainder of the paper, to prevent any confusion with extrinsic and intrinsic evaluation methods, we will use the term "contradiction" to refer to intrinsic hallucination and "hallucination" to refer to extrinsic hallucination.

While automatic metrics, such as ROUGE, are commonly used to evaluate textual similarity, they are inadequate for assessing faithfulness. This inadequacy arises because a high degree of similarity does not necessarily imply faithfulness (Gehrmann et al., 2023). Therefore to evaluate the factual accuracy of generated texts, it is necessary to have a robust human evaluation methodology in place (Thomson and Reiter, 2020).

3 System Implementation

We created a minimum viable system with data selection, generation with LLMs and post processing, illustrated in figure 1.

3.1 Data Selection

For data selection, our focus was on using English accommodation descriptions and reviews from various accommodation types (hotels, resorts, motels, etc.). Descriptions tend to contain a lot of information about different aspects of the accommodation, such as location, amenities, room types, and activities. Therefore, we favoured verbose descriptions over shorter ones. With traveller reviews, recency was of primary importance, as the experiences of a stay can change seasonally and are reflected in what travellers say about it. We also chose reviews to be slightly verbose (with a minimum threshold set at 25 characters in length) to guarantee a sufficient level of detail. Additionally, we considered multiple reviews per accommodation, as traveller experiences can be subjective. This approach aimed to provide a representative and aggregated view of the experiences.

3.2 Generation of Highlights

Figure 1 describes a detailed scheme of our Hotel Highlights system.

We used descriptions and reviews as the input to generate highlights for each accommodation.

Prompt Design: We experimented with zero-shot and one-shot variants. Zero-shot prompting led to less control over the desired format of the output. Therefore, we opted for one-shot prompting as it allowed the output format to be influenced by reference examples. The prompt included a summarisation task, generation criteria, and reference examples with input content and output highlights in the one-shot setting. Copywriters aided in shaping the phrasing of the highlights, providing feedback to ensure brief, third-person titles and descriptions. Due to commercial sensitivity, we cannot share the exact prompt used.

We generated highlights for sample input texts and visually inspected them to check for divergences, fluency, and phrasing.

LLM Selection: We assessed both ChatGPT 3.5 `text-davinci-003` (Brown et al., 2020) and PaLM2 `text-bison` (Anil et al., 2023) models, and compared aspects such as the quality of generation, token limits, and data sharing agreements. For the same prompt and input data, we generated highlights with both models for a sample set of 25 accommodations.

To decide which LLM to use, we designed a human annotation task to rank the highlights using the following rating criteria: *good*, *satisfactory*, *bad*, and *unsure*. Eight internal-company annotators performed the evaluation. Around 75% of the highlights from PaLM2 were ranked between good and satisfactory, compared to 47% from ChatGPT 3.5. Inter-annotator agreement was low (κ=0.208), as some annotators were more conservative in assigning subjective ratings than others.

3.3 Post-Processing

To enable product decisions on which highlights to show to travellers, we included additional metadata after generation. This metadata contained information on the input source (i.e. hotel descriptions or traveller reviews), the sentiment of the highlight, and the category or theme of the highlight.

For sentiment analysis, we used multiple off-the-shelf sentiment classification models (Akbik et al., 2018; Camacho-Collados et al., 2022) to classify sentiment and determined the final sentiment based on a majority consensus among the labels. The initial goal was to classify the sentiment into one of three labels: *positive*, *neutral*, and *negative*. However, based on a sample human evaluation task, we observed that both humans and classification mod-

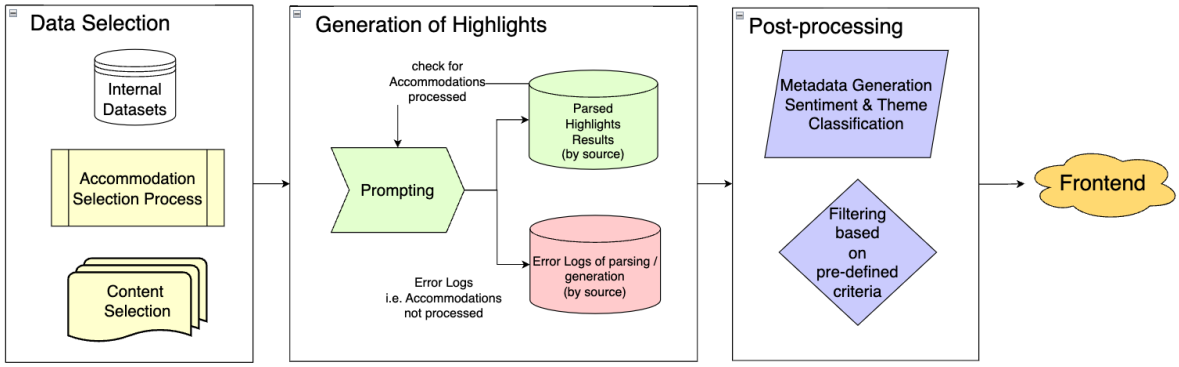

Figure 1: System Process Diagram

els struggled with the nuances between positive and neutral labels. Hence, we decided to use only two labels: *positive* and *negative*.

For theme classification, we devised a rule-based multi-label classification approach based on keyword patterns associated with company-defined categories, complementing the LLMs' ability to pick multiple pertinent data points from the input content to generate highlights. Since the input data contained both *objective* aspects of an accommodation (e.g. facilities and amenities, dining, location, etc.) and *subjective* aspects based on traveller experiences (e.g. staff, perks, experiences, cleanliness, etc.), we formulated classes to identify both types of themes. Additionally, we performed manual quality assurance checks to identify patterns and remove undesirable highlights from a business/traveller perspective.

4 Human Evaluation

We conducted a human evaluation experiment to better understand the quality of the generated highlights. We sampled 40 accommodations by limiting descriptions between 400 and 1000 characters in length. Description length was restricted to minimise annotators' cognitive load while still containing a decent amount of information about the accommodation.

From each accommodation, we selected three highlights, resulting in a total of 120 highlights evaluated in this experiment (40 accommodations * 3 highlights = 120). Figure 2 shows an example of a hotel description, along with highlights with no divergence, hallucination, or contradiction.

4.1 Design

The 40 accommodations were divided into four batches, with each batch containing 10 accommodations and their respective three highlights (30 highlights per batch). Each batch was evaluated by 30 participants, where each participant was shown a hotel description and a highlight (example shown in Appendix A), and asked to specify whether there were any divergences between the two. Participants could decide for a given highlight as a multiple-choice question if there was a *hallucination*, a *contradiction*, *both hallucination and contradiction*, or *no divergence*. For the participants, we defined *hallucination* as 'what is mentioned is nowhere in the input' and *contradiction* as 'what is mentioned contradicts the input'.

Following this, participants rated each highlight for three intrinsic features on a 7-point Likert scale: *clarity* (how clearly does the highlight express the details of the description?), *informativeness* (is the generated highlight informative?), and *grammaticality* (is the highlight grammatically correct?). As an optional step, participants could also suggest alternative highlights.

4.2 Experimental Procedure

The experiment was designed using Google Forms and conducted on Prolific. A validation task was provided to assess the participants' understanding of the task. They were presented with a hotel description and a highlight containing a very clear hallucination. Participants who correctly identified the hallucination received an extra bonus at the end.

Hotel description: Set along a sandy beach, this genteel hotel is 5 km from the Aquarium of Reunion, **and 2 km from both the sandy Plage de l'Hermitage and Eden Garden**. Featuring balconies or terraces, the relaxed rooms offer free Wi-Fi, flat-screen TVs and safes, plus minibars, and tea and coffee-making facilities. Suites add living areas. **Breakfast is served every morning for a surcharge.** Other amenities include 3 restaurants, a cafe and a bar, plus an outdoor pool, direct access to the beach, and meeting and event space. There's also a spa, gardens and a tennis court.

Highlight with no divergence: This accommodation has 3 restaurants.

Highlight with hallucination: Situated along a sandy beach, with **direct access to Plage de l'Hermitage.**
Explanation: There is no explicit mention of direct access in the description and therefore, this is regarded as hallucination.

Highlight with contradiction: Breakfast is served daily in the dining room.
Explanation: According to the description, breakfast is served with a surcharge, but this is not mentioned in the highlight, making it seem free of charge. This creates a contradiction.

Figure 2: Examples of hotel descriptions and generated highlights in different conditions.

5 Results

Out of 119 participants (whole group), 84 answered the validation question with *Hallucination* (henceforth, the success group), 13 with *Contradiction*, 19 with *both*, and 3 with *No Divergence*. In the remainder of this section, results will be reported for the whole group, with references to the success group when there are noticeable differences.

In more than half of the cases (53.22%), participants did not detect any divergence in the highlights. Among the cases marked as divergent, hallucinations were the most common (23.39%), followed by contradictions (13.67%), and lastly both hallucination and contradiction (9.72%). Furthermore, we evaluated the average rating scores for each intrinsic feature across the four batches. The results showed that grammaticality consistently received the highest ratings. Notably, batch 3 received the lowest ratings on all questions, which may suggest differences in the participants or the difficulty of the questions. Detailed per-batch results can be found in Appendix B.

Correlation between Divergence and the Three Intrinsic Ratings: We expect that when participants identify divergences, they will give lower ratings to the highlights, particularly in terms of clarity and informativeness. Therefore, we conducted a correlation analysis using Pearson correlation coefficients to assess the relationship between the divergence scores and the ratings for the three intrinsic features. In this context, divergence is treated as a binary variable: *divergent* (hallucination, contradiction, or both) versus *not divergent*.

The correlation analysis in table 1 confirms this assumption. The presence of divergences is negatively correlated with clarity (Cl), informativeness (In), and grammaticality (Gr), with the strongest negative correlation between divergence and clarity. There is also a very strong positive correlation between clarity and informativeness, indicating consistent evaluations across these questions. Both clarity and informativeness have positive correlations with grammaticality, though the correlation is less strong. All these correlations are statistically significant (p-values < 0.05). Full results can be found in Appendix C.

	Div	Cl	In	Gr
Div	1.00	-0.73	-0.63	-0.29
Cl	-0.73	1.00	0.89	0.58
In	-0.63	0.89	1.00	0.53
Gr	-0.29	0.58	0.53	1.00

Table 1: Correlation analysis between divergence and the three intrinsic ratings. Div, Cl, In, and Gr stand for Divergence, Clarity, Informativeness, and Grammaticality, respectively.

Theme Analysis We want to understand which themes have the most hallucinations and the highest intrinsic ratings. For this analysis, we focus on the following themes: facilities and amenities, location, dining and cuisine, activities, and wellness.

From our analysis, wellness highlights have the highest clarity and informativeness, and the lowest divergence (29.36%). In contrast, location highlights have the highest divergence (44.83%), closely followed by activities highlights (44.63%). Per-theme scores can be found in Appendix D.

Inter-Annotator Agreement: We computed separate Krippendorff's alpha reliability scores for each question type in each batch (n=16), obtaining an averaged score of $\alpha = 0.169$ for multi-class divergence, $\alpha = 0.267$ for binary divergence, $\alpha =$

0.071 for clarity, α = 0.074 for informativeness, and α = 0.003 for grammaticality. These results suggest near-zero inter-annotator agreement for the intrinsic features. However, there is a weak positive agreement for detecting different types of divergences. When considering divergence as a binary feature, agreement increases slightly, implying that people may have difficulty discerning different divergence types. Furthermore, we limited the analysis to those who answered the validation question correctly. We see an increase in their agreement rate for detecting divergence (α = 0.201 for multi-class divergence, α = 0.313 for binary divergence).

6 Conclusion & Future Work

As perceived by annotators, while 53.22% of cases show no divergence, there is still a significant number of hallucinations and contradictions, with the majority coming from the location theme as compared to other objective themes. Given the low inter-annotator agreement, this suggests that even with training, the task of evaluating divergences is difficult. An observation also seen by Zhang et al. (2023) in trying to obtain high agreement with not just crowd workers, but also with experts.

We expected that highlights with divergences would receive lower intrinsic ratings, and this expectation was confirmed in the evaluation. Additionally, the average rating of grammaticallity is relatively higher compared to the other intrinsic qualities, which aligns with the assumption that LLMs have high grammatical correctness.

Future Work: We would like to focus on better understanding the cases where highlights have been judged as containing divergences and how these divergences can be mitigated. Additional improvements planned for the human evaluation include more training for annotators with diverse examples for better calibration and an expanded sense check task for better filtering of annotators. Follow-ups include evaluation tasks around categorising type of divergences, along with an analysis of the suggested highlights written by annotators.

Given the known caveats with human evaluations (Thomson et al., 2023), we also intend to explore the use of LLMs to identify divergences in generated highlights, assessing the feasibility and scalability of this approach as an alternative or complement to human evaluation.

7 Limitations

One of the limitations of this work is that we did not perform a granular annotation of the divergence types. Additionally, we did not inspect the severity of the divergences as annotated by participants.

Another limitation concerns our human evaluation. Humans may find it difficult to identify hallucinations and contradictions. This challenge may be due to the complexity of the task itself, or it may indicate that more time and resources are required for proper training and calibration (Thomson et al., 2023). This raises the question of whether crowd workers are truly suitable for such evaluation tasks, given the nuanced and challenging nature of the assessments required.

8 Ethical Considerations

In total, 119 participants were recruited through Prolific. Based on pilot studies, the task was expected to take over half an hour, so a minimum threshold of 20 minutes was set for accepting responses, with no upper bound defined. Participants were compensated at a rate of £6 per hour, with an additional £3 bonus for correctly answering the validation test question.

Supplementary Material Statement: Source code for our Hotel Highlights system cannot be made available due to our commercialisation of the software. Human evaluation dataset cannot be made available as it incorporates private user data. However, a suitably anonymised version may be made available under a license, upon contact with the authors.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv e-prints*, pages arXiv–2209.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Saad Mahamood and Maciej Zembrzuski. 2019. Hotel scribe: Generating high variation hotel descriptions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 391–396, Tokyo, Japan. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech Language*, 80:101482.

Minh Tien, François Portet, and Cyril Labbé. 2015. Hypertext Summarization for Hotel Review.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Item Example

Figure 3 shows an example of one the questions shown to participants in our human evaluation.

B Average Intrinsic Ratings

Tables 2 and 3 show the average ratings of the three intrinsic features (i.e. clarity, informativeness, and grammaticality) for each batch for both the whole group and the success group. Tables 4 and 5 show the sum of answers to the divergence question for each batch for both the whole group and the success group.

Batch	Cl	Gr	In	Av
B1	5.09 (1.59)	5.97 (1.27)	4.92 (1.66)	5.33
B2	4.94 (1.6)	5.82 (1.4)	4.85 (1.6)	5.2
B3	4.67 (1.58)	5.66 (1.33)	4.59 (1.64)	4.97
B4	5.0 (1.62)	5.91 (1.23)	4.93 (1.58)	5.28

Table 2: Average ratings per intrinsic questions per batch for the whole group. Cl, Gr, In, and Ave stand for Clarity, Grammaticality, Informativeness, and Total Average, respectively. Standard deviations are presented in parentheses.

Batch	Cl	Gr	In	Av
B1	5.05 (1.65)	6.12 (1.22)	4.95 (1.7)	5.37
B2	4.97 (1.55)	5.71 (1.47)	4.95 (1.53)	5.21
B3	4.48 (1.64)	5.57 (1.4)	4.36 (1.72)	4.8
B4	4.92 (1.62)	5.95 (1.19)	4.89 (1.58)	5.25

Table 3: Average ratings per intrinsic questions per batch for the success group. Cl, Gr, In, and Ave stand for Clarity, Grammaticality, Informativeness, and Total Average, respectively. Standard deviations are presented in parentheses.

Batch	Both	Cont	Hall	No_div	Total
B1	84	119	238	459	900
B2	87	130	201	482	900
B3	100	104	220	446	870
B4	76	135	176	513	900

Table 4: Sum of the answers to the divergence questions per batch for the whole group. Cont, Hall, and No_div stand for contradiction, hallucination, and no divergence, respectively.

C Correlation Analysis

Table 6 presents the correlation analysis between the three intrinsic ratings and the divergence question for the success group.

Batch	Both	Cont	Hall	No_divergence	Total
B1	53	85	151	281	570
B2	37	100	162	361	660
B3	75	84	161	310	630
B4	57	94	131	378	660

Table 5: Sum of the answers to the divergence questions per batch for the success group. Cl, Cont, Hall, and No_div stand for contradiction, hallucination, and no divergence, respectively.

	Div	Cl	In	Gr
Div	1.00	-0.71	-0.61	-0.27
Cl	-0.71	1.00	0.90	0.54
In	-0.61	0.90	1.00	0.48
Gr	-0.27	0.54	0.48	1.00

Table 6: Correlation analysis between divergence and the three intrinsic ratings for the answers by the success group. Div, Cl, In, and Gr stand for Divergence, Clarity, Informativeness, and Grammaticality, respectively.

D Theme Classification Results

Tables 7 and 8 present the aggregated mean ratings and divergence counts for different themes for the whole group and the success group.

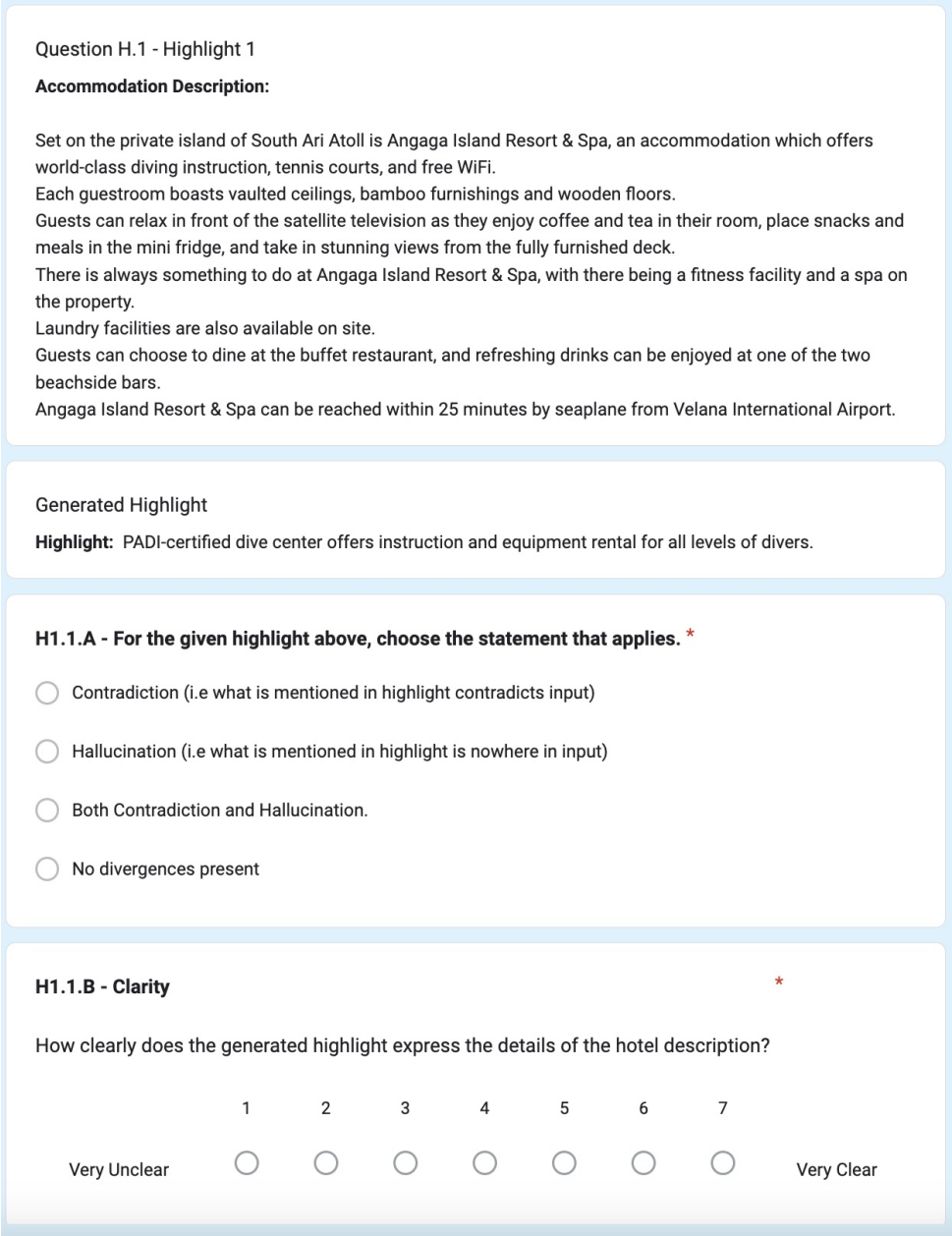

Figure 3: Example of one of the experimental items.

Theme	Av_Cl	Av_Gr	Av_In	total_count	No_div%	Div%
activities	4.71	5.77	4.75	475	55.37	44.63
dining and cuisine	5.01	5.85	5.01	922	59.87	40.13
facilities and amenities	5.12	5.86	5.03	1756	62.19	37.81
location	4.96	5.87	4.84	1073	55.17	44.83
wellness	5.12	5.86	5.11	327	70.64	29.36

Table 7: Aggregated mean ratings and divergence counts for different themes for the whole group. Table columns: Average Clarity (Av_Cl), Average Grammaticality (Av_Gr), Average Informativeness (Av_In), Total Count (total_count), No Divergence Percentage (No_div%), and Divergence Percentage (Div%).

Theme	Av_Cl	Av_Gr	Av_In	total_count	No_div%	Div%
activities	4.65	5.76	4.71	338	56.51	43.49
dining and cuisine	4.96	5.84	4.99	647	59.81	40.19
facilities and amenities	5.06	5.86	5	1236	62.62	37.38
location	4.93	5.87	4.87	764	55.5	44.5
wellness	5.09	5.83	5.11	233	71.67	28.33

Table 8: Aggregated mean ratings and divergence counts for different themes for the success group. Table columns: Average Clarity (Av_Cl), Average Grammaticality (Av_Gr), Average Informativeness (Av_In), Total Count (total_count), No Divergence Percentage (No_div%), and Divergence Percentage (Div%).