CONFERENCE PAPER (FULL TEXT) · PRESENTED AT THE CONGRESS (NOT PRINTED IN THE PROCEEDINGS)

When the Measure Fails Before the Model Does: Generative AI as a Source of Tourist Information in Kyrgyz, Turkish and English

Жасалма интеллект туристтик маалымат булагы катары: Генеративдик моделдердин кыргыз, түрк жана англис тилдериндеги жообунун сапатын салыштырма баалоо

Mahmut Vural · Zamirgul Kazakbaeva · Nasikat Alieva · Raşit Yılmaz · Altynai Toktomatova
Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan

Congress	8th International Turkic World Tourism Congress (8. Uluslararası Türk Dünyası Turizm Kongresi)
Date / Venue	25–27 September 2026, Ankara Hacı Bayram Veli University, Faculty of Tourism, Ankara, Türkiye
Status	Full-text paper presented at the congress; not printed in the proceedings book. This file is the authors' final full text.
Type	Full-text paper (unpublished)
Language	English (title also in Kyrgyz)
Keywords	Generative artificial intelligence; Tourist information; Kyrgyz language; Less-resourced languages; Coder reliability; Kyrgyzstan
ORCID	Mahmut Vural: 0000-0001-8946-2540 Zamirgul Kazakbaeva: 0000-0003-4493-0843 Nasikat Alieva: 0009-0002-6209-1553 Raşit Yılmaz: 0009-0009-1471-3929 Altynai Toktomatova: 0000-0001-7503-8749

HOW TO CITE (APA 7th ed.)

Vural, M., Kazakbaeva, Z., Alieva, N., Yılmaz, R., & Toktomatova, A. (2026, September 25–27). *When the measure fails before the model does: Generative AI as a source of tourist information in Kyrgyz, Turkish and English* [Paper presentation]. 8th International Turkic World Tourism Congress, Ankara, Türkiye.

IEEE

M. Vural, Z. Kazakbaeva, N. Alieva, R. Yılmaz, and A. Toktomatova, “When the measure fails before the model does: Generative AI as a source of tourist information in Kyrgyz, Turkish and English,” presented at the *8th Int. Turkic World Tourism Congr.*, Ankara, Türkiye, Sep. 2026.

When The Measure Fails Before the Model Does: Generative Ai as A Source of Tourist Information in Kyrgyz, Turkish and English

Mahmut VURAL, Assist. Prof. Dr. Kyrgyz-Turkish Manas University, Vocational School, Bishkek, Kyrgyzstan ORCID: 0000-0001-8946-2540 · mahmut.vural@manas.edu.kg
Zamirgul KAZAKBAEVA Kyrgyz-Turkish Manas University, Vocational School, Bishkek, Kyrgyzstan ORCID: 0000-0003-4493-0843 · zamira.kazakbaeva@manas.edu.kg
Nasikat ALIEVA Kyrgyz-Turkish Manas University, Vocational School, Bishkek, Kyrgyzstan ORCID: 0009-0002-6209-1553 · nasikat.alieva@manas.edu.kg
Rasit YILMAZ Kyrgyz-Turkish Manas University, Vocational School, Bishkek, Kyrgyzstan ORCID: 0009-0009-1471-3929 · rasit.yilmaz@manas.edu.kg
Altynai TOKTOMATOVA Kyrgyz-Turkish Manas University, School of Foreign Languages, Bishkek, Kyrgyzstan ORCID: 0000-0001-7503-8749 · altynai.toktomatova@manas.edu.kg

ABSTRACT

This study set out to compare the quality of the answers that generative artificial intelligence models give about the tourism resources of Kyrgyzstan in Kyrgyz, Turkish and English. Fifteen standardised tourist questions were put to ChatGPT, Gemini and DeepSeek in the three languages on 26 July 2026, and the resulting 135 answers were examined through content analysis. Three results held firm. Answer length followed the same order in every model, from Kyrgyz at the short end to English at the long end. Several answers carried serious factual errors, the most striking of which described a lake in Türkiye in place of Lake Issyk-Kul and placed the city of Osh inside Uzbekistan. When the same questions were repeated on 14 September 2026 under identical settings, those errors had gone. The fourth result was not the one that had been planned. The attempt to rate the answers on five criteria did not produce workable agreement between coders. The figure reached 0.93 for two coders in English but fell to 0.49 once a third expert joined, and it stood at 0.02 in Turkish and 0.03 in Kyrgyz. It was determined that the coders were in fact close to one another, since their scores fell within one point of each other in ninety to one hundred per cent of cases; the difficulty was that almost every answer received a four or a five, which leaves a reliability coefficient nothing to work with. From this it can be understood that a five-point rubric does not discriminate between answers in these languages, and that counting and fact-checking carry the evidential weight that rating was expected to carry.

Keywords: generative artificial intelligence, tourist information, Kyrgyz language, less-resourced languages, coder reliability, Kyrgyzstan

1. INTRODUCTION

It is known that the way a traveller looks for information has changed several times since the middle of the 1990s. Printed brochures gave way to destination websites, destination websites gave way to search engines, and search engines were joined by review platforms carrying content written by other travellers. Each of these steps left the traveller with a set of sources to compare. With the public release of ChatGPT at the end of 2022, that comparison step began to disappear. The traveller now receives a single answer, written in a confident register, with no competing version beside it.

This shift carries a particular consequence for destinations. A destination could once shape part of what a visitor read, through its official site, its social media accounts and its partner platforms. In this connection, that influence weakens. The answer a model gives is assembled from the material inside its training corpus, and neither the volume nor the accuracy of the Kyrgyzstan-related material in that corpus lies in the hands of the Kyrgyz Ministry of Culture, Information, Sport and Youth Policy or of any tour operator in Bishkek. The gatekeeper has moved.

For Kyrgyzstan the question is not an abstract one. According to the National Statistical Committee, 5.3 million people took a recorded holiday in the republic during 2025, tourism reached 3.8 per cent of gross domestic product, and the sector produced 74.4 billion soms in

gross value added. More than 3,100 firms and sole traders work in the field. Issyk-Kul, Son-Kul, Ala-Archa, Tash-Rabat and Sulaiman-Too in Osh carry most of the inbound flow. A sector of that size depends on the accuracy of what is said about it online, and increasingly on what is said about it by a language model.

Language enters the problem directly. Large language models do not spread their capacity evenly across languages, since the English portion of a training corpus is several times larger than that of most other languages, and languages with thin digital resources sit at the edge of the model (Joshi et al., 2020). Kyrgyz belongs to that group, and the limitation of its digital resources has been documented in the literature (Alekseev & Turatali, 2024). The useful question is therefore not whether a model speaks Kyrgyz but how well it speaks Kyrgyz about a specific place.

The aim of this study was to answer that question empirically. Fifteen tourist questions were prepared, translated into three languages and put to three publicly available models, and the resulting answers were examined. What the study found in the course of that work was not confined to the answers. A second finding emerged from the measuring process itself, and it is noteworthy that it turned out to be the more consequential of the two.

The paper proceeds in five further parts. The second part reviews the two bodies of literature that meet in this problem. The third describes the method and states plainly where it broke down. The fourth reports four findings. The fifth discusses what they mean for destinations and for research, and the sixth closes with recommendations and limitations.

2. LITERATURE REVIEW

2.1. Generative models in tourism

Automation in tourism was discussed for many years in terms of recommender systems, chatbots and service robots. Gretzel, Sigala, Xiang and Koo (2015) set out the smart tourism framework, in which a destination collects data from physical infrastructure, social connections and organisational sources, and converts that data into on-site experience and business value. In that framework the system offered options and the traveller chose among them. A generative model removes the choosing. It supplies one answer in continuous prose, and the traveller reads it as a conclusion rather than as a menu.

Carvalho and Ivanov (2024) mapped the application areas of generative models across tourism and hospitality and set the benefits against the risks. Efficiency gains in front-of-house and back-of-house work were identified, alongside the risk of overreliance on the output, the displacement of human workers, cybersecurity exposure and the loss of human contact. It is noteworthy that the same paper warns explicitly against passing sensitive customer information to such systems. The warning concerns the enterprise. The present study concerns what the traveller is told before any enterprise is involved.

Dwivedi et al. (2023) raised a point that bears directly on the design used here. There is no reliable link between how fluent an answer sounds and whether it is true. The smoother the prose, the more readily a reader accepts it, yet smoothness carries no information about accuracy. In a travel context this gap matters more than in most, because the reader usually has no way to check.

The experimental evidence comes from Kim, Kim, Park, Kim, Jhang and King (2025), whose six-stage study in the *Journal of Travel Research* measured what happens when an answer contains something untrue. It was determined that when the incorrect element sat in a prominent position within the answer, or belonged to the same domain as the question, participants became less willing to visit the recommended destination. The effect weakened when the errors were scattered across unrelated domains. A single factual error, placed where the reader is looking, therefore costs a destination something measurable.

Stergiou (2024) approached the same field through accessibility and diagnosticity, showing that a model's suggestions carry different weight at different stages of a trip. Before departure they shape the choice of destination; during the trip they shape logistics. What these studies share is a method: the user is surveyed, and the perception is modelled. What the answer itself contains is rarely opened. Almost all of this work has been done in English.

2.2. Language resource and its measurement

Performance in a given language tracks the share of that language in the training corpus. Joshi et al. (2020) sorted the world's languages into six classes by digital resource level and showed

that the great majority sit in the lowest classes. Alekseev and Turatali (2024) catalogued the natural language processing resources available for Kyrgyz and reported that the language remains in a difficult position despite having millions of speakers, with effort scattered across uncoordinated projects.

Three mechanisms are usually named for the gap. The first is corpus volume, which is the simplest. The second is tokenisation: in agglutinative languages, word forms break into many pieces and the model holds the meaning less securely. The third is cultural fit, since concepts grounded in local practice appear thinly in training material and the answer drifts towards the generic. All three bear on tourist content, which is built from place names, traditional terms and small logistical details.

Isbarov et al. (2025) tested Claude, Gemini, GPT and LLaMA across nine Turkic languages and more than 38,000 questions in the TUMLU benchmark. Understanding was found to be adequate; generation was not. The models often answered correctly but built the answer in a neighbouring resource-rich language, most often Turkish or Russian, rather than in the language of the question. For a traveller this distinction is decisive, since what is wanted is a usable answer and not evidence of comprehension.

Turatali, Turdubaeva, Izmailov, Alekseev and Nikolenko (2026) produced the first large-scale evaluation of models in Kyrgyz. KyrgyzLLM-Bench pairs two natively written datasets, KyrgyzMMLU with 7,977 curriculum-aligned items and KyrgyzRC with 400 reading-comprehension questions, with post-edited Kyrgyz versions of four established benchmarks, and tested twenty-six open and closed models. The picture is more differentiated than a simple deficit account. Model rankings transfer broadly from English to Kyrgyz on WinoGrande and BoolQ, while HellaSwag shows a wide gap that the authors attribute to translation-induced shifts in plausibility rather than to competence alone. The overall gap nonetheless remains substantial and is traced to imbalance in multilingual pretraining data.

2.3. Judging the output

If the quality of a model's answer is to be compared across languages, someone has to judge it, and the judging is itself a method with its own literature. Chiang and Lee (2023) showed that evaluations produced by a language model can align with expert human evaluation when the model receives the same instructions and the same rating questions as the human raters, and that the resulting scores stay stable across different phrasings of the instruction. Gu et al. (2026) surveyed the paradigm and catalogued its artefacts, of which position bias and verbosity bias are reported most often. Verbosity bias matters here, because answer length varies systematically by language in the data collected for this study.

Human coding brings its own requirements. Elo and Kyngäs (2008) describe content analysis as a systematic and replicable classification of written material, and note that its internal validity can be assessed through agreement coefficients. Cohen (1960) introduced kappa as a measure of agreement between two judges that removes the share attributable to chance. Hayes and Krippendorff (2007) argued that among the many coefficients proposed, Krippendorff's alpha should serve as the standard, since it holds across any number of coders, any level of measurement and any pattern of missing data. Their paper also sets the practical bar: data should not be relied upon when reliability falls short, and a coefficient below roughly 0.667 leaves conclusions unsafe.

A further caution appears in recent work on less-resourced languages. TukaBench (2026) reports that the reliability of language-model evaluators declines in exactly those languages where such evaluators would be most useful. That observation was noted during the design of this study and was expected to apply to the artificial component of the coding. It was not expected to apply to the human component.

2.4. Model behaviour over time

One more strand belongs in this review, and it was added after the data had been collected. Chen, Zaharia and Zou (2024) evaluated GPT-3.5 and GPT-4 in March 2023 and again in June 2023 on the same tasks. Performance moved, and not always upward: on one identification task accuracy fell from 84 per cent to 51 per cent across the three months. The conclusion drawn by the authors is that the behaviour of a given model can change substantially in a short period, and that continuous monitoring is therefore necessary.

For a study of the present kind the implication is uncomfortable and unavoidable. A measurement taken on one date describes the model on that date. Whether it describes the

model three months later is an empirical question, and one that can be tested rather than assumed.

2.5. The gap

Placing the two main bodies of work side by side leaves an evident gap. Tourism research examines generative models from the user's side and does not treat language as a variable. Computational linguistics measures the cross-linguistic gap precisely, but does so through school-curriculum items and general knowledge tests, far from the situation of a person planning a trip. The practical question — what quality of answer a traveller receives when asking in Kyrgyz — has fallen between them.

This study addresses that question, and in the course of addressing it encountered a second one that the literature had not prepared us for.

3. METHOD

3.1. Design

Content analysis was used as the method of data collection and analysis, since the object of study is the text itself rather than the opinion of a user (Elo & Kyngäs, 2008). The unit of analysis is one complete answer to one question. Secondary material was drawn from the scientific literature and from the publications of the National Statistical Committee and the Department of Tourism under the Ministry of Economy and Commerce. No human participants were involved and no ethics committee permission was required.

3.2. Question set

Fifteen questions were developed and grouped under five thematic headings: practical entry conditions, destinations and attractions, culture and etiquette, season and safety, and economy and logistics. Three conditions governed the selection. The question had to resemble what a real traveller would ask, and preliminary interviews with staff at Bishkek tour operators were used to identify recurring enquiries. The answer had to be checkable against an independent source, which excluded questions of opinion such as which place is best. The set had to cover the pre-trip, in-trip and cultural dimensions of a visit. Content validity was checked by three experts — a tourism practitioner, a language specialist and an instructor in information technologies — and two questions from the initial list were replaced on their advice.

3.3. Models, languages and collection protocol

Three publicly available models were selected: ChatGPT, Gemini and DeepSeek. The criterion was free accessibility for users in Kyrgyzstan together with the ability to accept questions in all three languages. Each question was asked in Kyrgyz, Turkish and English, and the three versions were equalised in length and detail through back-translation, since the size of a prompt can influence the size of an answer.

The following protocol was applied. Each question was asked in a fresh session so that earlier answers could not carry over, with memory and personalisation switched off. All questions were put within a single day, 26 July 2026, and the date was recorded. Reasoning level was left at the platform default, which is Auto in ChatGPT and 3.6 Flash in Gemini.

Answers were archived complete and unedited. No follow-up or clarifying question was asked, and only the first answer was taken, since most real users stop at the first answer.

It should be noted that one asymmetry exists. Web search was switched off in ChatGPT and in DeepSeek. In Gemini the function could not be switched off, so the Gemini cells were collected with search active. Each model remains internally consistent across its three languages, which leaves the cross-linguistic comparison intact; the comparison between models, however, cannot separate the contribution of the model from that of the retrieval layer. In total, 15 questions × 3 models × 3 languages produced 135 answers, and every one of them returned.

3.4. Measures that did not depend on rating

Two measures were taken without any judgement being exercised. Answer length was counted automatically in words. Factual accuracy was checked against independent sources: the electronic visa portal of the Kyrgyz Republic, the publications of the Department of Tourism, the data of the National Statistical Committee, the travel advice issued by foreign ministries, and the official sites of transport operators and mobile networks. An answer key

covering the verifiable facts of each question was prepared in advance from these sources, and every factual claim in the corpus was checked against it.

3.5. The rating procedure and what happened to it

Beyond those two measures, each answer was to be rated on a five-point scale for accuracy, completeness, currency, cultural appropriateness and linguistic quality. The rubric combined information-quality dimensions from the tourism literature with evaluation practice from computational linguistics.

A first pass across all 135 answers was carried out with the assistance of a large language model acting as evaluator, following the paradigm described by Chiang and Lee (2023). Human verification was then organised. Since answer length differs systematically across the three languages, the completeness criterion was anchored to coverage of the question rather than to volume of text, in order to limit the verbosity bias reported by Gu et al. (2026).

Human rating proceeded in several rounds between 13 and 29 September 2026. In the first two rounds, thirty answers drawn at random across all model–language cells were rated by members of the research team. In the following rounds, independent experts from outside the author group rated the same material in Turkish, English and Kyrgyz, and a further expert rated a Russian set collected separately. Coders did not see one another's scores and did not see which model had produced a given answer. Agreement was calculated through Cohen's kappa and Krippendorff's alpha (Cohen, 1960; Hayes & Krippendorff, 2007), with 0.667 taken as the minimum defensible level.

Ratings in Turkish, English and Russian were collected through a standalone web form, and ratings in Kyrgyz through Google Forms; the questions, the answers and the criteria were identical in every version.

3.6. The follow-up measurement

After the September ratings, a targeted check was carried out. The questions on which serious errors had been recorded in July were put again to the same model, in the same language, with the same settings, on 14 September 2026. The purpose was to separate a language effect from a time effect, following the logic set out by Chen, Zaharia and Zou (2024).

3.7. A note on Russian

A Russian set of 45 answers was collected on 14 September 2026 using the same protocol. Because seven weeks separate it from the main corpus, and because the follow-up check described above showed that at least one model changed materially in that interval, the Russian data are not entered into the main comparison. They are reported here only where they bear on the temporal question.

4. FINDINGS

4.1. Answer length

All 135 answers were obtained, with no blocked or empty return. The first finding required no judgement at all and is presented in Table 1.

Table 1
Mean Answer Length by Model and Language (*words*)

Model	Kyrgyz	Turkish	English
ChatGPT	183	261	353
Gemini	277	322	372
DeepSeek	378	398	713

Note. Counts are mean words per answer across the fifteen questions in each cell. Data collected 26 July 2026.

The ordering is identical in all three models and admits no exception. The ratio between the shortest and the longest condition reaches 1:1.9 inside a single model, with ChatGPT producing an average of 183 words in Kyrgyz against 353 in English, and DeepSeek producing 378 against 713. Because the prompts were equalised across the three languages through back-translation, the difference cannot be traced to the wording of the questions. What varies is what the model has to say, not what it was asked.

4.2. Factual errors

Errors were recorded against the answer key and fall into six categories: factual error, outdated information, over-general answer, language switching, distortion of place names, and misrepresentation of cultural context. Four cases deserve to be set out individually.

The visa question produced the clearest pattern. The correct rule for holders of ordinary Turkish passports, resting on the bilateral agreement of 10 February 2012, permits a stay of up to ninety days within any period of one hundred and eighty days. Gemini reported this correctly in all three languages. ChatGPT gave the ninety-day allowance in all three but omitted the reference window. DeepSeek gave three different and incorrect answers to the same question: sixty days in Kyrgyz, thirty days in Turkish, and thirty days per entry with a sixty-day ceiling in English. One model, one question, three mutually inconsistent claims about a rule published on an official portal.

Two failures in the Turkish cell of DeepSeek were of a different order. Asked for five places worth seeing around Lake Issyk-Kul, the model described Lake İznik in Türkiye, covering the ancient city walls, the Hagia Sophia of İznik and a Roman necropolis, with no element referring to Kyrgyzstan. Asked about Sulaiman-Too, the same model stated that Osh is a city in southern Uzbekistan. Following the taxonomy of Huang et al. (2025), both are factuality hallucinations rather than faithfulness hallucinations, since the generated content contradicts verifiable facts about the world.

The fourth case is shared rather than isolated. The Western Bus Station of Bishkek closed in October 2024, and services moved to the new terminal. Eight of the nine model–language cells still described the Western Station as the departure point for Issyk-Kul transport. The single exception was the Kyrgyz answer from Gemini, which named the new terminal correctly and also supplied the departure time and ticket classes of the seasonal Bishkek–Balykchy train. On this point the answer in the less-resourced language was the most current of the nine.

A fifth observation came from the Kyrgyz-speaking members of the research team and could not have been produced by automated checking. In a DeepSeek answer in Kyrgyz, the Russian idiom *в сухом остатке*, meaning "in summary", had been rendered word by word as *Сүзүмдө*, which in Kyrgyz means "in strained yoghurt". No Russian text appears in the answer. The contamination sits at the level of the idiom, which indicates that the Kyrgyz output has been produced through Russian and then translated. In this connection, the phenomenon reported by Isbarov et al. (2025) is present in the corpus, but in a form that a character-level check for language switching cannot detect.

4.3. Change over seven weeks

The follow-up measurement of 14 September 2026 produced a clear result, shown in Table 2.

Table 2
DeepSeek in Turkish, Identical Settings, Two Dates

Question	26 July 2026	14 September 2026
Visa period for Turkish citizens	30 days	90 days within 180 days
Five places around Issyk-Kul	Lake İznik, Türkiye	Cholpon-Ata, Ruh Ordo, Karakol, Jeti-Oguz, Skazka
Sulaiman-Too and its significance	Osh in southern Uzbekistan	Osh in southern Kyrgyzstan
Daily budget and currency exchange	correct	correct

Note. Same model, same language and identical settings on both dates.

Three serious errors disappeared together, while the control question that had been correct in July remained correct. Three independent failures do not resolve themselves simultaneously by chance. The model had been updated in the interval. It is noteworthy that this outcome, while it removes one explanation for the Russian results reported in passing, delivers something more valuable: a dated and documented instance of the instability described by Chen, Zaharia and Zou (2024).

4.4. What happened to the ratings

The fourth finding concerns the measuring instrument rather than the models. As it can be seen in the Table 3, agreement among human coders on the same thirty answers is reported by language.

Table 3
Agreement Among Human Coders

Language	Coders	Krippendorff's α	Note
English	2	0.93	high
English	3	0.49	falls when independent expert added
Turkish	3	0.02	effectively no agreement
Kyrgyz	3	0.03	effectively no agreement

Note. Krippendorff's alpha at ordinal level. Values below 0.667 are not considered defensible (Hayes & Krippendorff, 2007).

The English pair reached 0.93, which is comfortable by the standard set out in Hayes and Krippendorff (2007). Adding a third coder from outside the author group brought the figure down to 0.49, below the minimum defensible level. In Turkish the three-coder figure was 0.02 and in Kyrgyz it was 0.03, both of which are indistinguishable from chance.

The disagreement is not random, and its shape is informative. The external experts compressed their judgements into the middle of the scale, with standard deviations of 0.66 in English and 0.86 in Turkish, and used the top score rarely. The author-coders spread across the full range and leaned upward, with one Turkish coder assigning the top score to thirty-seven of fifty ratings. Rank order did not survive the difference either: Spearman correlations between the external experts and the other coders sat near zero.

The Kyrgyz data make the mechanism visible. The three Kyrgyz coders gave mean scores of 4.72, 4.54 and 3.96, with standard deviations between 0.50 and 0.63, and the two closest coders scored within one point of each other in every single case. They were not disagreeing about whether the answers were good; they agreed that almost all of them were. What they did not agree on was whether a given answer deserved a four or a five. When nearly every rating falls on the top two points of a scale, a chance-corrected coefficient such as alpha has almost no variance to explain and collapses towards zero. Such a result can occur because of two reasons: the answers may be genuinely similar in quality, or the rubric may not give coders a way to separate a good answer from a very good one. The data cannot separate these two explanations, but in either case the rubric fails to discriminate. In contrast to this situation, the English coders used the whole range of the scale, with a standard deviation of 1.25 for one of them, and agreement between the first two was high.

One element of the rating exercise nonetheless held. One external expert rated the complete English set of forty-four answers rather than a sample of ten. At the level of individual answers his scores correlated only weakly with the first-pass scores. At the level of the model, however, both produced the same ordering: ChatGPT ahead of Gemini, and Gemini ahead of DeepSeek. From this it can be understood that noise at the level of a single rating averages out across forty-four answers, even though it swamps any comparison based on ten.

5. DISCUSSION

5.1. Depth rather than correctness

The expectation formed from Joshi et al. (2020) and repeated for Kyrgyz by Turatali et al. (2026) was that the less-resourced language would perform worse across every dimension. The evidence here does not support that shape. What the data show instead is a difference in how much is said. Answer length follows the resource ordering exactly, in every model, without exception, and this is a count rather than a judgement.

Two explanations can be offered, and they are not exclusive. The material available in Kyrgyz, though limited in volume, is written natively rather than translated, so what the model does say about the threshold of a yurt or the contents of a national dish comes from sources that describe those things from the inside. At the same time the volume is insufficient to support the

elaboration, enumeration and qualification that produce a long answer. For the traveller the practical consequence is not a wrong answer but a thin one.

5.2. The instrument is part of the finding

The more consequential result of this study was not planned. Three competent, engaged coders produced three different orderings of the same ten answers. In Turkish the agreement figure sat at 0.02, and in Kyrgyz it sat at 0.03.

The Kyrgyz ratings behave like the Turkish ones. In both languages the coders concentrated on the upper end of the scale, and in both the coefficient fell to chance level. The difficulty is therefore not specific to Kyrgyz. It concerns the rubric itself when it is applied to answers that coders judge to be broadly acceptable.

The pattern across all the coders points the same way. Scale-use differences of this kind — external assessors clustering in the middle, internal assessors leaning high — are documented in the rater literature as leniency and range-restriction effects. What matters for a comparative study is that they make absolute scores incomparable across coders. A rubric that leaves this much room is not measuring the thing it names.

5.3. What survives and what does not

The claims that rest on counting survive. Answer length differs by language in a consistent direction. Specific factual errors occurred, are documented, and can be verified by anyone with the archived answers. The model changed within seven weeks in a way that removed three of those errors.

The claims that rest on rating do not survive. This study cannot state that the quality of Kyrgyz answers is lower than the quality of English answers, because the measurement that would support such a statement did not hold. It can say something weaker: coders in every language judged most answers to be acceptable, and the rubric did not allow them to rank those answers consistently. Nor can it report a defensible mean score for any language. Reporting those figures, as an earlier version of this work did, would attach a number to something the data do not contain.

Kim et al. (2025) showed that a single prominent factual error reduces the intention to visit a recommended destination. The errors recorded here satisfy both conditions identified in that study: they are prominent within the answer, and they sit at the centre of the question asked. A traveller told that Osh lies in Uzbekistan, or shown a lake in Türkiye in answer to a question about Issyk-Kul, is not receiving a marginally degraded service. The results of this study are therefore compatible with the studies in the literature that link factual error in generative output to reputational cost for a destination, even though the rating exercise designed to quantify that cost did not succeed.

5.4. Measurement in a moving field

The September re-test carries an implication that reaches beyond this paper. Three serious errors vanished in seven weeks, without any announcement and without any change on the user's side. Any study that measures a generative model at one point in time is describing that point in time. Chen, Zaharia and Zou (2024) made this argument from a three-month interval; the present data narrow the window to seven weeks.

This also offers the most plausible reading of a negative result reported earlier. Isbarov et al. (2025) found frequent drift into a neighbouring language when Turkic languages were used. No such switching appears in the forty-five Kyrgyz answers collected here. The idiom calque described in Section 4.2 suggests that Russian-mediated generation has not disappeared but has become less visible. Eighteen months separate the two studies, and in a field that moves this quickly, a benchmark result should be read with its date attached.

5.5. Implications for destinations

Three groups can act on these findings. For destination management organisations, official information in Kyrgyz needs to be published in a textual, machine-readable form. Content placed online as an image or inside a PDF cannot be read by a machine and therefore has no route into the training corpora that will produce tomorrow's answers. The KyrText corpus (2026) shows that targeted pretraining on natively written Kyrgyz material yields measurable improvement, which makes the availability of such material a policy question rather than a technical one.

For tourism enterprises, the categories in which errors clustered here — entry rules, prices and transport — are precisely the categories in which a wrong answer produces a stranded traveller. Verifying that information through the enterprise's own channels costs little and prevents the specific failures documented above.

In our opinion, the recommendation for the research community and for the Organization of Turkic States is less obvious and more important. The discussion of digital sovereignty in the Turkic world has focused on building shared corpora. This study suggests that shared evaluation instruments belong beside them. Without a rubric that has been tested in Kyrgyz and shown to discriminate between answers, comparisons of the kind attempted here cannot be made, and the absence of such instruments is not visible until someone tries.

6. CONCLUSION

The findings of this study, based on content analysis of 135 answers obtained from three generative models in three languages, can be summarized as follows:

Answer length follows the resource ordering of the three languages in every model without exception, from 183 words for ChatGPT in Kyrgyz to 713 words for DeepSeek in English. Serious factual errors occurred in the corpus, including the substitution of a lake in Türkiye for Lake Issyk-Kul, the placement of Osh inside Uzbekistan, and three mutually inconsistent visa periods given by one model across three languages. Eight of the nine model–language cells described a Bishkek bus station that closed in October 2024. When the same questions were repeated seven weeks later under identical settings, the three most serious errors had disappeared. Agreement among human coders did not reach a defensible level in any language. In Turkish and Kyrgyz it fell to chance level because coders placed nearly every answer on the top two points of the scale,.

According to the findings of this research the following recommendations can be given to practitioners and to policy makers:

- Official tourism information should be published in Kyrgyz in textual, machine-readable form, and the practice of posting information as images or PDF files should be abandoned.
- Tourism enterprises should verify through their own channels the information travellers obtain from generative models, with particular attention to entry rules, prices and transport.
- Open Kyrgyz-language corpora should be developed at state level and released under open licence.
- Shared evaluation instruments, and not only shared corpora, should be developed for the languages of the Turkic world within the framework of the Organization of Turkic States.
- A module on digital information literacy should be introduced into vocational tourism programmes, so that students learn to treat a fluent answer as a claim requiring verification.

6.1. Limitations and Suggestions for Further Researches

Several limitations bound these findings. The models are updated continuously, and the results describe the week in which the data were collected; the September re-test demonstrates this rather than merely warning about it. The free tiers of these platforms alter the underlying model without informing the user once a usage quota is reached, which limits what any study conducted on such interfaces can claim.

Web search could be disabled in ChatGPT and DeepSeek but not in Gemini, so the comparison between models cannot separate the model from its retrieval layer. A set of fifteen questions cannot cover every kind of tourist enquiry, and questions concerning travellers with particular needs were not included. The paid versions were not tested. The evaluation was carried out by researchers and experts rather than by travellers, so the study measures what the output contains and not what a traveller makes of it.

The most substantial limitation is the one reported as a finding. The rating component of the design did not discriminate between answers, and the conclusions drawn here rest on the measures that required no rating.

Further work should therefore begin with the instrument. A rubric validated in Kyrgyz, with worked examples and a calibration session before coding begins, is a prerequisite for any quantitative comparison in this field. Beyond that, the number of questions should be increased, the measurement should be repeated at defined intervals given the seven-week change observed here, and a design that tests the same model with and without retrieval would isolate the layer that this study could not. A version of the study covering several countries of the Turkic world would extend the picture. It can be offered to those who are interested in this subject that the question set, the answer key and the coding form developed here may be adapted for other less-resourced destination languages.

DECLARATION OF GENERATIVE AI USE

During the preparation of this manuscript the authors used generative artificial intelligence tools for drafting, for language editing and for the first-pass coding of the response corpus described in Section 3.5. All scientific content, the interpretation of the data and the final text are the responsibility of the authors, who reviewed and edited the output and take full responsibility for the content of the publication.

REFERENCES

Alekseev, A., & Turatali, T. (2024). *KyrgyzNLP: Challenges, progress, and future*. arXiv:2411.05503. https://arxiv.org/abs/2411.05503

Carvalho, I., & Ivanov, S. (2024). ChatGPT for tourism: Applications, benefits and risks. *Tourism Review*, 79(2), 290–303. https://doi.org/10.1108/TR-02-2023-0088

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time? *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.5317da47

Chiang, C.-H., & Lee, H.-Y. (2023). Can large language models be an alternative to human evaluations? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). Opinion paper: "So what if ChatGPT wrote it?" *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179–188. https://doi.org/10.1007/s12525-015-0196-8

Gu, J., et al. (2026). A survey on LLM-as-a-judge. *The Innovation*, 7(6), 101253. https://doi.org/10.1016/j.xinn.2025.101253

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. https://doi.org/10.1080/19312450709336664

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. https://doi.org/10.1145/3703155

Isbarov, J., et al. (2025). *TUMLU: A unified and native language understanding benchmark for Turkic languages*. arXiv:2502.11020. https://arxiv.org/abs/2502.11020

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of ACL 2020*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kim, J. H., Kim, J., Park, J., Kim, C., Jhang, J., & King, B. (2025). When ChatGPT gives incorrect answers: The impact of inaccurate information by generative AI on tourism decision-making. *Journal of Travel Research*, 64(1), 51–73. https://doi.org/10.1177/00472875231212996

KyrText: A multi-domain large-scale corpus for Kyrgyz language. (2026). *Proceedings of the LoResLM Workshop*. https://aclanthology.org/2026.loreslm-1.39/

National Statistical Committee of the Kyrgyz Republic. (2026). *Tourism in Kyrgyzstan: Main indicators*. Bishkek.

Stergiou, D. P. (2024). ChatGPT and tourist decision-making: An accessibility–diagnosticity theory perspective. *International Journal of Tourism Research*, 26(5), e2757. https://doi.org/10.1002/jtr.2757

TukaBench: A culturally grounded benchmark for African languages. (2026). arXiv:2606.01322. https://arxiv.org/abs/2606.01322

Turatali, T., Turdubaeva, A., Izmailov, R., Alekseev, A. M., & Nikolenko, S. I. (2026). *KyrgyzLLM-Bench: Benchmarking Kyrgyz language understanding*. arXiv:2607.17173. https://arxiv.org/abs/2607.17173