

RESEARCH WHITE PAPER

Language Tasks

Arabic-aware language intelligence. Precise error locations, controlled corrections and interpretable decisions.



Detect

Locate the source span

Correct

Preserve intended meaning

Explain

Keep the decision traceable

System design, real Arabic examples and a measured comparison of language-task outputs.

OVERVIEW

Measured correction. Controlled intervention.

A precision-oriented assessment on matched, expert-reviewed Arabic sentence pairs.

75.67%

Detection F0.5
Diacritic-insensitive

95.12%

Normalized correction accuracy
Conditional on localized errors

46.50%

Exact sentence match
Diacritic-insensitive

Executive overview

L3 supports enterprise language tasks through Arabic-aware analysis: preserving valid wording, locating errors and making controlled corrections. This evaluation compares Lisan with 25 language-model configurations against an expert reference for each sentence.

The principal ranking uses diacritic-insensitive F0.5 because unnecessary changes to valid text carry a cost. F0.5 emphasizes precision relative to recall. Diacritics and tatweel are removed symmetrically so harmless surface variation does not dominate detection or whole-sentence results.

Lisan records a detection F0.5 of 75.67%. GPT-6 Astra, the second-ranked system, achieves 65.60%. Conditional correction quality and exact sentence delivery are presented separately to keep error coverage visible.

Evaluation protocol. All quality measures use matched, expert-reviewed source-reference pairs and normalized reference edits. Results describe the tested configurations and outputs. The linked protocol explains the evaluation and scoring methods.

Reading guide

Language Tasks and system design	03–04
Methods and measurement layers	05–07
Measured quality	08–09
Observed Arabic outputs	10–13
Throughput and energy	14–15
Scope and complete matrices	16–22

01 · LINGUISTIC PROBLEM

Arabic language tasks across the sentence

A valid word form can still be incorrect in the sentence where it appears.

A valid word can still be wrong in its sentence. Arabic forms encode morphology, grammatical roles and attached particles; the right decision depends on their relationships.

FORM

Spelling and orthography

Hamza placement, letter forms and word spelling depend on the intended form and its context.

STRUCTURE

Morphology and agreement

Number, gender and grammatical case connect words across the sentence.

RELATION

Verb complementation

Verb choice and prepositions affect how a phrase expresses its meaning.

EXPRESSION

Lexical and editorial choice

The expert reference defines preferred wording. Other valid expressions can still differ from that target.

FUNCTION

Punctuation

A sentence's intent determines whether its ending is a statement, a question or a continuation.

INFORMATION

Diacritics and preservation

Optional marks can carry useful information. A correction should preserve valid content rather than erase it.

Preserve what is valid

The task is controlled correction. Locating an error, choosing its replacement and delivering an exact complete sentence remain separate questions.

The examples later in this paper show these decisions in real evaluated sentences, including cases that require stronger contextual review.

02 · SYSTEM DESIGN

An Arabic-first neuro-symbolic path

Linguistic structure and neural context support controlled, localized correction.

Lisan combines Arabic-aware tokenization, morphological and syntactic information, explicit linguistic rules and neural components. The design pairs learned coverage with explicit constraints and interpretable structure for language tasks.

FIGURE 1 · ARABIC LANGUAGE-TASK PIPELINE

- 01 Arabic-aware tokenization**
 Establish the linguistic text units used by the engine.
- 02 Morphology and tagging**
 Use word-internal structure and grammatical information.
- 03 Syntactic and rule analysis**
 Apply sentence-level dependencies and explicit linguistic constraints.
- 04 Neural and contextual checks**
 Combine learned coverage with the analyzed context.
- 05 Correction and explanation**
 Attach a correction and available linguistic metadata to a source span.

Error localization

Corrections are tied to explicit source spans rather than supplied only as a rewritten sentence.

Controlled alternatives

The engine may propose multiple valid suggestions. Only its first correction is scored here to match the single-output LLM protocol.

Explainable output

Rule and analysis metadata can describe the linguistic reason for a correction.

Production orientation

The architecture is designed for low latency and efficient batched processing. Observed service throughput is assessed separately from architectural intent.

03 · BENCHMARK CONSTRUCTION

One corpus. One expert reference per sentence.

Every system is evaluated against the same source-reference pairs.

Material was segmented with the sentence tokenizer used by the Lisan engine. A linguistic expert reviewed every sentence, either rewriting it as the corrected reference or retaining it unchanged.

FIGURE 2 · FROM SOURCE TO SCORING

Sentence boundaries

Apply the sentence tokenizer used by the Lisan engine.

Expert reference

Review each sentence, correcting it or retaining it unchanged.

Shared evaluation

Compare each tested output with the same expert reference.

Changed references assess correction. Unchanged references assess whether the system preserves acceptable input.

Evaluation protocol: [scope and scoring methods](#) ↗

Tested model families

Lisan

Claude

GPT family

Gemini family

DeepSeek

Llama

Qwen

Controlled language-task instruction

Correct only when needed

Correct spelling or grammar. Do not paraphrase, remove words or add diacritics.

Preserve correct input

Return the source sentence exactly when no correction is needed. Return sentence-only output.

Output standardization

Some configurations returned commentary, wrappers or a statement that the sentence was already correct. The preparation pipeline identified these cases and mapped them to the original sentence when appropriate, so commentary was not scored as Arabic edits.

The model versions in the result matrices identify the evaluated systems; the comparison does not claim coverage of every model in each family.

04 · EVALUATION METHODOLOGY

Detection begins in source coordinates

A correct error location and a correct replacement are separate evaluation questions.

The primary evaluator extracts a reference edit set from source to expert correction and a system edit set from source to model output. Token-aware alignment maps changes back to source character offsets.



Five edit-operation classes

Replacement	Deletion	Insertion	Split	Merge
-------------	----------	-----------	-------	-------

These are alignment operation classes. The retained results do not provide an error-type frequency distribution.

Detection counts

Count	Operational definition
TP	A system edit whose exact source span is present in the reference edit set.
FP	A system edit whose source span is absent from the reference edit set.
FN	A reference edit span absent from the system edit set.
TN	Not computed: the number of possible non-error spans is not well-defined.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad \text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1} = 2\text{PR} / (\text{P} + \text{R})$$

$$\text{F0.5} = 1.25\text{PR} / (0.25\text{P} + \text{R})$$

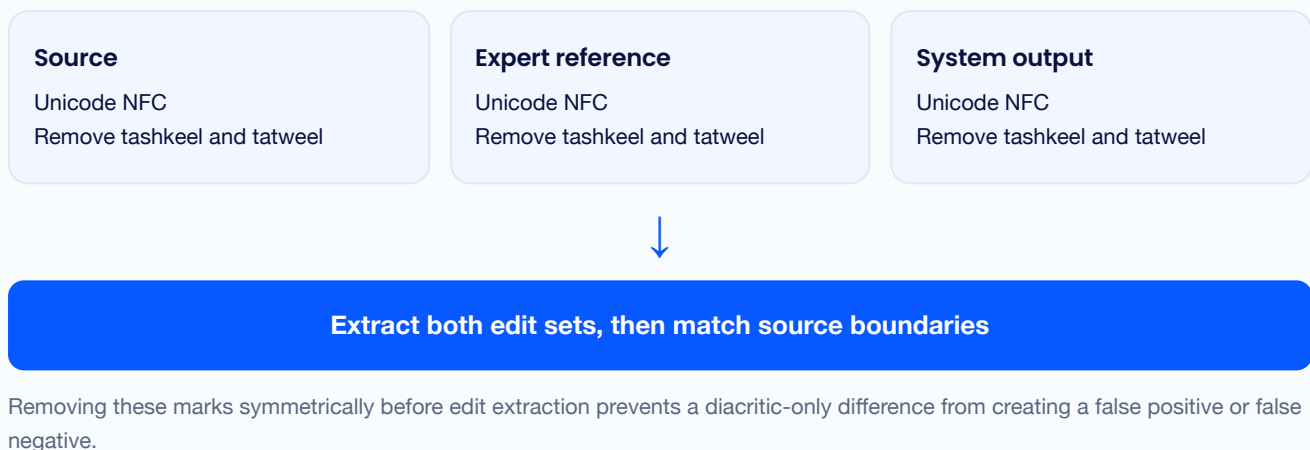
The formulas use P and R on the same scale. F0.5 is the primary ranking measure; precision and recall remain visible alongside F1.

04 · EVALUATION METHODOLOGY

Normalization and correction fidelity

Three measurement layers preserve the distinction between finding, fixing and delivering.

FIGURE 3 · SYMMETRIC DIACRITIC-INSENSITIVE EVALUATION



1. Detection: did the system find the error?

Detection compares source-span boundaries. It is scored independently of whether the proposed replacement equals the expert's replacement.

2. Correction: was the replacement right?

Normalized replacement accuracy is evaluated only after a true-positive source span is localized. Arabic normalization is applied before comparison, so superficial variation such as an additional or missing diacritic does not mark an otherwise correct suggestion as wrong.

Conditional denominator. Lisan's 95.12% normalized correction accuracy is conditional on correctly localized source spans. It does not measure missed errors. Detection precision, recall, F1 and F0.5 retain that separate coverage information.

3. Sentence delivery: was the whole output exact?

Diacritic-sensitive exact sentence match requires every character in the complete output to equal the expert reference. **Diacritic-insensitive exact sentence match** removes tashkeel and tatweel from both before the same complete-sentence comparison. The latter is used in the main scorecard; both are retained in Appendix A2.

Whole-output similarity and edit-distance measures supplement these layers. They leave the primary aligned-edit precision, recall and F-scores unchanged.

05 · PRIMARY RESULTS

Precision-weighted detection leadership

Ranked by diacritic-insensitive F0.5 across all 26 tested configurations.

FIGURE 4 · TOP FIVE DETECTION F0.5



Higher is better. Values are percentages on a common 0–100 scale.

FIGURE 5 · TOP FIVE DETECTION F1



F1 balances precision and recall. The top-five membership differs from the F0.5 ranking.

+10.07 pp

F0.5 lead over GPT-6 Astra

+15.60 pp

F1 lead over GPT-6 Astra

Lisan detection profile: 78.24% precision, 66.88% recall, 72.12% F1 and 75.67% F0.5. Precision and recall preserve the distinction between unnecessary changes and missed reference edits.

Appendix A1 preserves precision, recall, F1 and F0.5 for every model, including the five leading configurations displayed here.

Replacement quality and complete-sentence fidelity

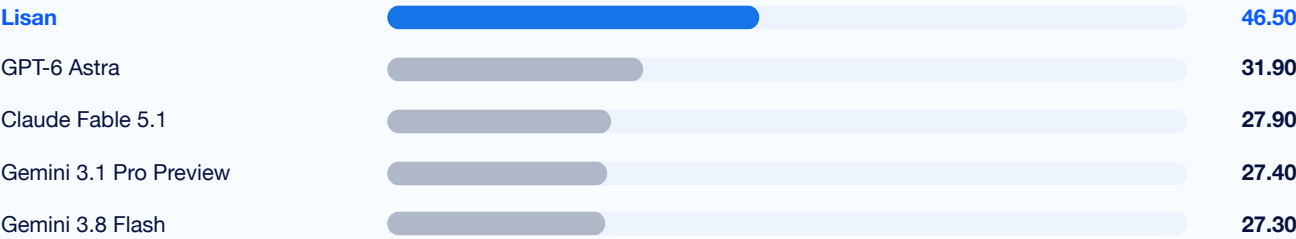
Strong conditional correction quality is presented alongside end-to-end sentence delivery.

FIGURE 6 · TOP FIVE NORMALIZED CORRECTION ACCURACY



Conditional on true-positive span localization. Higher is better; 0–100 scale.

FIGURE 7 · TOP FIVE EXACT SENTENCE MATCH



Diacritic-insensitive whole-sentence comparison. Higher is better; 0–100 scale.

95.12% Lisan normalized correction accuracy; 0.70 percentage points below Gemini 3.8 Flash at 95.82%.

46.50% Lisan exact sentence match; 14.60 percentage points above GPT-6 Astra at 31.90%.

Supporting whole-output indicators

Metric	Lisan	Strongest LLM result	Comparison
Character error rate	1.08%	GPT-6 Astra · 1.71%	37.0% lower
Word error rate	3.50%	GPT-6 Astra · 4.51%	22.4% lower
No-op accuracy	64.85%	Claude Fable 5.1 · 94.55%	unchanged references

Lower CER and WER are better. Relative differences use the displayed rounded values. Appendix B1 retains the complete whole-output and no-op results.

OBSERVED OUTPUTS / ORTHOGRAPHY

The right form. The same meaning.

Exact Arabic text from the evaluated sentence and its saved model outputs.

Source

نعم يمكن لكتاب واحد يقرأه الإنسان في مرحلة من عمره فيجد فيه ما يضيء له طريقه إلى النجاح

Expert reference

نعم يمكن لكتاب واحد يقرأه الإنسان في مرحلة من عمره فيجد فيه ما يضيء له طريقه إلى النجاح.

Lisan (L3)

نعم يمكن لكتاب واحد يقرأه الإنسان في مرحلة من عمره فيجد فيه ما يضيء له طريقه إلى النجاح.

GPT-6 Astra

نعم يمكن لكتاب واحد أن يقرأه الإنسان في مرحلة من عمره فيجد فيه ما يضيء له طريقه إلى النجاح

Gemini 3.8 Flash

نعم يمكن لكتاب واحد يقرأه الإنسان في مرحلة من عمره فيجد فيه ما يضيء له طريقه إلى النجاح

Reference decision

The reference changes the verb form and final hamza, then adds a full stop. Lisan matches that complete reference.

Alternative outputs

The other outputs fix the final hamza but retain a different verb construction. The evaluator records those differences against the reference.

Blue emphasis marks changed source words in the reference and differences from the reference in system outputs. The original characters, punctuation and spacing are retained.

OBSERVED OUTPUTS / GRAMMATICAL RELATIONS

Agreement depends on the relationship.

A local word form and a sentence-level dependency lead to different correction decisions.

A shared correction

Source	إليكم هذا الكلام ذو المعنى الجميل:
Expert reference	إليكم هذا الكلام ذا المعنى الجميل:
Lisan (L3)	إليكم هذا الكلام ذا المعنى الجميل:
GPT-6 Astra	إليكم هذا الكلام ذا المعنى الجميل:

The expert reference selects ذَا; both displayed systems make that correction.

Preserving the reference's agreement

Source	بين منطوق لم يُقصد، ومقصود لم يُنطق تضيع الكثير من المحبة.
Expert reference	بين منطوق لم يُقصد، ومقصود لم يُنطق تضيع كثير من المحبة.
Lisan (L3)	بين منطوق لم يُقصد، ومقصود لم يُنطق تضيع كثير من المحبة.
GPT-6 Astra	بين منطوق لم يُقصد، ومقصود لم يُنطق يضيع الكثير من المحبة.

Lisan follows the reference's wording and feminine verb. The compared output changes the verb as well. The difference is visible at the source span.

OBSERVED OUTPUTS / DIACRITICS

Keep the information. Correct the marks.

Normalization for scoring and preservation in delivered text are different decisions.

Source

قطر الذرة الجَاهِز في الأسواق مع القليل من الجوز الأمريكي أو الجَّوز العادي المحمَّص.

Expert reference

قطر الذرة الجَاهِز في الأسواق مع القليل من الجوز الأمريكي أو الجَّوز العادي المحمَّص.

Lisan (L3)

قطر الذرة الجَاهِز في الأسواق مع القليل من الجوز الأمريكي أو الجَّوز العادي المحمَّص.

GPT-6 Astra

قطر الذرة الجَاهِز في الأسواق مع القليل من الجوز الأمريكي أو الجوز العادي المحمَّص.

Gemini 3.8 Flash

قطر الذرة الجَاهِز في الأسواق مع القليل من الجوز الأمريكي أو الجوز العادي المحمَّص.

Delivered text

Lisan retains the reference's corrected marks. The displayed comparison outputs remove marks more broadly.

Evaluation lens

Diacritic-insensitive detection removes tashkeel and tatweel symmetrically. The paper presents normalized correction quality alongside both whole-sentence views.

This observed example explains why the paper shows several measurement layers. It does not add a new aggregate score for diacritic handling.

OBSERVED OUTPUTS / REVIEW BOUNDARIES

Context can change the best decision.

The same evidence also shows where the reference is better preserved by another system.

Preserving the intended number

Source	حبتان بندورة بحيث تكون معصورة.
Expert reference	حبتا بندورة بحيث تكونان معصورتين.
Lisan (L3)	حبة بندورة بحيث تكون معصورة.
GPT-6 Astra	حبتا بندورة بحيث تكونان معصورتين.

The reference keeps the dual meaning across the phrase. Astra matches it; Lisan changes the noun to singular.

Recognising a question

Source	من هو أفضل شريك لبرج القوس
Expert reference	من هو أفضل شريك لبرج القوس؟
Lisan (L3)	من هو أفضل شريك لبرج القوس.
GPT-6 Astra	من هو أفضل شريك لبرج القوس؟

The expert reference ends with a question mark. Astra retains that function; Lisan supplies a full stop.

These are evaluated sentence examples, not a separate comparison score. They support keeping contextual review and complete-sentence fidelity visible alongside detection.

06 · OPERATIONAL RESULTS

Observed corrected-text throughput

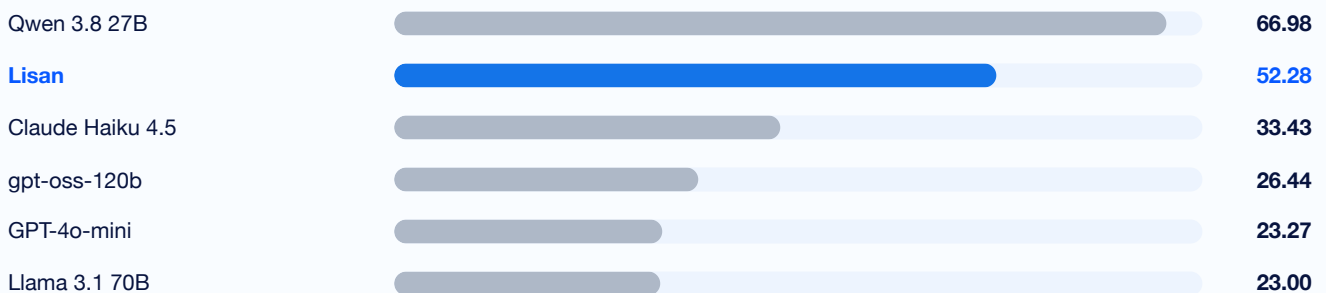
A shared text-unit denominator across providers with different native tokenizers.

The operational measure applies Lisan's tokenizer to each completed corrected output, then divides the standardized token count by accumulated end-to-end request time. This creates one text-unit denominator across providers whose native tokenizers are incompatible.

$$\text{Standardized throughput} = \frac{\text{Corrected-output tokens, counted with Lisan's tokenizer}}{\text{Accumulated end-to-end request time in seconds}}$$

The accumulated time includes network, request, prompt-processing and generation overhead. It is therefore a delivered-text service measure rather than raw decoder throughput.

FIGURE 8 · SIX HIGHEST OBSERVED SERVICE-THROUGHPUT RESULTS



Standardized corrected-output tokens per second. Bars share a 0–70 scale.

52.28

Lisan standardized output tokens/s
Second in the available logs

1.18

GPT-6 Astra tokens/s
Observed service throughput

Operational scope. This service-throughput measure includes each provider's serving stack and request overhead. The separate A10 test on the next page measures sustained batch processing and accelerator energy.

Appendix B2 retains observed throughput for every tested configuration. Quality and operational results should be interpreted using their respective measurement scopes.

07 · EFFICIENT DEPLOYMENT

Efficient inference. Measured at the accelerator.

A steady-state A10 test separates processing efficiency from service throughput.

2,000

Tokens per second
Sustained batch processing

100 W

Mean GPU power
NVIDIA A10 · 24 GB

POWER ÷ PROCESSING RATE

100 W ÷ 2,000 tokens/s

0.05 J per token

Energy at the accelerator, derived from the measured power and throughput.

ENERGY SCOPE: MEASUREMENT AND PLANNING SCENARIO

Accelerator

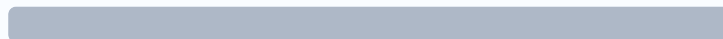
Measured power ÷ throughput



0.000014

Full-stack scenario

Accelerator result × 3



0.000042

kWh per 1,000 processed tokens · The full-stack scenario models overhead; it is not a full-system power measurement or a competitor benchmark.

Controlled steady state

At least 95% GPU utilization. A 30-second warmup is followed by a 10-minute measurement window. Power is sampled at 1 Hz with nvidia-smi; dynamic frequency scaling is enabled, with no throttling recorded.

Read the scopes separately

The 52.28 tokens/s service result includes request and network overhead. The 2,000 tokens/s figure measures saturated batch processing. Host, cooling and data-center overhead are outside the accelerator measurement.

08 · INTERPRETATION

Scope, limitations and reproducibility

Interpret results through their task, reference and measurement conditions.

01 One expert reference

Each sentence has one expert correction. A different valid correction can be penalized even when linguistically acceptable.

02 A finite benchmark

The expert-reviewed Arabic corpus provides a comparative snapshot. Broader claims require further domains, writers, dialect mixtures and error types.

03 Tested model snapshots

Hosted systems, configurations and provider behavior can change. The results describe the tested outputs rather than every future version of a model family.

04 One Lisan suggestion

Only the first proposed correction was scored, even when the engine could provide several valid alternatives.

05 Point estimates

The evaluation reports point estimates for the tested material.

How each result is grounded

Measurement layer	Retained outcomes	Denominator and scope
Primary aligned-edit evaluator	Precision, recall, F1 and F0.5	Normalized reference spans across the shared source-reference pairs.
Replacement comparison	Normalized correction accuracy	Conditional on correctly localized TP spans. Lisan's normalized correction score is 95.12%.
Whole-output evaluation	Exact match, CER, WER and no-op accuracy	Diacritic/tatweel-stripped strings across the shared benchmark; no-op uses unchanged references.
Operational timing logs	Standardized output tokens/s	End-to-end service throughput retokenized with Lisan; not identical-hardware latency.

The expert reference defines the evaluation target. The appendices preserve every tested configuration and its measured quality and service-throughput results.

APPENDIX A1 · COMPLETE DETECTION RESULTS

Detection matrix

Ranks 1–13

Precision, recall, F1 and F0.5 are shown as percentages.

Rank · System	P	R	F1	F0.5
01 Lisan	78.24%	66.88%	72.12%	75.67%
02 GPT-6 Astra	73.47%	45.92%	56.52%	65.60%
03 Gemini 3.8 Flash	74.08%	33.89%	46.51%	59.88%
04 Claude Fable 5.1	78.18%	30.63%	44.01%	59.66%
05 Gemini 3.1 Pro Preview	75.43%	31.79%	44.73%	59.18%
06 Claude Opus 5	61.03%	40.64%	48.79%	55.46%
07 Gemini 3 Flash Preview	62.48%	36.55%	46.12%	54.72%
08 GPT-5.6 Luna	57.62%	39.95%	47.18%	52.94%
09 DeepSeek-V4-Flash-0731	56.89%	35.65%	43.83%	50.83%
10 Gemini 3.5 Flash-Lite	65.96%	25.47%	36.75%	50.05%
11 GPT-5.6 Terra	54.32%	37.84%	44.61%	49.97%
12 GPT-5.1	53.10%	37.50%	43.96%	49.02%
13 DeepSeek-V4-Pro	54.49%	33.63%	41.59%	48.48%

Reading this matrix. P is precision; R is recall. All detection metrics are micro-averaged and diacritic-insensitive. TP + FN covers the same normalized reference edit set for every system. Rows are ordered by F0.5, with model versions retained.

The table continues across pages 17–18. Correction and whole-sentence metrics for the same rank order follow in Appendix A2.

APPENDIX A1 · COMPLETE DETECTION RESULTS

Detection matrix

Ranks 14–26

Precision, recall, F1 and F0.5 are shown as percentages.

Rank · System	P	R	F1	F0.5
14 GPT-5.6 Sol	48.47%	43.00%	45.57%	47.27%
15 GPT-5.4	48.10%	39.13%	43.15%	45.99%
16 Gemini 2.5 Flash	62.72%	21.82%	32.38%	45.62%
17 Claude Sonnet 5	50.87%	31.40%	38.83%	45.26%
18 Qwen 3.6 27B	47.99%	24.66%	32.58%	40.35%
19 GPT-4o-mini	38.12%	34.24%	36.07%	37.27%
20 Qwen 3.8 27B	39.17%	27.49%	32.31%	36.10%
21 Gemma 3 27B	35.43%	38.40%	36.86%	35.99%
22 Claude Haiku 4.5	34.06%	28.99%	31.32%	32.91%
23 gpt-oss-120b	31.68%	29.98%	30.81%	31.33%
24 Llama 4 Maverick	27.14%	30.20%	28.59%	27.70%
25 Llama 3.1 70B	26.16%	22.77%	24.35%	25.40%
26 Llama 3.3 70B	13.61%	21.52%	16.67%	14.69%

Reading this matrix. P is precision; R is recall. All detection metrics are micro-averaged and diacritic-insensitive. TP + FN covers the same normalized reference edit set for every system. Rows are ordered by F0.5, with model versions retained.

The table continues across pages 17–18. Correction and whole-sentence metrics for the same rank order follow in Appendix A2.

APPENDIX A2 · COMPLETE CORRECTION RESULTS

Correction and delivery

Ranks 1–13

Conditional replacement accuracy and complete-sentence exact match, with both normalization views.

Rank · System	Correction Normalized	Sentence Insensitive	Sentence Sensitive
01 Lisan	95.12%	46.50%	44.30%
02 GPT-6 Astra	95.70%	31.90%	20.60%
03 Gemini 3.8 Flash	95.82%	27.30%	14.20%
04 Claude Fable 5.1	95.51%	27.90%	24.00%
05 Gemini 3.1 Pro Preview	95.41%	27.40%	13.90%
06 Claude Opus 5	94.29%	25.30%	21.20%
07 Gemini 3 Flash Preview	94.83%	27.00%	15.20%
08 GPT-5.6 Luna	94.95%	24.50%	13.10%
09 DeepSeek-V4-Flash-0731	92.77%	22.70%	12.40%
10 Gemini 3.5 Flash-Lite	94.94%	24.40%	18.60%
11 GPT-5.6 Terra	94.21%	23.20%	15.30%
12 GPT-5.1	93.81%	21.50%	11.10%
13 DeepSeek-V4-Pro	92.46%	23.50%	14.20%

Correction normalized

Normalized replacement accuracy is conditional on correctly localized true-positive source span. The normalization policy prevents superficial Arabic variation from deciding replacement correctness.

Sentence columns

Both columns compare the entire output with the expert reference. “Insensitive” removes diacritics and tatweel; “Sensitive” requires every character to match.

The same F0.5 rank order is used across Appendix A1, Appendix A2 and Appendix B1. The table continues across pages 19–20.

APPENDIX A2 · COMPLETE CORRECTION RESULTS

Correction and delivery

Ranks 14–26

Conditional replacement accuracy and complete-sentence exact match, with both normalization views.

Rank · System	Correction Normalized	Sentence Insensitive	Sentence Sensitive
14 GPT-5.6 Sol	92.81%	21.60%	10.50%
15 GPT-5.4	92.86%	20.30%	9.50%
16 Gemini 2.5 Flash	92.91%	21.40%	14.40%
17 Claude Sonnet 5	93.02%	21.70%	15.60%
18 Qwen 3.6 27B	87.28%	18.10%	13.70%
19 GPT-4o-mini	90.09%	17.50%	10.50%
20 Qwen 3.8 27B	85.62%	14.80%	10.60%
21 Gemma 3 27B	89.60%	16.60%	7.70%
22 Claude Haiku 4.5	88.44%	16.60%	8.70%
23 gpt-oss-120b	86.68%	12.00%	7.00%
24 Llama 4 Maverick	88.48%	8.50%	2.50%
25 Llama 3.1 70B	85.47%	8.30%	4.70%
26 Llama 3.3 70B	78.84%	3.00%	1.00%

Correction normalized

Normalized replacement accuracy is conditional on correctly localized true-positive source span. The normalization policy prevents superficial Arabic variation from deciding replacement correctness.

Sentence columns

Both columns compare the entire output with the expert reference. “Insensitive” removes diacritics and tatweel; “Sensitive” requires every character to match.

The same F0.5 rank order is used across Appendix A1, Appendix A2 and Appendix B1. The table continues across pages 19–20.

APPENDIX B1 · COMPLETE WHOLE-OUTPUT RESULTS

Whole-output quality

All 26 configurations, ordered by primary F0.5 rank. Diacritic-insensitive metrics.

Rank · System	Exact match	CER ↓	WER ↓	No-op accuracy
01 Lisan	46.50%	1.08%	3.50%	64.85%
02 GPT-6 Astra	31.90%	1.71%	4.51%	88.48%
03 Gemini 3.8 Flash	27.30%	1.79%	4.98%	76.97%
04 Claude Fable 5.1	27.90%	1.81%	4.97%	94.55%
05 Gemini 3.1 Pro Preview	27.40%	1.82%	5.02%	79.39%
06 Claude Opus 5	25.30%	1.94%	5.46%	73.33%
07 Gemini 3 Flash Preview	27.00%	1.92%	5.38%	80.00%
08 GPT-5.6 Luna	24.50%	2.00%	5.69%	70.30%
09 DeepSeek-V4-Flash-0731	22.70%	2.08%	5.84%	73.94%
10 Gemini 3.5 Flash-Lite	24.40%	1.91%	5.55%	86.67%
11 GPT-5.6 Terra	23.20%	2.02%	5.95%	70.91%
12 GPT-5.1	21.50%	2.06%	6.02%	58.18%
13 DeepSeek-V4-Pro	23.50%	2.18%	6.01%	73.94%
14 GPT-5.6 Sol	21.60%	2.14%	6.50%	63.64%
15 GPT-5.4	20.30%	2.16%	6.43%	57.58%
16 Gemini 2.5 Flash	21.40%	2.04%	5.87%	92.73%
17 Claude Sonnet 5	21.70%	2.14%	6.15%	62.42%
18 Qwen 3.6 27B	18.10%	2.21%	6.57%	67.88%
19 GPT-4o-mini	17.50%	2.74%	7.62%	44.85%
20 Qwen 3.8 27B	14.80%	2.72%	7.50%	61.82%
21 Gemma 3 27B	16.60%	2.90%	8.24%	41.21%
22 Claude Haiku 4.5	16.60%	2.64%	7.82%	49.70%
23 gpt-oss-120b	12.00%	2.95%	8.56%	45.45%
24 Llama 4 Maverick	8.50%	3.07%	9.38%	24.85%
25 Llama 3.1 70B	8.30%	3.22%	9.18%	28.48%
26 Llama 3.3 70B	3.00%	4.40%	13.99%	10.30%

CER is character error rate; **WER** is word error rate. Lower is better for both. Exact match and no-op accuracy are higher-is-better measures. No-op accuracy assesses preservation on the unchanged references; the full-output measures supplement the aligned-edit evaluator.

APPENDIX B2 · COMPLETE OPERATIONAL RESULTS

Observed service throughput

Every tested configuration, ranked by standardized corrected-output tokens per second.

Rank	System	Output tokens/s
01	Qwen 3.8 27B	66.98
02	Lisan	52.28
03	Claude Haiku 4.5	33.43
04	gpt-oss-120b	26.44
05	GPT-4o-mini	23.27
06	Llama 3.1 70B	23.00
07	GPT-5.4	22.73
08	GPT-5.1	21.67
09	DeepSeek-V4-Flash-0731	21.16
10	DeepSeek-V4-Pro	20.71
11	GPT-5.6 Terra	16.28
12	Gemini 2.5 Flash	16.05
13	Claude Fable 5.1	15.07
14	Gemma 3 27B	14.56
15	GPT-5.6 Luna	12.58
16	Claude Sonnet 5	12.29
17	GPT-5.6 Sol	8.38
18	Gemini 3.1 Pro Preview	8.05
19	Qwen 3.6 27B	7.90
20	Claude Opus 5	7.43
21	Gemini 3.8 Flash	6.25
22	Llama 3.3 70B	5.98
23	Llama 4 Maverick	5.53
24	Gemini 3.5 Flash-Lite	2.91
25	Gemini 3 Flash Preview	2.43
26	GPT-6 Astra	1.18

Corrected outputs are tokenized with Lisan and divided by accumulated end-to-end request time. Values include service overhead and are observed logs rather than a hardware-identical experiment. Each row retains the observed service-throughput result.

EVIDENCE YOU CAN INSPECT

From the workbook to the comparison

The comparison pairs each source sentence with its expert correction and the output of each tested configuration. The reader and downloadable paper present the same results.

01 Complete source and output coverage

The accompanying comparison preserves every source/reference pair and tested system output from the supplied workbook. The text matches the workbook cells exactly, including punctuation and diacritics.

02 Separate finding from fixing

Lisan's detection scorecard distinguishes correct localization, unnecessary edits and missed reference edits. The 95.12% replacement score applies only after a true-positive span has been localized; 46.50% measures exact complete-sentence delivery.

03 Inspect alternatives in context

The tests let readers compare the original, expert reference, Lisan and another tested configuration. A different wording can be linguistically valid while differing from the expert reference.

What full reproduction still requires

The workbook preserves the comparison outputs. Recreating the scores additionally needs the exact tokenizer and span evaluator, output-cleanup rules, pinned provider requests and raw timing logs. The results retain the original evaluation outputs.

Model names identify the tested outputs. The accelerator-energy result uses the A10 measurement conditions; the full-stack estimate is a separate planning scenario.

[Read the evaluation scope and scoring methods ↗](#)

[Open the sentence and model comparisons ↗](#)