

MAGADIR at WMT 2026: Cross-System QE Re-ranking and Entity-Aware Selection for Low-Resource Arabic–English Translation

Hossam Boudraa
Tampere University
hossam.boudraa@tuni.fi

Abstract

We describe the MAGADIR submissions to the WMT 2026 Low-Resource Arabic–Asian Language Translation Shared Task for English $\leftrightarrow$ Arabic. Our systems combine a small official parallel corpus, open-weight multilingual translation models, and reference-free quality estimation (QE) to select outputs for a blind Challenge Test set. Selection is segment-level QE re-ranking over a heterogeneous cross-system candidate pool; for AR $\rightarrow$ EN the primary system additionally applies a Wikidata-backed *entity-recall floor* that vetoes candidates omitting known proper names before the QE utility (MetricX-24-QE) decides. On the official Challenge Test set, the AR $\rightarrow$ EN primary obtains 27.45 BLEU, 56.61 chrF₂, 64.49 TER, 81.94 COMET, and 81.28 BERTScore; the EN $\rightarrow$ AR primary obtains 14.59 BLEU, 51.48 chrF₂, 80.85 TER, 83.88 COMET, and 83.13 mBERT. Relative to our own contrastive runs, QE-guided primaries gain up to +1.94 COMET (AR $\rightarrow$ EN) and +0.81 COMET (EN $\rightarrow$ AR) while sometimes conceding BLEU, exposing a consistent trade-off between lexical-overlap metrics and neural adequacy objectives. Because the AR $\rightarrow$ EN primary differs from its no-floor contrastive in QE utility as well as in the floor, those deltas (+0.37 COMET, −0.60 BLEU) measure the two jointly; we therefore isolate the floor with a post-hoc control that applies the primary’s utility without the floor, completing a 2×2 on the blind pool. Reference-free, the floor delivers 83% of the entity-recall gain while rewriting 7% of outputs at no COMETKiwi cost (+0.0002), whereas the utility switch rewrites 74% for the remaining 17% and carries the entire −0.0085 COMETKiwi loss: the veto is close to free, and the adequacy cost belongs to the utility. Reference-free entity auditing of this kind requires only the source, the hypothesis and a gazetteer, and we recommend it as a routine check on QE-selected systems.

1 Introduction

Low-resource machine translation (MT) increasingly depends on adapting and selecting among large multilingual systems rather than training task-specific models from scratch. This setting is especially demanding for Arabic-centered translation, where orthographic normalization, rich morphology, dialectal variation, and named-entity transliteration make surface-overlap metrics fragile. The WMT 2026 Low-Resource Arabic–Asian Language Translation Shared Task (Laskar et al., 2026) provides a compact English–Arabic parallel corpus and blind Challenge Test sources, requiring systems that generalize beyond a small supervised sample.

MAGADIR submitted systems for Sub-Task 1A, English $\rightarrow$ Arabic (EN $\rightarrow$ AR), and Sub-Task 2A, Arabic $\rightarrow$ English (AR $\rightarrow$ EN) (Laskar et al., 2026). The official data contain 21,000 training sentence pairs, 1,000 development pairs, and 500 devtest pairs; the blind Challenge Test contains 1,500 source segments per direction. Because no references were available for the blind set at selection time, we frame final system construction as a *reference-free decision problem*: generate a diverse candidate pool, score candidates with QE models, and select the candidate that best balances semantic adequacy, fluency, and entity preservation (Figure 1).

Our contributions are threefold:

1. A compact heterogeneous candidate pool.

We combine supervised LoRA adapters over NLLB and Hunyuan-family models with zero-shot decoder-only translators from the Hunyuan/HY-MT, Gemma, and Aya families, deliberately mixing encoder–decoder and LLM inductive biases; the full pool is enumerated in Appendix B.1 (§4).

2. Cross-system reference-free selection. We

use segment-level QE re-ranking over the cross-system pool under neural QE utilities (COMETKiwi, MetricX), and quantify its metric trade-offs on both official and internal evaluations (§4.1, §6).

3. **An entity-aware selection layer.** We introduce a Wikidata-derived bilingual gazetteer and an entity-recall floor that filters candidates omitting known proper names before QE selection. We report the development evidence behind it *per utility*, state exactly what was and was not measured before submission, and audit entity recall reference-free on the blind Challenge outputs (§4.1, §6.1, §7).

The paper reports only MAGADIR submissions and MAGADIR development experiments; we do not compare against other participating teams.

2 Related Work

Multilingual translation models. NLLB-200 scales multilingual neural MT to 200 languages and remains a strong encoder–decoder baseline for low-resource translation (NLLB Team et al., 2022). Recent decoder-only multilingual translators, including Hunyuan-MT and HY-MT1.5, have made LLM inference competitive for direct translation and zero-shot transfer (Zheng et al., 2025b,a). We also use Gemma 3 and Aya Expanse as open-weight multilingual instruction-tuned models with broad cross-lingual coverage (Gemma Team, 2025; Dang et al., 2024).

Parameter-efficient adaptation. LoRA adapts a pretrained model through low-dimensional update matrices rather than full-parameter training (Hu et al., 2022); QLoRA combines LoRA with low-bit quantization, enabling adaptation of large models at substantially lower memory cost (Dettmers et al., 2023). Our training experiments use both for NLLB- and Hunyuan-family candidates, while the final selection stage remains model-agnostic.

MBR decoding and QE re-ranking. MBR decoding selects the hypothesis with the best expected utility instead of the highest model probability (Eikema and Aziz, 2020), and neural-metric utilities improve translation quality when they correlate with human judgments better than log-probability or lexical overlap (Freitag et al., 2022). When the pairwise utility is replaced by a single reference-free QE score, the expected-utility

Split	Segments	Role in our pipeline
Train	21,000	LoRA/QLoRA adaptation only
Dev	1,000	
Devtest	500	training-time monitoring and qualitative analyses
Challenge	1,500×2	reference-based internal evaluation; selector ablations; choice of τ
		blind official scoring; reference-free selection; reference-free entity audit (§6.1)

Table 1: Official data splits (Laskar et al., 2026) and the strictly separated role of each split. Segment counts for the Challenge Test are per direction.

argmax reduces to QE re-ranking; this is the form we use, and it is the reference-free member of the quality-aware decoding family. In a blind shared-task setting reference-free utilities are required; we therefore rely on COMETKiwi (Rei et al., 2022b) and MetricX-24-QE (Juraska et al., 2024). We report official BLEU (Papineni et al., 2002; Post, 2018), chrF₂ (Popović, 2015), TER (Snover et al., 2006), COMET (Rei et al., 2022a), and BERT-style embedding metrics (Zhang et al., 2020).

Entity grounding. Named entities pose a distinctive challenge for Arabic–English translation: the correct output may require transliteration, an established exonym, or a conventional English label. Wikidata provides a multilingual entity graph connecting Arabic surface forms to English labels (Vrandečić and Krötzsch, 2014). We use it as a lightweight gazetteer rather than as a full retrieval-augmented generation system.

3 Task, Data, and Evaluation

Data. The official corpus, released by the shared-task organizers (Laskar et al., 2026), contains 21,000 training pairs, 1,000 development pairs, and 500 devtest pairs; the blind Challenge Test contains 1,500 source segments per direction. Table 1 summarizes the splits and the distinct role each one plays in our pipeline: only the training split touches model weights, the devtest split is the sole source of selection decisions (including the entity-floor threshold τ), and the Challenge split is scored exclusively by the organizers. This strict role separation matters in a reference-free setting, because any leakage of the selection split into training would inflate exactly the QE signals we later rely on for blind-set decisions.

Exploratory observations. Three properties of the data directly shaped our system design. First, the Arabic side exhibits substantial sur-

face variation for identical content — Unicode presentation-form variants, inconsistent whitespace, and definite-article attachment to borrowed names (البرنايو vs. بيرنايو) — motivating the conservative normalization pipeline applied to all data (Unicode normalization, whitespace normalization, removal of empty lines) and the article-aware matching of Appendix C.2. Second, paired source–target length-ratio checks were applied to guard against misaligned pairs, a standard failure mode of small parallel corpora. Third, development-set inspection showed a high density of named entities — people, clubs, venues, and competitions — whose English renderings are conventional rather than compositional; this observation, quantified later by the entity-recall column of Table 6, is what motivated the Wikidata gazetteer and the entity-recall floor. We did not train a new subword vocabulary; each model used its native tokenizer, which keeps the candidate pool heterogeneous at the segmentation level as well.

Challenge set. All six final output files contain exactly 1,500 lines; the AR→EN outputs were verified to contain no Arabic-script residue and the EN→AR outputs to contain Arabic script on every line (Appendix A).

Evaluation. The official metrics reported by the organizers (Laskar et al., 2026) are BLEU, chrF₂, TER, COMET, and an embedding metric (BERTScore for AR→EN, mBERT for EN→AR). Internally, we additionally used reference-based COMET-22 on the 500-segment devtest split and reference-free COMETKiwi/MetricX scores for model selection.

4 System Overview

Figure 1 summarizes the design. The same high-level recipe is used in both directions: generate candidates from multiple models, normalize and sanitize outputs, score candidates with reference-free metrics, and submit one primary plus two contrastive systems chosen as ablations of the primary (Table 2; §4.2 explains why one of the AR→EN ablations is not fully controlled). Algorithm 1 gives the complete selection procedure.

4.1 Selection Rule

Cross-pool QE re-ranking. Let $\mathcal{C}(x) = \bigcup_{m \in \mathcal{M}} \{y_m(x)\}$ be the pool of candidate translations of source x produced by the model set $\mathcal{M}$ (at

most one deterministic candidate per model), and let $Q(x, y)$ denote a reference-free QE score. Our core selector is the segment-level decision rule

$$\hat{y}_{\text{QE}}(x) = \arg \max_{y \in \mathcal{C}(x)} Q(x, y). \quad (1)$$

Equation (1) is QE re-ranking over a *cross-system* pool. It can be read as the limiting case of MBR decoding (Eikema and Aziz, 2020) in which the pairwise expected utility over pseudo-references is replaced by a single reference-free utility, and our submission identifiers (“MBR-COMETKiwi”) were coined from that reading; throughout the paper we use the plainer term *QE re-ranking*, as it describes exactly what Equation (1) computes. Both COMETKiwi and MetricX-24-QE appear as utilities Q in the submitted systems: the AR→EN primary uses MetricX-24-QE under the entity floor of Equation (4), while AR→EN Contrastive 2 and all EN→AR systems use COMETKiwi without a floor (§4.2, §4.3).

Entity-recall floor. Let $G(x)$ be the set of gazetteer entities whose Arabic surface forms occur in x , and let $\ell(e)$ be the canonical English label of entity e . We define surface entity recall

$$R_{\text{ent}}(x, y) = \frac{|\{e \in G(x) : \ell(e) \sqsubseteq y\}|}{|G(x)|}, \quad (2)$$

where $\sqsubseteq$ denotes normalized surface containment (Appendix C.2), and set $R_{\text{ent}}(x, y) := 1$ when $G(x) = \emptyset$. Because R_{ent} depends only on the source, the hypothesis, and the gazetteer, it can be computed on the blind Challenge outputs; we do so in §6.1. The entity-constrained selector restricts Equation (1) to the feasible set

$$\mathcal{F}_{\tau}(x) = \{y \in \mathcal{C}(x) : R_{\text{ent}}(x, y) \geq \tau\}, \quad (3)$$

$$\hat{y}(x) = \begin{cases} \arg \max_{y \in \mathcal{F}_{\tau}(x)} Q(x, y) & \mathcal{F}_{\tau}(x) \neq \emptyset, \\ \hat{y}_{\text{QE}}(x) & \text{otherwise,} \end{cases} \quad (4)$$

with $\tau = 0.85$ selected on the devtest split. The floor acts as a *precision-oriented veto*: it never forces an entity-matching candidate into first place on its own; it only removes candidates that demonstrably drop known entities before the QE utility decides. Two consequences follow directly from Equation (4) and are used in the analyses of §6.1 and §7. First, the floor can only change the decision on a segment where (a) $G(x) \neq \emptyset$, (b) the unconstrained winner has $R_{\text{ent}} < \tau$, and (c) at

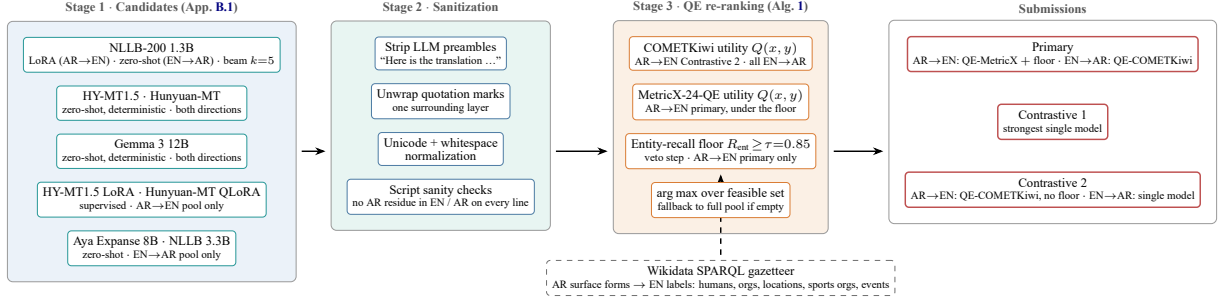


Figure 1: MAGADIR system pipeline. Candidate generation, sanitization, and reference-free scoring are shared across both directions; the dashed Wikidata entity-recall constraint is applied only to the AR→EN primary system. Both QE utilities are used in submitted systems: the AR→EN primary maximizes MetricX-24-QE over the floor-feasible set, while AR→EN Contrastive 2 and all EN→AR systems maximize COMETKiwi over the full pool (§4.2). The EN→AR pool is fully zero-shot (§4.3).

least one other candidate has $R_{\text{ent}} \geq \tau$; on every other segment it is the identity. Second, since R_{ent} takes values in $\{k/|G(x)|\}$, the threshold $\tau = 0.85$ is equivalent to *omit no known entity* for all $|G(x)| \leq 6$ ($5/6 = 0.833 < 0.85$) and tolerates exactly one omission only from $|G(x)| \geq 7$ ($6/7 = 0.857$). In the football-domain Challenge data, where segments rarely contain more than six recognized entities, the floor therefore behaves almost everywhere as a strict veto on any omission of a gazetteer-known name.

Choice of τ . $\tau = 0.85$ is the threshold adopted in the development run in which the floor was first introduced (Table 6, row “QE-MetricX + floor”). A devtest sweep over $\tau \in \{0.50, 0.75, 0.85, 1.00\}$ confirms the discretization argument above directly: $\tau = 0.85$ and $\tau = 1.00$ produce *byte-identical* output files (BLEU 27.69, chrF 56.78, COMET-22 0.8460, COMETKiwi 0.8226), $\tau = 0.75$ differs on a handful of segments (27.70 / 56.79 / 0.8459 / 0.8226), and only $\tau = 0.50$ moves the selector appreciably (27.54 / 56.84 / 0.8461 / 0.8217) by admitting candidates that drop half the recognized entities. Every setting improves COMETKiwi over the unconstrained MetricX selector (0.8192). The practical sensitivity to τ in this domain is therefore very low, and the choice was not re-tuned afterwards. Consistently with this, the packaged Challenge run was in fact invoked with a floor of 1.00 rather than 0.85; because the two are decision-equivalent here, the submitted file is exactly the one the $\tau = 0.85$ configuration would have produced (see also the Limitations).

Algorithm 1 Entity-aware cross-pool QE re-ranking (AR→EN primary)

Require: source x ; model set $\mathcal{M}$; QE utility Q ; gazetteer $\mathcal{G}$; threshold τ

- 1: $\mathcal{C}(x) \leftarrow \{\text{SANITIZE}(y_m(x)) : m \in \mathcal{M}\}$ ▷ Stage 2
- 2: $G(x) \leftarrow \{e \in \mathcal{G} : \text{some Arabic form of } e \text{ occurs in } \text{NORM}(x)\}$
- 3: **for all** $y \in \mathcal{C}(x)$ **do**
- 4: $q_y \leftarrow Q(x, y)$ ▷ MetricX-QE in the AR→EN primary; COMETKiwi elsewhere
- 5: $r_y \leftarrow R_{\text{ent}}(x, y)$ by Eq. (2) ($:= 1$ if $G(x) = \emptyset$)
- 6: **end for**
- 7: $\mathcal{F} \leftarrow \{y : r_y \geq \tau\}$
- 8: **if** $\mathcal{F} = \emptyset$ **then**
- 9: $\mathcal{F} \leftarrow \mathcal{C}(x)$ ▷ fallback: floor is the identity
- 10: **end if**
- 11: **return** $\arg \max_{y \in \mathcal{F}} q_y$

4.2 Arabic → English

The AR→EN candidate pool contains at most one candidate per model from the systems marked as pool members in Table 10 (Appendix B.1): the NLLB-200 distilled 1.3B LoRA adapter trained for two epochs (beam search, $k=5$), HY-MT1.5 zero-shot, Hunyuan-MT zero-shot, Gemma 3 12B zero-shot, and the two supervised LLM adapters (HY-MT1.5 LoRA and the Hunyuan-MT QLoRA adapter), giving $|\mathcal{M}| = 6$ and a median of six candidates per segment. Aya Expanse 8B and the five-epoch NLLB LoRA checkpoint were development-stage systems only and were *not* part of the submitted AR→EN pool. LLM candidates use deterministic (greedy) decoding with a translation-only

prompt and a maximum output length of 256 new tokens. We remove common preambles such as “Here is the translation” and strip one layer of surrounding quotation marks.

The primary AR→EN submission is **cross-pool QE re-ranking plus entity floor**, i.e. Equation (4). Contrastive 1 is HY-MT1.5 zero-shot after output cleaning, a single-model reference point, so that Primary–Contrastive 1 measures the value of cross-system pooling. Contrastive 2 is cross-pool QE-COMETKiwi without the entity floor (Equation (1)); it was intended to isolate the floor, but as explained next it also differs from the primary in QE utility.

Which utility the primary actually used. One property of the packaged primary requires an explicit correction, because it is not what the submission identifier suggests. The submitted file is named `cross_pool_mbr_cometkiwi_entity_floor`, but the entity-floor implementation that produced it scores the candidates surviving the floor with **MetricX-24-QE**, not COMETKiwi: the COMETKiwi reference in the filename comes from a `--cometkiwi_fallback` command-line flag that the script accepts but never reads, the fallback branch of Equation (4) being implemented as “re-admit all candidates and score them with the same MetricX utility”. In Equation (4) terms, the submitted primary is $Q = \text{MetricX-QE}$ with the floor applied, whereas Contrastive 2 is $Q = \text{COMETKiwi}$ with no floor. The Primary–Contrastive 2 delta therefore measures the floor *and* a change of QE utility together, and does not isolate the floor as originally claimed; §6.1 quantifies how large the resulting confound is. We report the systems as they were actually run rather than as they were labelled, and we treat the missing floor-only ablation on the official metrics as the main piece of unfinished business for this submission.

4.3 English → Arabic

The EN→AR candidate pool comprises HY-MT1.5, Hunyuan-MT, Gemma 3 12B, Aya Expanse 8B, NLLB-200 1.3B, and NLLB-200 3.3B, all zero-shot. The primary output is selected segment-by-segment with Equation (1); the contrastive systems are the two strongest single-model zero-shot runs in our internal evaluation (Aya Expanse 8B and Gemma 3 12B).

Why zero-shot despite the parallel data. Parallel data were available for both directions, and the decision not to fine-tune for EN→AR was a budget allocation rather than a methodological preference. Three factors drove it. (i) Under a single-GPU budget shared by both sub-tasks, adapter training and, more importantly, the candidate-generation and QE-scoring passes it multiplies, were prioritized for AR→EN, the direction in which the gazetteer’s canonical labels are on the target side and the entity floor is applicable. (ii) The Arabic-target references in the official corpus exhibit the surface variation described in §3; with 21K pairs, a LoRA adapter tends to learn that surface style (a benefit for BLEU) faster than it improves adequacy (what COMETKiwi sees), which is the over-specialization signature we observed for the Hunyuan-MT QLoRA adapter in the AR→EN direction (§7). (iii) The zero-shot decoder-only translators were already competitive under COMETKiwi on Arabic output, so the marginal value of an adapter was expected to be smaller than that of adding model diversity to the pool, which is what the official EN→AR deltas (Table 4) confirm *ex post*. We regard direction-specific EN→AR adaptation as the most obvious next step and list it as a limitation. We note for completeness that a reverse-direction NLLB-600M LoRA adapter was in fact trained (`nllb600-reverse-en-ar-lora`) but was never decoded or scored on a development set, so point (iii) rests on the zero-shot comparison alone and not on a measured fine-tuned EN→AR baseline.

5 Experimental Setup

Fine-tuning. For NLLB, we trained LoRA adapters on the official parallel data with rank $r = 16$, scaling $\alpha = 32$, dropout 0.05, AdamW optimization, cosine learning-rate scheduling, bf16 precision, and gradient accumulation. Hunyuan/HY-MT-family adaptation used (Q)LoRA with 4-bit NF4 quantization and LoRA modules on attention and feed-forward projections. These adapters were useful during development and two of them are in the submitted AR→EN pool, but the final EN→AR system relies on zero-shot candidates rather than direction-specific fine-tuning (§4.3). Full hyperparameters are listed in Appendix B.

Decoding and scoring. NLLB uses beam search with beam size 5; LLM inference uses determinis-

Task	Direction	Submission	System description
Sub-Task 2A	AR→EN	Primary	Cross-pool QE-MetricX with Wikidata entity-recall floor (Eq. 4); submission ID <code>cross_pool_mbr_cometkiwi_entity_floor</code> (the identifier names COMETKiwi, but the selector is MetricX-QE — see §4.2)
Sub-Task 2A	AR→EN	Contrastive 1	HY-MT1.5 zero-shot, cleaned translation-only output
Sub-Task 2A	AR→EN	Contrastive 2	Cross-pool QE-COMETKiwi without entity-recall floor (Eq. 1)
Sub-Task 1A	EN→AR	Primary	Cross-pool QE-COMETKiwi over zero-shot multilingual candidates
Sub-Task 1A	EN→AR	Contrastive 1	Aya Expanse 8B zero-shot
Sub-Task 1A	EN→AR	Contrastive 2	Gemma 3 12B zero-shot

Table 2: MAGADIR submission mapping. Contrastive 1 removes the cross-system pool and is a clean single-model reference point. Contrastive 2 was intended to remove only the entity floor, but also differs from the primary in QE utility (COMETKiwi against MetricX-24-QE), so it is not a controlled floor ablation; §6.1 separates the two effects reference-free. Submission identifiers retain the historical “MBR” naming (§4.1). The table lists only MAGADIR systems.

tic decoding (`num_beams=1`, `temperature=0.0`, `do_sample=False`) with a fixed prompt and 256 maximum new tokens (Table 11). Candidate scoring uses `Unbabel/wmt22-cometkiwi-da` for COMETKiwi and `google/metricx-24-hybrid-xl-v2p6-bfloat16` for MetricX-QE. All selection is segment-level; we did not use document-level context in the submitted systems.

Entity gazetteer. The Wikidata gazetteer maps Arabic surface forms to English labels for humans, organizations, locations, sports organizations, and events, retrieved with type-restricted SPARQL queries over Arabic labels and aliases. We used it as a precision-oriented filter rather than a recall-complete named-entity recognizer. Development logs showed that remote SPARQL access can be rate-limited, so the resulting gazetteer should be interpreted as a partial-coverage resource (Appendix C).

6 Official MAGADIR Results

Table 3 reports all MAGADIR Challenge Test submissions, Figure 2 plots them in BLEU–COMET space, and Table 4 reports the pairwise deltas between submissions.

AR→EN. The primary entity-floor system obtains the highest official COMET among our AR→EN submissions (81.94). Compared with the no-floor cross-pool contrastive it gains +0.37 COMET and +0.48 chrF₂ while losing 0.60 BLEU

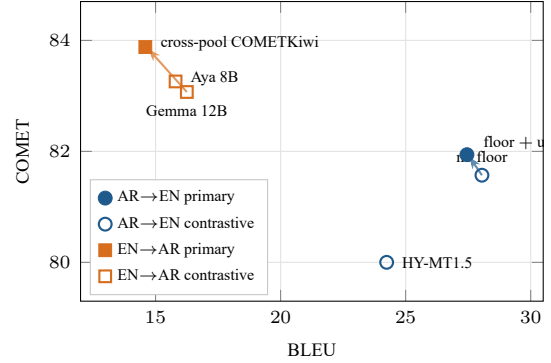


Figure 2: Official MAGADIR submissions in BLEU–COMET space. Arrows show the movement induced by the primary selectors: both trade lexical overlap for neural adequacy, i.e. they move up-left. The AR→EN arrow combines the entity floor and a change of QE utility (§4.2).

and increasing TER by 1.88 (Table 4). This delta must be read with care: the two systems differ both in the entity floor and in the QE utility (MetricX-24-QE against COMETKiwi; §4.2), so it is a joint effect and we do not attribute it to the floor. Reference-free evidence that *does* separate the two components is given in §6.1, and it assigns the adequacy cost to the utility rather than to the veto. What the row does show is the direction of the combined change: selected outputs move in a direction favored by the neural adequacy metric but not uniformly by lexical metrics. Against the single-model contrastive, cross-system pooling is uniformly positive (+3.21 BLEU, +1.94 COMET), confirming that model diversity, not any

Task	Direction	Submission	System	BLEU	chrF ₂	TER↓	COMET	Emb.
Sub-Task 2A	AR→EN	Primary	Cross-pool QE-MetricX + entity floor	27.45	56.61	64.49	81.94	81.28
Sub-Task 2A	AR→EN	Contrastive 1	HY-MT1.5 zero-shot	24.24	54.49	67.69	80.00	80.00
Sub-Task 2A	AR→EN	Contrastive 2	Cross-pool QE-COMETKiwi	28.05	56.13	62.61	81.57	81.57
Sub-Task 1A	EN→AR	Primary	Cross-pool QE-COMETKiwi	14.59	51.48	80.85	83.88	83.13
Sub-Task 1A	EN→AR	Contrastive 1	Aya Expanse 8B zero-shot	15.80	51.43	77.63	83.26	83.43
Sub-Task 1A	EN→AR	Contrastive 2	Gemma 3 12B zero-shot	16.25	51.53	76.94	83.07	83.22

Table 3: Official MAGADIR-only Challenge Test results. “Emb.” denotes BERTScore for AR→EN and mBERT for EN→AR; lower is better for TER. Bold marks the best MAGADIR submission per direction and metric. In both directions the QE-guided primary wins COMET while a contrastive system wins BLEU, the central trade-off analyzed in §8.

Comparison	ΔBLEU	ΔchrF ₂	ΔTER	ΔCOMET	ΔEmb.
<i>AR→EN</i>					
Floor + utility (P – C2)	−0.60	+0.48	+1.88	+0.37	−0.29
Pooling effect (P – C1)	+3.21	+2.12	−3.20	+1.94	+1.28
<i>EN→AR</i>					
Pooling (P – Aya C1)	−1.21	+0.05	+3.22	+0.62	−0.30
Pooling (P – Gemma C2)	−1.66	−0.05	+3.91	+0.81	−0.09

Table 4: Pairwise deltas between official submissions (P = primary; positive ΔTER is worse). The AR→EN P–C2 row is *not* a floor ablation: the two systems differ in the entity floor *and* in the QE utility (§4.2), so it measures both together and cannot be attributed to either; §6.1 separates them reference-free. Cross-system pooling over a weak single baseline (HY-MT1.5) improves every metric simultaneously.

individual translator, drives the AR→EN result.

EN→AR. The primary cross-pool COMETKiwi system obtains the highest official COMET among our EN→AR submissions (83.88). The Gemma contrastive has higher BLEU and lower TER, but the primary gains +0.81 COMET relative to it. This is consistent with primary selection optimizing reference-free semantic quality rather than reference n -gram overlap: with all-zero-shot candidates whose wording diverges from reference style, QE selection favors adequate paraphrases that lexical metrics undercount.

6.1 Reference-Free Entity Audit of the Blind Outputs

Because R_{ent} (Equation 2) needs only the source, the hypothesis, and the gazetteer, it is the one entity-preservation statistic that *can* be computed on the Challenge Test, and it measures the floor under exactly the utility that was submitted. Table 5 reports the protocol of Algorithm 2 for the three AR→EN submissions and one post-hoc control. Of the 1,500 blind sources, 773 (51.5%) contain at least one of the 93 Arabic surface forms the

gazetteer knows, for 1,113 entity occurrences in total; the remaining 727 segments have $G(x) = \emptyset$, where Equation (4) is provably the identity.

On the quantity it controls, the floor behaves as intended: the Primary reaches $\bar{R}_{\text{ent}} = 0.4264$ against 0.3931 for the unconstrained cross-pool selector and 0.4038 for the single-model HY-MT1.5 contrastive, and renders *every* recognized entity on 35.2% of entity-bearing segments against 31.8%. The absolute level is moderate for the reason developed in Appendix C: the gazetteer stores one canonical English label per entity, so any acceptable variant rendering is scored as a miss. The gap between systems, not the level, is the quantity of interest.

Separating the floor from the utility change.

The audit also recovers the floor’s *action set*. Partitioning the 773 entity-bearing segments by the state of $\mathcal{F}_\tau(x)$: on 501 (64.8%) no candidate clears τ and the fallback of Equation (4) fires; on 206 (26.7%) every candidate clears it and the floor is inert; and on only **66 (8.5%)** does the floor remove at least one candidate while leaving another, which is the sole configuration in which it can change a decision. The partition is identical at $\tau = 0.85$ and $\tau = 1$, as §4.1 predicts. At the candidate level the veto is correspondingly light: it removes 172 of the 8,734 pooled hypotheses and retains 98.0%. That retention figure is recorded independently in the submission job log, and our audit reproduces it exactly, which confirms that the matching used here agrees with the matching the selector actually applied.

This makes the raw intervention rate interpretable, and it is far larger than a veto can explain: the Primary and Contrastive 2 files differ on 1,087 of 1,500 segments (72.5%) while the floor could act on 66. The explanation is the one identified in

§4.2 — the packaged Primary also switched QE utility. To separate the two we scored the same pool with the Primary’s utility and the floor disabled, giving the control C2’ of Table 5. Since C2’ shares its utility with the Primary and its floor setting with Contrastive 2, the difference decomposes exactly:

$$\underbrace{P - C2'}_{\text{floor}} = +0.0276, \quad \underbrace{C2' - C2}_{\text{utility}} = +0.0057,$$

summing to the +0.0333 of $\bar{R}_{\text{ent}}$ that the Primary gains over Contrastive 2. **The floor contributes 83% of the entity improvement and the utility switch 17%**, even though the utility switch is responsible for essentially all of the output churn: C2’ differs from Contrastive 2 on 74.1% of entity-bearing segments, whereas the Primary differs from C2’ on only 7.1%.

This is the paper’s central quantitative claim about the mechanism, and it is worth stating plainly: *the floor is a surgical intervention*. It rewrites a twentieth of the output while delivering the large majority of the entity gain, whereas changing the QE utility rewrites three quarters of the output for a sixth of it. Restricting to the 66-segment action set makes the concentration sharper still: the 26 segments on which the floor actually overrides the unconstrained winner supply +20.8 of the +21.3 segment-equivalents of the floor-only gain $P - C2'$, or 97.7%.

The floor’s adequacy cost is not the floor’s. The same control settles a second question, and reverses the natural reading of Table 5. The Primary scores 0.8168 COMETKiwi against Contrastive 2’s 0.8251, a drop of 0.0083 that would naturally be charged to the entity floor. Decomposed against C2’ (0.8166), the drop is almost entirely elsewhere:

$$\underbrace{P - C2'}_{\text{floor}} = +0.0002, \quad \underbrace{C2' - C2}_{\text{utility}} = -0.0085.$$

The floor is *free* on this axis — marginally positive, in fact — and the entire adequacy cost comes from selecting with MetricX-QE instead of COMETKiwi. In hindsight this is what one should expect: Contrastive 2 selects the argmax of COMETKiwi itself, so any move away from that utility must lose COMETKiwi almost by definition, whereas the floor only removes candidates and leaves the utility to choose among the rest.

Combining the two decompositions gives the paper’s clearest statement about the mechanism. The entity floor buys +0.0276 entity recall for +0.0002 COMETKiwi, changing 7% of segments; the utility switch buys +0.0057 entity recall for −0.0085 COMETKiwi, changing 74%. The floor is close to a free improvement on the axis it targets, and the component that actually cost adequacy is the one adopted by accident (§4.2).

A caveat on the control. The Primary and C2’ differ on 77 segments, of which only 26 lie inside the floor’s action set; on the other 51 the floor is provably the identity, so those differences cannot be the floor. They reflect run-to-run variation in MetricX scoring between two separately submitted jobs, whose candidate batches differ. The variation is entity-neutral — it accounts for +0.5 of the +21.3 segment-equivalents, about 2% — so it does not disturb the decomposition, but it does mean the +0.0276 attributed to the floor should be read with a tolerance of roughly ± 0.001 rather than as an exact quantity.

Taken with the controlled devtest ablation of §7, which varies the floor against a fixed utility and pool and moves in the same direction, this establishes the floor’s behaviour from two independent directions and on two test sets. The same analysis bounds the claim. The veto is inert on the 64.8% of entity-bearing segments where no pooled candidate preserves every recognized entity, so its reach is set by pool diversity and gazetteer coverage rather than by the threshold. And the decomposition above is available only for R_{ent} and COMETKiwi, which are reference-free: the official metrics of Table 4 require the withheld references, so C2’ cannot be scored on them and those deltas remain a joint measurement of floor and utility. We report them as such.

7 Development Experiments and Ablations

Table 6 consolidates the key AR→EN development experiments, and Figure 3 plots the resulting adequacy–entity landscape. These experiments used the 500-segment devtest split with references and served internal selection and ablation, not official scoring. Every cross-pool row states the QE utility it was scored under and whether the floor was applied.

System	$\bar{R}_{\text{ent}}$	$R_{\text{ent}=1}$ (%)	Chg. (%)	COMETKiwi
P: floor+MetricX	0.4264	35.2	73.1	0.8168
C2': MetricX	0.3988	32.0	74.1	0.8166
C2: COMETKiwi	0.3931	31.8	—	0.8251
C1: HY-MT1.5	0.4038	32.6	77.7	0.8146

Table 5: Reference-free entity audit (Algorithm 2) on the 773 AR→EN Challenge segments with $G(x) \neq \emptyset$. “Chg.” is the share of those segments whose output differs from Contrastive 2. C2' is a post-hoc control, not a submitted system: it applies the Primary's utility (MetricX-24-QE) to the same pool with the floor disabled, completing the 2×2 that separates the floor from the utility switch (§6.1).

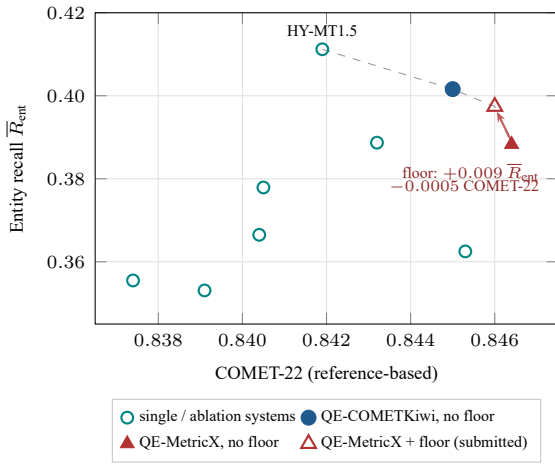


Figure 3: Development-set relation between reference-based COMET-22 and entity recall $\bar{R}_{\text{ent}}$, on the uniform scale of Table 6. Both red points use the MetricX-QE utility. The hollow triangle is the same selector with the entity floor applied, which is the configuration submitted as the AR→EN primary: the floor lifts it by +0.009 $\bar{R}_{\text{ent}}$ at a COMET-22 cost of 0.0005 (arrow), to within 0.004 of the COMETKiwi selector. The dashed line joins the non-dominated systems.

Selection lessons. The highest-COMETKiwi development system was cross-pool QE-COMETKiwi (0.8289), which also has the second-highest entity recall (0.4016, behind HY-MT1.5 zero-shot at 0.4112) and lies on the adequacy–entity front of Figure 3; this position motivated its use as the core selector. Hunyuan-MT QLoRA achieved by far the best BLEU/chrF on the devtest split (33.56/60.10) but the *lowest* COMETKiwi (0.7993), a signature of over-specialization to the small supervised sample: its wording matches the references it was tuned toward while the reference-free metric is less impressed. The MetricX selector produced

the strongest COMET-22 (0.8464). During development, the surface entity diagnostic then in use indicated a drop of ~ 0.33 in entity preservation relative to the COMETKiwi selector over the *same* pool; because that diagnostic relied on exact gazetteer labels, we treated the drop as a warning signal rather than a definitive quality measure. That caution proved well founded: the drop was almost entirely an artifact of the diagnostic, and on the uniform scale of Table 6 the true gap is 0.0133 (next paragraph). It was nevertheless this apparent contrast — two utilities, one pool, opposite verdicts on entities — that motivated the entity floor of Equation (4) at the time, and we describe the design as it was actually driven rather than retrofitting a rationale. An additive entity *reward* variant (row “QE-COMETKiwi + additive entity reward”) was also tried and rejected: it raised BLEU but lowered COMETKiwi, COMET-22, and entity recall itself, which is the empirical basis for the veto-not-reward design argued in Appendix C.1.

The entity “collapse” was a measurement artifact. The diagnostic used during development reported entity preservation collapsing from 0.48 to 0.15 under MetricX. That reading does not survive scrutiny. The COMETKiwi rows and the MetricX/floor rows had not been scored with the same diagnostic: the later implementation required *all* ranked Wikidata candidates for an Arabic surface form to appear simultaneously, and additionally injected a non-matchable `wikidata_search` sentinel into every expected set — a column-indexing defect, that value being the gazetteer’s `match_type` field rather than an alias list. The expected set grew from 157 items to 1,246 while correctly rendered entities *rose* from 149 to 184, so a strictly better system scored far worse.

Re-scoring every row of Table 6 with one gazetteer and the selector’s own matching code removes the artifact. Under $\bar{R}_{\text{ent}}$, the COMETKiwi selector reaches 0.4016 and the MetricX selector 0.3883: a real but small gap of 0.0133, not 0.337. Roughly 96% of the apparent collapse was measurement. The floor then recovers about two thirds of what remains, lifting MetricX to 0.3974. Two conclusions follow. The motivation for the floor was sound but the magnitude that motivated it was inflated; and the entity column is now comparable across all rows, so it can be read as a single scale rather than only within blocks.

Family	System	Utility	Floor	BLEU	chrF	COMET-22	COMETKiwi	$\bar{R}_{\text{ent}}$
NLLB	NLLB-1.3B LoRA, 2 epochs	—	No	30.22	57.38	0.8405	0.8142	0.3779
NLLB+QE	NLLB-1.3B LoRA, 5 epochs, n -best QE re-ranking	COMETKiwi	No	29.63	57.02	0.8432	0.8214	0.3887
LLM	HY-MT1.5 zero-shot	—	No	23.77	54.67	0.8419	0.8167	0.4112
LLM+LoRA	Hunyuan-MT QLoRA	—	No	33.56	60.10	0.8453	0.7993	0.3625
LLM	Gemma 3 12B zero-shot	—	No	26.46	54.58	0.8391	0.8146	0.3531
LLM	Aya Expanse 8B zero-shot	—	No	25.86	53.69	0.8374	0.8131	0.3555
Cross-pool	QE-COMETKiwi [Contrastive 2]	[= COMETKiwi	No	27.55	56.04	0.8450	0.8289	0.4016
Cross-pool	QE-COMETKiwi + additive entity reward	COMETKiwi+rew.	No	29.79	57.06	0.8404	0.8131	0.3665
Cross-pool	QE-MetricX	MetricX	No	27.40	57.16	0.8464	0.8192	0.3883
Cross-pool	QE-MetricX + floor ($\tau=0.85$) [= submitted config.]	MetricX	Yes	27.69	56.78	0.8460	0.8226	0.3974

Table 6: Selected internal AR→EN development results (devtest, 500 segments). *Utility* is the reference-free score maximized by the selector; *Floor* marks whether Eq. (4) was applied. Bold marks the column best. The entity column is $\bar{R}_{\text{ent}}$ (Eq. 2) over the 278 devtest segments with $G(x) \neq \emptyset$, recomputed for *every* row with a single gazetteer and the selector’s own matching code, so one scale applies across the whole table (§7). The entity-floor implementation scores surviving candidates with MetricX-24-QE, so the row corresponding to the submitted primary is the MetricX+floor row (§4.2); the last two rows differ only in the floor and thus isolate its effect. The cross-pool rows use the wide devtest ablation pool of Appendix B.1, not the six-system Challenge pool.

The floor in isolation. The row “QE-MetricX + floor” of Table 6 applies the entity floor on top of the MetricX-QE utility, not COMETKiwi, and it is the development counterpart of the submitted AR→EN primary, which uses the same utility-plus-floor combination (§4.2); the two differ only in the candidate pool, six Challenge systems rather than the wider devtest ablation pool of Appendix B.1. No COMETKiwi-plus-floor system was built, and no claim in this paper rests on one. Because the last two rows of Table 6 share a pool *and* a utility and differ only in the floor, they form a controlled ablation whose effect is directly readable: $+0.0091 \bar{R}_{\text{ent}}$ and $+0.0034$ COMETKiwi, at -0.38 chrF, $+0.29$ BLEU, and essentially no change in COMET-22 (-0.0005). In Figure 3 this is the arrow that lifts the MetricX selector onto the adequacy–entity front, almost to the entity level of the COMETKiwi selector, at a negligible COMET-22 cost and a small chrF cost. On the blind Challenge set, the matched control C2’ of §6.1 completes the corresponding 2×2 and shows the same pattern: the floor supplies 83% of the entity-recall gain over Contrastive 2 at $+0.0002$ COMETKiwi, and the utility switch carries the -0.0085 COMETKiwi cost. The official reference-based deltas of Table 4 ($+0.37$ COMET, $+0.48$ chrF₂, -0.60 BLEU) are the one measurement that still mixes floor and utility, because C2’ cannot be scored against the withheld references. Appendix D illustrates on a constructed example

when the floor can and cannot change a decision.

How often the floor could act. The devtest split admits the same partition as the blind set, recomputed with the matcher of Appendix E so that the two are directly comparable.¹ Of the 500 devtest segments, the gazetteer recognizes at least one entity on **278** (55.6%); the distribution is 181 segments with one entity, 68 with two, 25 with three, 3 with four and 1 with five, and 222 segments with none, on which the floor is the identity by definition. Among the 278 entity-bearing segments at $\tau = 0.85$, *no* candidate clears the threshold on **179** (64.4%) and the fallback branch fires; *every* candidate clears it on 57 (20.5%), where the feasible set is the whole pool and the veto is inert; and only on **42** (15.1% of entity-bearing segments, 8.4% of all) does the floor remove at least one candidate while leaving another, which is the sole configuration in which it can change a decision. The devtest and Challenge fallback rates agree closely (64.4% against 64.8%, §6.1), so the ceiling on the veto’s reach is a property of the gazetteer and the candidate pool rather than of either test set. That the fallback is the majority outcome on entity-bearing segments is the single most informative number

¹The original devtest run logged a slightly different split (158 fallback, 120 floor-applied) because that script additionally normalized punctuation away before matching, which admits a few more candidates. We report the recomputed figures so that every number in this paper uses one matcher; the entity extraction itself is identical, reproducing the logged entity-count distribution exactly.

System	BLEU raw	BLEU norm.	Δ BLEU	Δ chrF
NLLB-1.3B LoRA	52.83	53.43	+0.60	+0.37
Ensemble-COMETKiwi	48.91	49.69	+0.78	+0.47
NLLB-3.3B LoRA	48.32	49.40	+1.08	+0.69
Chimera refined	38.10	44.29	+6.19	+2.66
Hunyuan-MT LoRA	39.56	39.86	+0.30	+0.29
HY-MT1.5 LoRA	34.93	35.76	+0.83	+0.41

Table 7: Custom-article normalization diagnostic from the entity-aware experiment, computed on that experiment’s internal evaluation subset (absolute scores are therefore not comparable with Table 6; only the deltas are of interest). “Norm.” applies the article and Unicode normalization used in all later experiments.

Error type	Behavior	Mitigation
Entity omission / paraphrase	High-QE candidate silently drops or rewrites a proper name	Entity-recall floor (Eq. 4)
Hallucinated proper name	Fluent, English-looking name absent from the source	Not vetoed by the floor (recall-only); flagged for future precision check
Transliteration variant	Output name differs from reference spelling but is not semantically wrong	Treated as warning only; motivates transliteration-aware equivalence

Table 8: Entity error taxonomy observed in the development experiments, with the mitigation adopted (or deferred) in the submitted systems. Note that R_{ent} is a recall statistic: it cannot penalize hallucinated names, only omitted ones.

about this component: more often than not the pool simply contained no candidate rendering every recognized entity in canonical surface form, so the veto had nothing to select and reduced to the identity. This is consistent with the wide but low-diversity devtest pool described in Appendix B.1, where an NLLB n -best list supplies most candidates and near-duplicates tend to share the same omissions, and it is the strongest argument in the paper for transliteration-aware matching (Limitations) rather than a stricter threshold.

Normalization and entity diagnostics. An additional custom-article experiment measured the effect of Arabic normalization and article handling. Table 7 shows that normalization improved every tested system, with the largest gain (+6.19 BLEU, +2.66 chrF) for the chimera-style post-edited system whose outputs contain the most surface variation: normalization is a cheap, system-agnostic win to apply before any metric comparison. The same experiment identified two frequent entity-related error types, summarized with the omission failure in Table 8.

8 Discussion

Metric trade-offs. The official results and dev experiments show our systems fall into two broad categories. Single-model systems and some LoRA systems score strongly on BLEU/chrF when their wording matches the reference style. Cross-pool selection, especially under COMETKiwi, favors semantically adequate paraphrases and therefore improves COMET more reliably than lexical metrics; graphically, the primary selectors move up-left in Figure 2. This explains why the primary systems were not always the highest-BLEU MAG-ADIR submissions, and mirrors the deliberate positioning of QE-guided decoding in prior quality-aware pipelines (Freitag et al., 2022).

What the evidence for the floor does and does not show. Three facts are established, all under the MetricX utility that the primary actually used. (i) On devtest, over one fixed pool, the floor moves every reference-free measure in the intended direction: +0.0034 COMETKiwi and +0.0091 $\bar{R}_{\text{ent}}$, at −0.38 chrF and essentially no change in COMET-22 (Table 6). (ii) On the blind Challenge set, against a matched no-floor control, the floor supplies +0.0276 of the +0.0333 entity-recall gain over Contrastive 2 — 83% — while altering only 7% of segments. (iii) On the same comparison the floor costs +0.0002 COMETKiwi, i.e. nothing; the −0.0085 that separates the two submitted systems is the utility switch, not the veto.

What is *not* established is a decomposition of the *official, reference-based* deltas of Table 4 (+0.37 COMET, +0.48 chrF₂, −0.60 BLEU, +1.88 TER), because the control cannot be scored against the withheld references. Those remain a joint measurement of floor and utility. Given (iii), however, we would caution against reading the lexical cost as the floor’s price: on every axis we *can* decompose, the floor is the cheap component and the utility switch is the expensive one. Where the floor does move lexical metrics, the mechanism is visible — it removes candidates that paraphrase or drop a name, and survivors carry the canonical Wikidata label, often longer than the reference’s rendering (*Real Madrid CF* vs. *Real Madrid*), which lengthens the hypothesis and raises TER while COMET and character-level chrF₂ credit the preserved content.

Named entities. The Wikidata entity floor was designed to reduce severe omission of names,

places, and organizations. The entity diagnostic exposed its limits: it is sensitive to transliteration variants, partial names, and incomplete gazetteer coverage (Table 8), and R_{ent} is a recall statistic that cannot penalize hallucinated names. Future work should combine Wikidata labels with fuzzy transliteration matching and model-based entity equivalence on both the recall and precision sides.

Direction asymmetry. AR→EN benefited from supervised adapters and entity labels with English targets. EN→AR relied on zero-shot model diversity plus COMETKiwi selection for the reasons given in §4.3. Its relatively high COMET among MAGADIR submissions indicates that cross-pool selection is useful even without direction-specific fine-tuning, while the lexical metrics suggest room for Arabic-target style tuning.

9 Conclusion

MAGADIR’s WMT 2026 English–Arabic submissions combine open-weight multilingual translators, parameter-efficient adaptation, cross-system reference-free QE re-ranking, and a Wikidata-based entity veto. The official MAGADIR-only results show that QE-guided primary systems achieved the strongest COMET in both directions (+0.37 over the no-floor cross-pool contrastive and +1.94 over the single-model baseline for AR→EN; +0.81 over the best single system for EN→AR), while contrastive systems sometimes achieved stronger lexical-overlap metrics. Because the AR→EN primary and its contrastive differ in QE utility as well as in the floor, we isolate the veto reference-free instead (§6.1): it supplies 83% of the entity-recall gain at no COMETKiwi cost, while the utility switch carries the adequacy loss. Two technical lessons follow. Blind-set selection in low-resource MT should not rely on a single metric, since neural QE, lexical metrics, and entity-preservation diagnostics each expose different failure modes. And an ablation is only as trustworthy as the bookkeeping behind it: our own floor-versus-utility confound, and an entity diagnostic whose expected-label convention changed midway, were both invisible until the released files were re-audited end to end. Future work will improve Arabic-target adaptation, expand the entity gazetteer, and replace exact entity matching with transliteration-aware equivalence.

Limitations

First, the AR→EN primary and Contrastive 2 differ in *two* respects rather than one: the primary applies the entity floor *and* selects with MetricX-24-QE, whereas Contrastive 2 selects with COMETKiwi and applies no floor (§4.2). The floor is nonetheless isolated three times in this paper — as a controlled devtest ablation against a fixed utility and pool (§7), via its action set on the blind outputs, and against the post-hoc C2’ control that completes the 2×2 on the Challenge pool (§6.1) — and all three agree in sign and rough magnitude, on both entity recall and COMETKiwi. What the confound still prevents is attributing the *official, reference-based* deltas of Table 4 to one component, because C2’ cannot be scored against the withheld references; that decomposition is available to the organizers but not to us. Since every axis we can decompose assigns the cost to the utility rather than the floor, we expect the same of the reference-based deltas, but we have not shown it. C2’ was produced after the evaluation period and is a diagnostic, not a submission. Relatedly, the packaged primary was invoked with a floor of 1.00 rather than the devtest-selected 0.85; the two are decision-equivalent on every segment with at most six recognized entities, which covers all devtest and Challenge segments we observed, but the equivalence is a property of this domain rather than of the method.

Second, the entity comparison that originally motivated the floor was inflated by a defective diagnostic: re-scoring every row of Table 6 under one gazetteer reduces the COMETKiwi-vs-MetricX entity gap from 0.337 to 0.0133 (§7). The floor remains beneficial but its motivating evidence was weaker than we believed at submission time. $\bar{R}_{\text{ent}}$ itself measures exact surface containment of gazetteer labels, so it undercounts legitimate transliteration variants and partial names; the reported values are lower bounds on semantic entity preservation, and comparisons between systems with very different output styles should be read with that caveat. R_{ent} is also recall-only and blind to hallucinated names.

Third, and most substantively, the floor is inactive where it is most needed: on 64.8% of entity-bearing Challenge segments (64.4% on devtest) no pooled candidate renders every recognized entity in canonical form, so the veto has nothing to select and reduces to the identity. Raising that ceiling

requires transliteration-aware matching or a more diverse pool, not a different threshold. The Wiki-data gazetteer is partial-coverage due to SPARQL rate limits (93 Arabic surface forms), and the entity floor consequently cannot veto omissions of entities it does not know.

Finally, all selection is segment-level and reference-free at submission time; QE models carry their own biases, and optimizing toward them can inflate COMET-family scores relative to human preference. The EN→AR systems are zero-shot for budget reasons rather than because adaptation was shown to be unhelpful. Our conclusions are drawn from a single low-resource English–Arabic task with 500 devtest segments, and the internal comparisons are MAGADIR-only.

Acknowledgements

We thank the WMT 2026 shared-task organizers for releasing the task data and official evaluation outputs, and the two anonymous reviewers, whose questions about the development evidence for the entity floor led directly to §6.1 and the reconciliation in §7. We also thank the maintainers of the open-source model, metric, and knowledge-base resources used in this work.

References

- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expand: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Firoj Alam, Bishwaraj Paul, Irfan Ahmad, Maya Silvi Lydia, and Pankaj Kundan Dadure. 2026. Findings of the WMT 2026 shared task on low-resource arabic-asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T.

Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, Massachusetts, USA.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025a. HY-MT1.5 technical report. *arXiv preprint arXiv:2512.24092*.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025b. Hunyuan-MT technical report. *arXiv preprint arXiv:2509.05209*.

A Submission Packaging and Validation Protocol

The final packaging stage enforced a four-layer validation protocol before any file was accepted into a submission slot:

1. **Structural integrity:** every output file must contain exactly 1,500 non-empty lines, one hypothesis per source segment, in source order.
2. **Script consistency:** AR→EN files must contain no residual Arabic-script characters (a Unicode-block regular expression scan over every line), and EN→AR files must contain Arabic script on *every* line, catching untranslated pass-throughs and empty-hypothesis failures.
3. **Encoding hygiene:** all files are UTF-8 with NFC-normalized Unicode, matching the normalization applied during preprocessing so that the organizers’ scorers and our internal scorers see identical strings.
4. **Cross-slot distinctness:** the three files per direction are checked pairwise for non-identity,

ensuring that the primary and contrastive slots encode genuinely different systems rather than redundant copies of the same selection.

Table 9 maps each submission to its output file and summarizes the outcome. All checks passed for all six files.

B Candidate Pool, Training, Decoding, and Selection Hyperparameters

B.1 Candidate Pool Inventory

Table 10 enumerates every candidate system that could be selected by Equation (1) in each direction, answering the question of exactly which hypotheses were available to the selector. Each model contributes at most one deterministic candidate per segment. Candidate generation was not loss-free, however, so $|\mathcal{C}(x)| = |\mathcal{M}|$ holds only on the majority of segments rather than on all of them: in AR→EN, $|\mathcal{C}(x)| = 6$ on 1,264 of the 1,500 Challenge segments, 5 on 214, and ≤ 4 on the remaining 22 (minimum 2); in EN→AR, $|\mathcal{C}(x)| = 6$ on 1,336 segments, 5 on 128, and ≤ 4 on 36 (minimum 1). The losses come from empty or non-parsing generations that the sanitizer discarded, concentrated in `hytm15_lora` (1,326/1,500 surviving) and `hunyuan7b_zshot` (1,427) for AR→EN, and in `hunyuan7b_zshot_enar` (1,381) for EN→AR. Equations (1) and (4) range over whatever candidates survived, so no segment is ever left without a hypothesis.

The devtest ablation pool is a different pool.

The submitted Challenge pools of Table 10 must not be confused with the pool over which the development ablations of Table 6 were run. The 500-segment devtest ablation pool (`cross_pool.devtest.jsonl`) contains **13,455 candidates**, i.e. a median of **27** candidates per segment (range 18–29; $|\mathcal{C}(x)|=27$ on 266 of 500 segments, 28 on 79, 29 on 52, 26 on 51, ≤ 25 on 52). It draws on eleven sources rather than six, and is dominated by an NLLB *n*-best list rather than by one candidate per model: `nllb_pool` (9,915 candidates, an *n*-best list contributing ~ 20 hypotheses per segment), `hunyuan7b_zshot` (494), `aya8b_zshot` (488), `hytm15_lora` (487), `gemma3_12b_zshot` (487), `chimera_refined` (469), `hytm15_zshot` (464), `nllb33b_lora` (448), `nllb13b_baseline` (123), `nllb13b_avg5` (76), and `nllb13b_mbr_qe` (4). Two consequences matter for reading Table 6. First, the ablation numbers are *not* a like-for-like

Direction	Submission	Output file	Lines	Script check	Distinct
AR→EN	Primary	cross_pool_mbr_cometkiwi_entity_floor.test.en	1,500	no Arabic residue	Yes
AR→EN	Contrastive 1	hymt15_zshot.test.clean.en	1,500	no Arabic residue	Yes
AR→EN	Contrastive 2	cross_pool_mbr_cometkiwi.test.en	1,500	no Arabic residue	Yes
EN→AR	Primary	cross_pool_mbr_cometkiwi.test.ar	1,500	Arabic on every line	Yes
EN→AR	Contrastive 1	aya8b_zshot_enar.test.ar	1,500	Arabic on every line	Yes
EN→AR	Contrastive 2	gemma3_12b_zshot_enar.test.ar	1,500	Arabic on every line	Yes

Table 9: Submission packaging: file mapping, line counts, script-level sanity checks, and pairwise cross-slot distinctness. The distinctness check guards against a subtle failure mode of ensemble pipelines, where two selection configurations silently converge to (near-)identical outputs and waste a submission slot. The AR→EN primary filename names COMETKiwi although the selector is MetricX-QE (§4.2).

Candidate system	Base model	Adaptation	Decoding	AR→EN pool	EN→AR pool
NLLB-1.3B LoRA (2 ep.)	NLLB-200 distilled 1.3B	LoRA, train split	beam $k=5$	1,500	—
NLLB-1.3B zero-shot	NLLB-200 distilled 1.3B	none	beam $k=5$	—	1,500
NLLB-3.3B zero-shot	NLLB-200 3.3B	none	beam $k=5$	—	1,441
HY-MT1.5 zero-shot	HY-MT1.5 7B	none	greedy, translation-only prompt	1,497	1,493
HY-MT1.5 LoRA	HY-MT1.5 7B	LoRA (NF4), train split	greedy	1,326	—
Hunyuan-MT zero-shot	Hunyuan-MT 7B	none	greedy, translation-only prompt	1,427	1,381
Hunyuan-MT QLoRA	Hunyuan-MT 7B	QLoRA (NF4), train split	greedy	1,493	—
Gemma 3 12B zero-shot	Gemma 3 12B IT	none	greedy, translation-only prompt	1,491	1,489
Aya Expanse 8B zero-shot	Aya Expanse 8B	none	greedy, translation-only prompt	—	1,480
Pool size $ \mathcal{M} $				6	6

Table 10: Every candidate system available to the selector, per direction, with the number of Challenge segments (of 1,500) on which that system contributed a surviving candidate. “Greedy” denotes deterministic decoding (num_beams=1, temperature=0.0) with a fixed prompt and 256 maximum new tokens; the 7B/12B LLMs were served in NF4 4-bit. Membership is taken from the packaged pool files `cross_pool.test.jsonl` for each direction. Aya Expanse 8B contributes to EN→AR only, and the AR→EN pool contains *two* HY-MT1.5 variants (zero-shot and LoRA) rather than a second NLLB LoRA checkpoint. The devtest ablation pool used in Table 6 is a different and much wider pool (see below).

preview of the submitted systems: the devtest selector chose from a pool roughly four and a half times wider, with a very different composition, so absolute devtest scores should not be read as predictions of Challenge behaviour. Second, because the NLLB n -best list supplies the large majority of candidates, the devtest pool is far less diverse in *model identity* than its raw size suggests, which works against the entity floor: near-duplicate n -best variants tend to share the same entity omissions, so a floor violation is rarely repairable from within the pool. The devtest ablation used the *same* gazetteer as every other experiment in this paper: 172 TSV rows, which resolve to 93 distinct Arabic surface forms once the up-to-three ranked Wikidata candidates per form are collapsed to the rank-0 canonical label (Appendix E). Its small size is a further limit on how often the floor can act.

B.2 Hyperparameters

Table 11 consolidates all hyperparameters and resources across the three stages of the pipeline. Two design choices deserve emphasis. First, all LLM candidates are decoded *deterministically* with a

fixed translation-only prompt: this makes every candidate pool exactly reproducible and attributes all output diversity to model identity rather than sampling noise, which is the property the cross-pool selector of Equation (1) exploits. Second, the entity floor τ was the only selection hyperparameter chosen on the devtest split; the QE utilities themselves are used off-the-shelf without task adaptation, so the selection stage introduces no additional trainable parameters.

C Wikidata Entity Gazetteer

C.1 Construction Pipeline

The gazetteer was built with type-restricted SPARQL queries over Wikidata (Vrandečić and Krötzsch, 2014), one query family per entity class: humans, organizations, locations, sports organizations, and events. For each class, the query collects the Arabic label *and* all Arabic aliases of every entity together with its canonical English label; aliases matter because Arabic sports text routinely uses short forms (نشواميني for a player whose canonical label carries the full given name). Figure 4 shows the simplified query template. The stored

Component	Setting
<i>Candidate generation</i>	
Pools	Table 10; $ \mathcal{M} = 6$ (AR→EN), 6 (EN→AR)
NLLB decoding	beam search, beam size 5
LLM decoding	greedy (num_beams=1, temperature=0.0), 256 max new tokens
LLM prompt	Translate the following segment into {English Arabic}, without additional explanation. \\n\\n{source}, wrapped in each model's own chat template
Output cleaning	preamble stripping; one layer of quote unwrapping; Unicode + whitespace normalization
<i>NLLB LoRA fine-tuning</i>	
Rank r / scaling α	16 / 32
Dropout	0.05
Optimizer / schedule	AdamW / cosine
Precision	bfl6, gradient accumulation
<i>Hunyuan/HY-MT-family (Q)LoRA</i>	
Quantization	4-bit NF4
LoRA target modules	attention + feed-forward projections
<i>Reference-free selection</i>	
Utility, AR→EN primary	MetricX-24-QE: google/metricx-24-hybrid-x1-v2p6-bfloat16
Utility, other systems	COMETKiwi: Unbabel/wmt22-cometkiwi-da (AR→EN Contrastive 2 and all EN→AR)
Granularity	segment-level, no document context
Entity floor τ	0.85 on devtest; the packaged Challenge run was invoked with 1.00, which is decision-equivalent here (§4.1)
<i>Data</i>	
Train / dev / devtest	21,000 / 1,000 / 500 pairs
Challenge Test	1,500 segments per direction
<i>Software and hardware</i>	
Frameworks	PyTorch; HuggingFace Transformers; PEFT
Hardware	single Blackwell-class GPU, bfl6 inference

Table 11: Hyperparameters and resources across the pipeline. Off-the-shelf QE utilities plus deterministic decoding make the entire selection stage reproducible from the released output files alone.

file keeps, for each Arabic surface form, up to three ranked Wikidata candidates; only the rank-0 English label is treated as canonical (Appendix E). The entity class is not stored as a field, so the five classes describe the query design rather than a per-entry annotation.

Remote SPARQL endpoints are rate-limited, and our development logs recorded throttled and truncated result sets even with paginated queries and retry-with-backoff. Coverage is therefore *partial by construction*: the floor of Equation (4) only constrains entities the gazetteer knows, and an omission of an uncovered entity passes the floor unpenalized. This asymmetry is intentional — a partial precision-oriented veto is safe, whereas a partial recall-oriented reward would systematic-

```
SELECT ?e ?arForm ?enLabel WHERE {
  ?e wdt:P31/wdt:P279* wd:<CLASS> .
  { ?e rdfs:label ?arForm .
    FILTER(LANG(?arForm) = "ar") }
  UNION
  { ?e skos:altLabel ?arForm .
    FILTER(LANG(?arForm) = "ar") }
  ?e rdfs:label ?enLabel .
  FILTER(LANG(?enLabel) = "en")
}
```

Figure 4: Simplified SPARQL template used to harvest (Arabic surface form, canonical English label) pairs, instantiated once per entity class (<CLASS> = Q5 for humans, and analogously for organizations, locations, sports organizations, and events). The UNION branch harvests Arabic aliases in addition to primary labels.

cally bias selection toward whatever slice of the entity space happened to be harvested. The rejected additive-reward variant of Table 6 is the empirical counterpart of this argument.

C.2 Matching Normalization

The containment test $\ell(e) \sqsubseteq y$ of Equation (2) operates on normalized strings: Unicode NFC, case-folding, and whitespace collapsing on the hypothesis side; on the source side, Arabic surface forms are matched after the same Unicode normalization used in preprocessing, so that presentation-form and definite-article variants (البرنايو vs. بيرنايو) map to the same gazetteer key. No fuzzy or transliteration-aware matching is applied in the submitted systems — a deliberate precision-first choice whose cost is quantified by the $\bar{R}_{\text{ent}}$ column of Table 6 and whose relaxation we identify as future work.

C.3 Representative Entries

Table 12 shows representative entries across the five entity classes, annotated with the translation phenomenon each one exercises. The examples make concrete why Arabic→English entity translation is *not* a character-level transliteration problem but a knowledge-grounded normalization problem, with at least six recurring sub-phenomena: (i) **diacritic recovery** — Arabic script encodes neither Latin accents nor vowel qualities, so *Tchouaméni*, *Güler*, and *Vinicius* must be restored from knowledge, not sound; (ii) **full-name recovery** — Arabic text uses bare surnames where the canonical English label carries given names; (iii) **definite-article stripping** — the Arabic article ال attaches to borrowed names and acronyms and must not surface in English; (iv) **acronym re-Latinization** — اليورفا is a phonetic rendering of an English

acronym, not a word to translate; (v) **brand-form restoration** — club and event names carry conventional prefixes/suffixes (*FC*, *CF*, *FIFA*) absent from everyday Arabic usage; and (vi) **descriptive-to-brand mapping** — أبطال دوري أوروبا literally reads “European Champions League” and must map to the branded *UEFA Champions League* rather than a compositional gloss. Every one of these sub-phenomena defeats naive string matching in a different way, which is precisely why $\bar{R}_{\text{ent}}$ in Table 6 is a lower bound on semantic entity preservation and why the floor is used as a veto rather than a score.

D Worked Example of the Entity Veto

Figure 5 traces Algorithm 1 on a constructed segment with $|G(x)| = 3$ recognized entities and four candidates. It illustrates the two regimes discussed in §6.1 and §7: in panel (a) two candidates omit a known entity and are vetoed, so the QE utility chooses among the survivors; in panel (b) every candidate renders one entity in a non-canonical variant, so $\mathcal{F}_\tau(x) = \emptyset$, the fallback fires, and the floor is the identity — while R_{ent} still counts the variant as a miss. Regime (b) is why the floor can leave $\bar{R}_{\text{ent}}$ almost unchanged on such segments without failing to do its job.

E Reference-Free Entity Audit Protocol

The audit is computed from the released output files, the Challenge Arabic sources, and the gazetteer, with no references and no model inference; it is therefore fully reproducible by the organizers or any reader. Algorithm 2 states the full protocol.

Two implementation details matter for reproducing the numbers in Table 5. First, $G(x)$ is built by longest-match-first, non-overlapping scanning of the raw Arabic source against the gazetteer keys, and $\ell(e)$ is the *first* English label listed for that key — the gazetteer stores up to three ranked Wiki-data candidates per Arabic form, and only the rank-0 label is treated as canonical. Second, the containment test runs on NFKD-normalized, diacritic-stripped, casefolded, whitespace-collapsed strings. We reuse the selector’s own matching functions verbatim rather than reimplementing them, so the audit measures the floor under exactly the criterion the floor enforces.

Three properties make the audit informative beyond the official deltas. First, when Primary and

(a) Veto regime: $\mathcal{F}_\tau(x) = \{c_2, c_4\}$

Source: سجل ريكيلي في كامب نو تحت أنظار اليوفا
 $G(x) = \{\text{Riquelme, Camp Nou, UEFA}\}$, $\tau = 0.85 \Rightarrow \text{need } 3/3$

c_1 : The Argentine playmaker scored at Camp Nou under UEFA’s eyes.
 $R_{\text{ent}}=2/3$ $Q=0.86$ veto

c_2 : Riquelme scored at Camp Nou as UEFA watched.
 $Q=0.84$ feasible $R_{\text{ent}}=1$

c_3 : Riquelme scored at Camp Nou in front of the officials.
 $R_{\text{ent}}=2/3$ $Q=0.85$ veto

c_4 : Riquelme scored at the Camp Nou under the eyes of UEFA.
 $R_{\text{ent}}=1$ $Q=0.85$ feasible $\Rightarrow$ selected

Unconstrained QE re-ranking would pick c_1 (highest Q), silently dropping the player’s name.

(b) Fallback regime: $\mathcal{F}_\tau(x) = \emptyset$

Source: هدف فينيسيوس في البرنابيو حسم كأس العالم
 $G(x) = \{\text{Vinicius Júnior, Bernabéu, FIFA World Cup}\}$

c_1 : Vinicius’s goal at the Bernabeu decided the World Cup.
 $R_{\text{ent}}=0/3$

c_2 : Vinicius Junior’s goal at the Bernabéu settled the World Cup.
 $R_{\text{ent}}=1/3$

c_3 : A goal by Vinicius at the Bernabéu sealed the FIFA World Cup.
 $R_{\text{ent}}=2/3 \Rightarrow$ fallback: arg max Q

No candidate reaches τ ; Eq. (4) returns the unconstrained winner. All three outputs are acceptable to a reader, yet every one scores < 1 on R_{ent} because $\text{Vinicius} \neq \text{Vinicius Júnior}$.

Figure 5: Constructed (not corpus-drawn) illustration of Algorithm 1. Dashed red boxes are candidates with $R_{\text{ent}} < \tau$; solid teal boxes are feasible; the shaded box is the returned hypothesis. Panel (b) shows the regime that accounts for the fallback branch firing on most entity-bearing segments (§6.1): the floor has nothing to veto, and R_{ent} penalizes variants that are not errors.

Contrastive 2 share a QE utility and differ only by the floor, $\bar{R}_{\text{ent}}(\text{Primary}) \geq \bar{R}_{\text{ent}}(\text{Contrastive 2})$ holds by construction. That precondition does not hold for the files actually submitted, which differ in utility as well, so the inequality observed here (0.4264 vs. 0.3931) is empirical rather than structural. It can nonetheless be attributed to the floor, because the matched control C2’ isolates it directly (+0.0276 of the +0.0333 total) and because the audit recovers the floor’s *action set* — the segments on which $\mathcal{F}_\tau(x)$ is a proper non-empty subset of $\mathcal{C}(x)$. Of the floor-only difference P–C2’, 97.7% lies on the 26 action-set segments where the floor overrides the unconstrained winner, and on those segments the Primary attains $R_{\text{ent}} = 1$ exactly; of the joint difference P–C2, 91% lies on the 66-segment action set as a whole (§6.1). Second, Changed bounds from above the num-

Arabic surface form	Canonical English label	Entity class	Phenomenon exercised
تشواميني	Aurélien Tchouaméni	human (athlete)	full-name recovery; French diacritics
جولر	Arda Güler	human (athlete)	full-name recovery; Turkish diacritics
مبابي	Kylian Mbappé	human (athlete)	full-name recovery; French acute accent
فينيسيوس	Vinicius Júnior	human (athlete)	Portuguese diacritics; suffixed short form
أنشيلوتي	Carlo Ancelotti	human (coach)	full-name recovery; ش → c sound-mapping ambiguity
ريكيكلي	Riquelme	human	conventional Spanish orthography (<i>qu</i>)
ريال مدريد	Real Madrid CF	sports organization	brand-suffix restoration (<i>CF</i>)
برشلونة	FC Barcelona	sports organization	brand-prefix restoration (<i>FC</i>); vowel mapping
اليويفا	UEFA	organization	acronym re-Latinization; article ال stripping
بيرنابيو	Bernabéu	location (stadium)	Spanish accent restoration; venue exonym
كامب نو	Camp Nou	location (stadium)	Catalan venue name; no article
دوري أبطال أوروبا	UEFA Champions League	event	descriptive-to-brand mapping
كأس العالم	FIFA World Cup	event	brand-prefix restoration (<i>FIFA</i>)

Table 12: Representative gazetteer entries across the five harvested entity classes, drawn from the football-heavy Challenge domain. Each Arabic surface form omits or re-encodes information — diacritics, given names, articles, brand affixes, Latin orthographic conventions — that the canonical English label must restore. This is why exact-surface entity matching is precision-oriented: a correct but variant transliteration in system output (e.g. *Vinicius* without diacritics, or *European Champions League* as a compositional gloss) would not match the canonical label and is scored as a miss by R_{ent} , while remaining acceptable to a human reader. The floor of Equation (4) therefore vetoes only hard omissions, never spelling variants alone.

Algorithm 2 Reference-free entity audit of a submission file

Require: sources $\{x_i\}_{i=1}^{1500}$; hypothesis file $H = \{h_i\}$; unconstrained file $U = \{u_i\}$ (Contrastive 2); pool files $\{y_{m,i}\}$; gazetteer $\mathcal{G}$; τ

- 1: $S \leftarrow \{i : G(x_i) \neq \emptyset\}$ $\triangleright$ segments with a recognized entity
- 2: $\bar{R}_{\text{ent}} \leftarrow \frac{1}{|S|} \sum_{i \in S} R_{\text{ent}}(x_i, h_i)$
- 3: $\text{Full} \leftarrow \frac{1}{|S|} |\{i \in S : R_{\text{ent}}(x_i, h_i) = 1\}|$
- 4: $\text{Changed} \leftarrow \frac{1}{|S|} |\{i \in S : h_i \neq u_i\}|$
- 5: $\text{Fallback} \leftarrow \frac{1}{|S|} |\{i \in S : \max_m R_{\text{ent}}(x_i, y_{m,i}) < \tau\}|$
- 6: **return** ($|S|$, $\bar{R}_{\text{ent}}$, Full , Changed , Fallback)

ber of segments that can account for the official -0.60 BLEU/ $+0.37$ COMET joint deltas, which converts a corpus-level metric difference into a per-segment intervention rate. Third, Fallback measures how often the pool contained no canonical-form candidate at all; a high fallback rate indicates that gazetteer coverage and matching strictness, not the veto rule, are the binding constraints, and it is what prioritizes transliteration-aware matching as future work.