The Broken Telephone Changes Tone: Examining Nuanced Linguistic Cues in LLM Chains-of-Translation

Anonymous Authors[1]

Abstract

As LLM-generated content proliferates online, texts are increasingly subject to repeated processing and translation by models, making it critical to understand how such iterative reprocessing reshapes language. Prior work has shown that this degrades factual content and reduces diversity, but the fine-grained linguistic shifts underlying these effects remain unexplored. We track changes in epistemic markers, grammatical voice, degree adverbs, and nominalisation density across 12 iterations of round-trip translation applied to 600 BBC News articles, varying intermediate language, translation model, and chain topology across 17 experimental configurations. We find a consistent epistemic shift: evidential and factive markers increase while hedges decline, potentially causing tentative claims to read as more certain. Concurrently, texts undergo register formalisation: informal degree adverbs give way to formal alternatives, active-voice density drops, with-agent passives attrite disproportionately, and nominalisation density rises. We also record clear model-specific patterns for certain settings. These shifts erode the markers of source, register, and agency, offering a fine-grained account of the factual degradation reported in previous studies.

1. Introduction

As large language models produce a growing share of online content, that content is increasingly reprocessed (summarised, translated, paraphrased) by other models. Understanding how text changes under such iterated processing is critical for assessing the reliability of LLM-mediated information.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Recent work has shown that iterated rephrasing drives text toward model-specific attractor states (Perez et al., 2025) and that iterated backtranslation rapidly degrades factual content (Mohamed et al., 2025). At the training level, recursive use of synthetic data causes model collapse (Shumailov et al., 2024; Dohmatob et al., 2025) and reduces epistemic diversity (Wright et al., 2025). However, these studies measure degradation through aggregate metrics (semantic similarity, factual accuracy, distributional statistics), leaving open the question of *how* the linguistic structure of the text is reshaped along the way.

We address this gap by tracking changes in epistemic markers, grammatical voice, degree adverbs, and nominalisation density across 12 iterations of backtranslation applied to 600 BBC News articles. We vary intermediate language (Chinese, French, Korean, Arabic), translation model (Qwen, Llama, Gemma), and chain topology (self-loop, two-player, multiplayer), yielding 17 configurations and over 122,000 translated texts.

We find three main patterns. First, evidential and factive markers increase while hedges decline, potentially causing originally tentative claims to read as more certain, eroding precisely the cues readers use to calibrate trust. Second, texts undergo formalisation: informal degree adverbs (*really*, *much*) give way to formal alternatives (*particularly*, *completely*), active-voice density drops, with-agent passives attrite faster than agentless passives, and nominalisation density increases. Third, the magnitude of these shifts varies by model (Qwen self-loops produce the largest nominalisation and booster increases, Llama the smallest), suggesting model-specific linguistic signatures.

These results show that iteratively processed text is not just less accurate but systematically restructured: uncertainty cues are stripped, register is formalised, and translation models may shift results in model-specific directions.

2. Related Works

Iterative processing with LLMs. Repeated processing of LLM outputs causes distributional narrowing across settings. Models trained on their own outputs undergo *model collapse*, progressively losing distributional tails (Shumailov et al.,

2024; Dohmatob et al., 2025), and Wright et al. (2025) show that LLM outputs are less epistemically diverse than basic web search, a narrowing Peterson (2025) terms *knowledge collapse*.

Analogous effects arise when text is iteratively reprocessed at inference time. Acerbi & Stubbersfield (2023) ran transmission chain experiments with ChatGPT and found that iteratively passed text amplifies human-like content biases for gender-stereotypical, negative, and threat-related information. Perez et al. (2025) extended this framework, finding that small per-step biases in toxicity, positivity, and reading difficulty compound into model-specific attractor states across iterated rephrasing. Mohamed et al. (2025) applied iterated backtranslation, showing that factual content degrades rapidly, with the rate influenced by intermediate language, model, and chain topology. Our work builds on Mohamed et al. (2025) but shifts from macro-level degradation metrics (BLEU, BERTScore, FActScore) to the fine-grained linguistic shifts (in epistemic marking, voice, and degree modification) that accumulate as factual loss occurs.

Hedging, degree adverbs, and grammatical voice. The epistemic markers we track draw on the Hyland (1998) taxonomy of hedges and boosters and the factive/non-factive verb distinction of Kiparsky & Kiparsky (1970). Recent work has shown that LLMs are highly sensitive to epistemic markers, with expressions of certainty paradoxically decreasing accuracy (Zhou et al., 2023), and that this overconfidence persists across languages (Rathi et al., 2025), inducing bias in LLM-based evaluations (Lee et al., 2025). Our work shows that iterated processing compounds such biases in originally human-authored text.

Degree adverbs, which modify scalar force and include both amplifiers and downtoners (Biber et al., 2000), are register-sensitive: informal registers favor items like *really* and *pretty*, while formal registers prefer *particularly* and *significantly*, making them a natural probe for formalisation under iterated translation. Grammatical voice has been studied as a marker of structural divergence in translation; MT systems tend toward more conservative, source-parallel active structures (Luo et al., 2024). Passive constructions vary substantially across languages in frequency and function (Siewierska, 2024): in Chinese, passives carry adversative connotations absent in English (Baker, 2018; Xiao et al., 2006), while Arabic passives are morphologically constrained and less frequent than their English counterparts (Farghal & Al-Shorafat, 1996). These cross-linguistic asymmetries make voice distribution a sensitive indicator of how iterated translation reshapes text structure.

3. Methodology

3.1. Data

In order to examine recent global trends across a variety of news domains, we use BBC News articles as collected by Li et al. (2024), but we restrict our dataset to entries on or after 1 January 2024, that belong to one of the following 12 topics: abortion, China, climate, COVID, economics, France, gender, immigration, Israel/Palestine, Russia, sports and the UK. For each topic, 50 articles were sampled (seed 42) under a length filter of 500 to 5,000 characters, yielding 600 articles per run. With one iteration 0 source and 12 back-translated iterations, our setup produced 7,800 texts per run.

Three models were used for translation, all accessed through OpenRouter: Qwen (`qwen3-30b-a3b-instruct-2507`) (Qwen Team, 2025), Llama (`llama-3.3-70b-instruct`) (Grattafiori et al., 2024), and Gemma (`gemma-3-27b-it`) (Team et al., 2025). We set the translation calls' temperature to 1.0. Four intermediate languages were investigated: Chinese (Simplified), French, Korean, and Modern Standard Arabic (hereafter Arabic).

3.2. Translation Chain Setups

The core of our experiments is iterated backtranslation: for each article, the English source is translated into an intermediate language (L), and then back into English (EN). This single round is repeated 12 times, producing translations at iterations 1 to 12, alongside the iteration 0 source.

On top of this design, three translation settings were run: **bilingual self-loop**, **bilingual two-player**, and **multilingual multiplayer**. The **bilingual self-loop** uses a single model for both forward and backward translation, with each iteration consisting of a translation from EN to L, and then back to EN by the same model; 3 models with 4 languages yields 12 total runs. The **bilingual two-player** splits the two directions model-wise: Qwen handles EN to L, while Llama takes care of the L to EN translation, run once per language for 4 runs in total. **Multilingual multiplayer** chains three models across two intermediate languages per iteration: Qwen translates from EN to the first L, Gemma handles translation from the first L to another L, and Llama translates from the second L back to English. Both the first and second L are sampled without replacement each iteration from our list. This constitutes 1 run.

In total, the experiments comprise 17 runs (12 self-loop + 4 two-player + 1 multiplayer), and 17 $\times$ 7,200 = 122,400 back-translated texts, excluding iteration 0 sources (7,200 = 600 articles $\times$ 12 iterations per run).

3.3. Metrics

A total of five evaluators were applied to the texts. Four of them use spaCy and regular expression (regex) only, whereas the fifth is a set of reference-based metrics computed against the iteration 0 source, including one neural metric.

The **epistemic markers** evaluator counts regex matches against six lexicons (hedges, boosters, factive, non-factive, evidential and non-evidential) drawn from the existing body of relevant literature (Hyland, 1998; Farkas et al., 2010; Kiparsky & Kiparsky, 1970; Karttunen, 1971); more details are in Appendix A. For **degree adverbs**, we identify adverbial modifier (advmod) dependencies whose lemma is part of a 14-word lexicon, derived empirically from adverbial modifier lemmas observed in the BBC corpus and model outputs combined (details in Appendix C). The **voice** evaluator similarly uses spaCy's dependency parsing to classify clauses as one of active (subject performs action on object, e.g., *the cat drank water*), agentless passive (subject receives action without agent being mentioned, e.g., *the water was drunk*) and with-agent passive (subject receives action from named agent, usually using a 'by-phrase', e.g., *the water was drunk by the cat*). This classification is based on the nominal subject (nsubj), nominal subject passive (nsubjpass), auxiliary passive (auxpass), and agent labels. The **agency** evaluator provides counts complementary to voice: existentials (expl), nominalisations (noun tokens where lemma carries a deverbal suffix; one of –tion, –sion, –ment, –ance, –ence, –al, or gerund –ing), and total verb + auxiliary tokens as a verb-density baseline.

On the other hand, the **similarity versus source** evaluator computes five reference-based metrics against the iteration 0 source: BLEU and chrF (Post, 2018), ROUGE-1 F1 from Google's rouge score, stemmed (Lin, 2004), METEOR through NLTK with WordNet (Miller, 1994), and BERTScore F1 with RoBERTa Large (Liu et al., 2019; Zhang et al., 2020). Since the iteration 0 source is the text that is meant to be preserved by the chain (i.e., the aforementioned reference), these metrics capture the cumulative drift from the original across iterations; they do not assess translation quality in the traditional machine translation sense. Thus, the evaluator skips iteration 0, which has no reference to compare against.

The first four evaluators run on every translated text as well as on the iteration 0 source, producing 7,800 records per evaluator each run. The similarity versus source evaluator, meanwhile, produces 7,200 records per run, excluding iteration 0. Each record stores the text word count, allowing per-1,000-words normalisation during analysis.

Separately, we use FActScore (Min et al., 2023) to track factual support, with the iteration 0 article as the reference and Llama-3.3-70B as the judge. Every per-run iter-0 to iter-12 delta we report comes with a bootstrap standard error, obtained by resampling articles within the run (B=2,000); the full SE tables are in Appendix F.

4. Results and Discussions

4.1. Length and Factuality

We confirm the finding of Mohamed et al. (2025) that factuality decays under iterated back-translation. Using FActScore, we extract a mean of 25.8 atomic facts per source article (15,473 facts across 600 articles) and check support against the iteration 12 back-translation in each of the 17 runs; of the resulting 263,041 (article, fact, run) triples, 57% of source facts remain supported at iteration 12, with the self-loop support rate varying widely across backward-pass models: 78.2% for Gemma, 55.1% for Qwen, and 42.8% for Llama. Reference-based similarity drops in the same direction: averaged across all 17 runs, BLEU falls from 36.8 at iteration 1 to 21.0 at iteration 12, chrF from 67.3 to 53.3, ROUGE-1 F1 from 0.76 to 0.61, METEOR from 0.60 to 0.43, and BERTScore F1 from 0.94 to 0.91. Article length also drops across iterations: iteration 12 outputs are 1 – 57% shorter than the iteration 0 source, with Korean and Arabic Llama runs shortest and Arabic Qwen closest to source length. This contrasts with Perez et al. (2025), who report stable-or-increasing length in their iterated-generation setup. Because per-article counts in our data are confounded with this length drift, we normalise all marker statistics per 1,000 words throughout the paper.

4.2. Hedging

Across the 17 experimental runs (12 single-model bilingual self-loops, 4 two-player runs, and the multiplayer run), evidential markers rise in every run from iteration 0 to iteration 12, with a mean increase of $+0.53$ against an iteration-0 baseline of 0.90; factive verbs rise in 15 of 17 runs (the exceptions, AR-Llama at -0.15 and FR-2P at -0.06, are small); hedges decrease in 13 of 17 runs, with CH-Llama as the positive outlier ($+2.53$); and boosters rise in 11 of 17 runs, with a clean by-model split among the self-loops: all four Qwen self-loops $+1.29$ to $+2.34$, all four Llama self-loops -0.49 to $+0.19$, Gemma in between. The evidential rise is the only change that is universal across bridge languages[1], models, and settings.

We attribute the evidential rise to literal back-translation of high-frequency attribution words from each bridge language's news register. Examples include Chinese 据/根

[1]By *bridge language* we mean the intermediate language used in each round-trip (Chinese, French, Korean, or Arabic in our setup), and by *bridge text*, the article's translation into that language at iteration 12.

Table 1. Co-occurrence of bridge-language attribution words and English evidential markers, per run (600 articles each). Run labels combine the bridge language (CH/FR/KO/AR) with the model (Qwen, Gemma, Llama for self-loops) or setting (2P, MM). Δ is the change in evidential rate from iteration 0 (baseline 0.90) to iteration 12, with bootstrap SE from resampling articles within each run (B=2,000). Columns 3 and 4 give the fraction of articles producing an English evidential, restricted to articles whose bridge text contained an attribution word and articles whose bridge text contained none, respectively. Column 5 gives the reverse: the fraction of articles with an English evidential whose bridge text also contained an attribution word.

		rate of English evidential		rate of attribution
Run	**Δ**	**if attribution**	**if no attribution**	**if English evidential**
CH-Qwen	$+0.26_{\pm 0.07}$	0.36	0.03	0.99
CH-Gemma	$+0.00_{\pm 0.06}$	0.33	0.04	0.99
CH-Llama	$+0.25_{\pm 0.08}$	0.34	0.01	0.99
FR-Qwen	$+0.81_{\pm 0.08}$	0.49	0.08	0.99
FR-Gemma	$+0.37_{\pm 0.05}$	0.41	0.03	1.00
FR-Llama	$+0.30_{\pm 0.07}$	0.35	0.03	0.98
KO-Qwen	$+1.05_{\pm 0.09}$	0.59	0.08	0.98
KO-Gemma	$+0.29_{\pm 0.06}$	0.41	0.11	0.98
KO-Llama	$+1.05_{\pm 0.13}$	0.41	0.04	0.97
AR-Qwen	$+0.65_{\pm 0.08}$	0.51	0.24	0.94
AR-Gemma	$+0.28_{\pm 0.06}$	0.40	0.11	0.98
AR-Llama	$+1.00_{\pm 0.12}$	0.39	0.16	0.86
CH-2P	$+0.24_{\pm 0.08}$	0.34	0.00	1.00
FR-2P	$+0.73_{\pm 0.09}$	0.45	0.15	0.94
KO-2P	$+1.01_{\pm 0.09}$	0.50	0.08	0.97
AR-2P	$+0.43_{\pm 0.08}$	0.44	0.13	0.95
MM	$+0.31_{\pm 0.08}$	0.38	0.00	1.00

据/表示/称/报道 (*jù*, *gēnjù*, *biǎoshì*, *chēng*, *bàodào*); French *selon*/*d'après*/*indique*; Korean 에 따르면/의 하면/밝혔 (*e ttareumyĕon*, *ŭihamyĕon*, *palkyĕ*); and Arabic *wafqan*, *biḥasab*, *qāla*, and *afāda* (more details in Appendix D). Each is rendered in the back-translation as "according to", "based on", or another phrase from our evidential lexicon. To test this, for every iteration 12 article we compare the bridge text against the corresponding English back-translation; in the multiplayer setting, which uses two intermediate languages per iteration, we examine both bridge texts. Across the 17 runs, articles whose bridge text contained at least one of the listed attribution words produced an English evidential in 33–59% of cases, whereas articles whose bridge text contained none of them did so in only 0–24%, which is a 2–27× contrast (Table 1, Figure 1). As these rates are different across the two groups (foreign attribution words present or absent), we see that evidential markers rise not just because of general stylistic drift. In the reverse direction, English evidentials in the back-translation co-occurred with a bridge attribution word in 86–100% of cases (mean 97%); this means our attribution lexicon does not miss any major source of evidentials. Cross-run variation in the size of the rise (Korean largest, Chinese smallest, Arabic and French intermediate) reflects the number of obligatory attribution forms in each bridge's news register. The 11–24% rate among Arabic articles whose bridge text contained none of our listed words reflects incomplete coverage of Arabic attribution verbs in our 8-item lexicon.

The combined effect of the evidential rise, factive rise, and hedge decline is a change in how source claims are framed. A claim that is hedged in the BBC source typically reappears at iteration 12 either attributed to a third party (*according to Y, X happened*) or asserted via a factive verb (*Y confirmed that X*); the back-translation thus reads as more confident than the source[2], which is something co-occurring with the degradation of factuality described earlier in Section 4.1. This shift in epistemic framing is one facet of a broader register shift that we trace through degree adverbs and voice in the next two sections.

4.3. Degree Adverbs

Table 2 summarises the changes in the degree adverb density from iteration 0 to iteration 12 across our experimental setups. The given deltas are calculated from the baseline counts of 2.68 degree adverbs per 1,000 words at iteration 0.

Bilingual self-loop. Across the 12 iterations, the total degree adverb density shows a model-dependent divergence. From iteration 0 to 12, Qwen shows a consistent increase in degree adverb density across all four languages, with more substantial gains in Korean (+1.92), Chinese (+1.65), and French (+1.32), and a shallower increase in Arabic (+0.10).

[2]For a detailed example, see Appendix B.

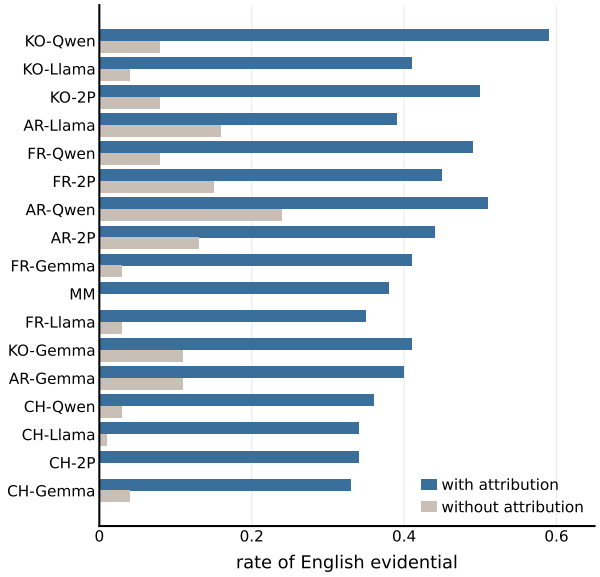

Figure 1. Rate at which an iteration 12 back-translation contains an English evidential marker, split by whether the bridge text contains a bridge-language attribution word. Runs sorted by Δ.

Table 2. Net change (Δ) in degree adverb density (per 1000 words) between iteration 0 (baseline 2.68) and iteration 12, per run (600 articles each). Run labels combine the bridge language (CH/FR/KO/AR) with the model (Qwen, Gemma, Llama for self-loops) or setting (2P for bilingual two-player and MM for multilingual multiplayer).

Run	Δ
CH-Qwen	+1.65
CH-Llama	−0.66
CH-Gemma	+0.23
FR-Qwen	+1.32
FR-Llama	−0.08
FR-Gemma	+0.18
KO-Qwen	+1.92
KO-Llama	−0.99
KO-Gemma	+0.33
AR-Qwen	+0.10
AR-Llama	−1.21
AR-Gemma	−0.11
CH-2P	+0.28
FR-2P	+0.56
KO-2P	+0.60
AR-2P	−0.75
MM	+1.32

In contrast, Llama self-loops decrease density (with deltas −0.08 to −1.21), with Arabic showing the greatest drop (−1.21). Gemma self-loops remain relatively stable across languages, with small increases in French (+0.18), Chinese (+0.23), and Korean (+0.33), and a case of decrease in Arabic (−0.11). Qwen systematically inflates the intensifiers, Llama suppresses them, and Gemma remains fairly steady.

At the lexical level, several adverbs display consistent shifts across the self-loops. "Really" (−0.01 to −0.39), "much" (−0.06 to −0.23), and "almost" (−0.03 to −0.23) decrease in all 12 runs. Conversely, "particularly" (−0.03 to +0.45) increases in 11 of the 12 self-loops, while "completely" and "nearly" increase in 9 of 12 (up to +0.36 and +0.28, respectively). The loss of "really", "much", and "almost" is notable because of their colloquial and approximate nature, which suggests that iterated translations progressively strip out the hedging-adjacent intensifier lexicons. The gain in "particularly", "completely", and "nearly" points toward a more formal and absolute text. The net effect would be a shift in the evaluative character of the text from more tentative towards more assertive intensification.

Bilingual two-player. All four runs show shifts in degree adverb density. Three out of the four runs, Chinese (+0.28), French (+0.56), and Korean (+0.60), exhibit increases.

Above, we observed in self-loops that Qwen consistently drives increases in degree adverb density, whereas Llama generally suppresses it. In contrast, in the bilingual two-player setting, three of the four runs, French (+0.56), Korean (+0.60), and Chinese (+0.28), result in increases, aligning with Qwen's self-loop direction, despite Llama producing the final output in English. However, when comparing these deltas to pure Qwen self-loops in French (+1.32), Korean (+1.92), and Chinese (+1.65), we see that the changes have certainly been dampened by Llama's suppression. Comparing Qwen's self-loops in Arabic (+0.10) with those of Llama (−1.21), the change in Arabic is also dampened in the bilingual two-player case.

Moreover, Arabic shows the only decrease (−0.75), diverging from the other three languages. Given that Arabic self-loops already exhibit the weakest increases or outright decreases across models, this points to language-level effects that operate independently of, and can override, model-level tendencies.

Multilingual multiplayer. The degree adverb density rises by +1.32, comparable to the Qwen self-loops with Chinese, French, and Korean. Across all topics, degree adverb density rises through iteration 12, with the largest increase in Sports and the smallest in Israel/Palestine, Russia, and economics.

Overall, across different experiment settings, we see that the total density of degree adverbs is model-dependent (increasing for Qwen, decreasing for Llama, and flat for Gemma, with dampening effects when more than one model is used). We also observe how informal items like "really", "much", and "almost" are stripped to make place for formal items such as "particularly", "completely", and "nearly" in all

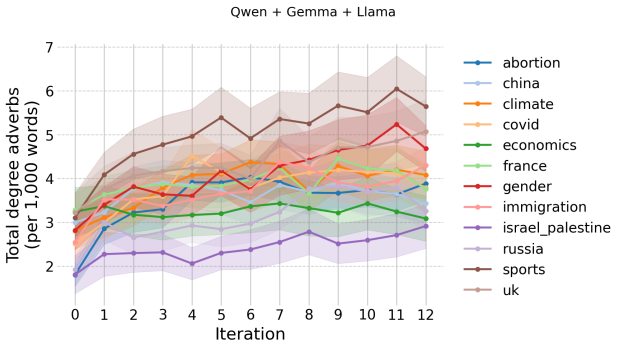

Figure 2. Line plot of degree adverb density per 1,000 words, grouped by the article topic.

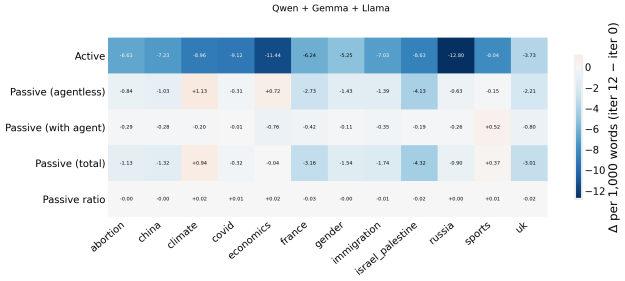

Figure 3. Heatmap of the deltas per 1,000 words (as calculated from the baseline counts) of topic-wise active and passive voice metrics.

self-loops. We argue that this underlying shift is register formalisation, where lexical substitution tells a more complete story than total density does, a pattern we also see with voice and nominalisation in Section 4.4.

4.4. Voice

Appendix Table 4 summarises the changes in active and passive voice from iteration 0 to 12 across our experimental set-ups. The given deltas are calculated from the baseline counts (per 1,000 words) at iteration 0: active 83.4, agentless passive 11.2, with-agent passive 1.9, and passive ratio (i.e., the number of passives divided by the total number of passives and actives) 0.137.

Bilingual self-loop. We find that, from iteration 0 to 12, active voice density per 1,000 words declines in 11 out of 12 self-loops, spanning −0.50 to −12.48, with CH-Llama being the sole exception at +0.85. The magnitude of this decline, however, is model-dependent; Qwen shows the steepest drops (−4.11 to −12.48), with Gemma (−0.50 to −4.37) and Llama (−5.43 to +0.85) showing smaller and more variable changes; Llama is also the only model with a positive case (CH-Llama, +0.85).

With-agent passive density similarly either drops or stays flat in 11 out of the 12 self-loops, spanning −0.96 to +0.12, with Korean and Arabic losing a larger proportion of with-agent passives versus agentless. This drop is most evident from iteration 0 to 1, after which the decline slows down.

Passive ratio, however, differs more by model than by language. Qwen self-loops decrease the passive ratio uniformly from −0.004 to −0.023, whereas Llama uniformly increases it across all four languages (+0.002 to +0.029). Gemma self-loops are small and mixed (−0.016 to +0.004). Agentless passive density similarly separates by model, with Qwen depicting a uniformly negative trend (−0.97 to −2.35), Llama showing mixed changes (−0.12 to +2.50) with the strongest change in Korean, and Gemma remaining mostly negative (−1.36 to +0.23).

Language also modulates the magnitude of passive ratio shift: Korean and Arabic show the largest effects within each model (active voice drops, and ratio shifts under Llama), Chinese is the most conservative across the three models with the smallest active voice declines, passive total changes, and passive ratio shifts, and French lies in between.

The direction of the ratio shift, however, is determined by the model regardless of the language. Qwen pushing the passive ratio down and Llama bringing it up is consistently evident across the four languages. Theoretically, Korean-style agent dropping tendencies ought to push up passive ratios across all models (Kim, 2007), yet only Llama reliably does so. Each model's internal conventions about English clause construction appear to supersede any structural affordance of the intermediate language.

Bilingual two-player. All four runs depict decreases in the proportion of active voice (−2.79 to −10.26) as well as increases in the passive ratio (+0.001 to +0.027). AR-2P shows both the largest passive ratio increase (+0.027) as well as the largest agentless passive increase across the two-player runs (+1.80).

In the self-loops, Qwen tends to drive the passive ratio down whereas Llama brings it up; however, in the bilingual two-player setting, the passive ratio increases across all four runs, following Llama's self-loop direction, despite Qwen handling the forward pass every iteration. This implies that the backward-pass model determines the voice direction, with the English-side linguistic conventions reappearing. This finding has a pertinent implication for any multimodel pipeline where content is translated back to English: the final English-producing model dictates the dominating stylistic priors, independent of the other participating models.

Multilingual multiplayer. We observe that the active voice density drops uniformly (−7.97 overall), with the topic-wise heatmap in Figure 3 showing losses in each of the 12 topics, the largest of which are observed for Russia

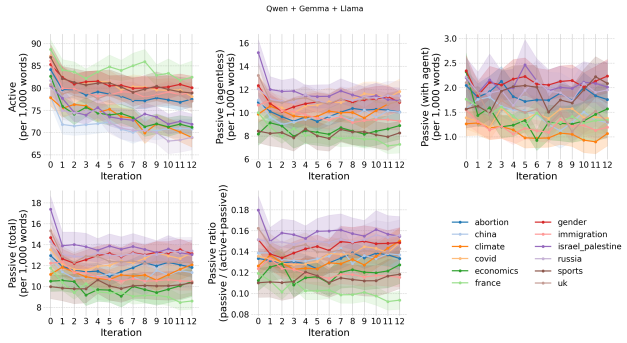

Figure 4. Line plot of the counts per 1,000 words of active and passive voice metrics, grouped by the article topic.

(−12.80), economics (−11.44) and COVID (−9.12). Passive totals, on the other hand, move only slightly (−1.35 overall), with the ratio barely budging (−0.002), meaning that the multilingual multiplayer case is dominated by attrition of the active voice; conversion from active to passive is not the mechanism.

Topic-wise, we find that Israel/Palestine, and (to a lesser extent) Russia, China and gender have the highest baseline passive totals as shown in Figure 4, agentless counts, and passive ratios, which is consistent with the conventions of reporting on war and politically-charged subjects (e.g., agent obscuring terms, like "was killed" or "were displaced"). These high-baseline topics depict a sharper drop from iteration 0 to 1, followed by a continued gradual decline, and especially in the case of Israel/Palestine, which shows the biggest decreases in passive total and passive agentless as per Figure 3 (−4.32 and −4.13 respectively).

High-baseline topics losing a larger proportion of the passive voice is in line with existing literature that shows LLMs' tendencies of regressing towards a linguistic mean; as such, more politically- or socially-charged domains, where the baseline sits far from this mean, experience more flattening. This implies that iterated LLM translations risk understating sensitive news coverage through homogenisation and neutralisation.

Nominalisation and verb attrition. Nominalisation density rises in all 17 runs (+0.9 to +16.1 per 1,000 words; mean +6.95, about 21% above the iteration 0 baseline of 33.0) while the total verb/auxiliary density drops in 15 of 17 runs (−16.16 to +3.26; the two exceptions are CH-Llama and CH-Gemma), as detailed in Appendix Table 5. This corroborates our reading that iterated translation is shedding clauses: verbs are lost from the per-1,000-words density and replaced by additional nouns. Conversion from active to passive is not the mechanism.

Looking into the highest-Δ run (KO-Qwen), we find that much of the nominalisation rise reflects lexical formalisation: Anglo-Saxon vocabulary is replaced with Latinate suffix-bearing forms (*jabs* becomes *injection treatments*, *chaos* becomes *confusion*). Several iteration 12 sentences in fact restore a verb where iteration 0 used a nominal phrase, indicating that strict verb-to-noun grammatical conversion is not the main mechanism. The signal is closer to register formalisation than to clause-to-noun-phrase rewriting, with iterated content drifting towards a more formal, institutional English.

Across all three modes of experimentation, we find that agentless passives survive while with-agent passive voices show attrition, most markedly in Arabic and Korean. This indicates that iterated translation removes "who-did-what" information even when the passive voice itself is preserved. English passives are known to be used in situations of ambiguity and accountability avoidance (Almahameed et al., 2022), and this reconstruction pattern replaces clearer sentence structures with precisely that. As such, unlike errors or factual drifts, agentless passives do not immediately register as harmful on surface level, yet they strip readers of the responsibility-assigning information (e.g., "protesters were dispersed" vs "the police dispersed protesters"). In the high-baseline topics that were flattened most prominently (Israel/Palestine, Russia, China and gender), this attribution erosion is especially problematic since the accountability language carries the greatest stakes.

Overall, voice drift under iterated back-translation is model-based, determined by the backward-pass model, with the intermediate language defining only its magnitude. The process sheds clauses, preferentially retains agentless passives, and substitutes Anglo-Saxon verbs with Latinate nominal forms; conversion from actives to passives is not the dominant mechanism. This replaces clearer constructions with clarity-reducing alternatives, silently eroding attribution while homogenising language in topics where accountability is of the essence.

5. Conclusions

We examined systematic linguistic shifts in LLM-mediated chains of translation applied to BBC News articles. Across 17 configurations of bridge language, model, and chain topology, evidential and factive marker density rises while hedge density declines, with the evidential rise driven by attribution-word calques from bridge-language news registers. Texts also undergo register formalisation: formal degree adverbs replace informal ones, active-voice density drops, with-agent passives attrite faster than agentless ones, and nominalisation density rises. The magnitude of these shifts varies by translation model, most pronounced in nominalisation and boosters. These shifts erode the markers of source, register, and agency, offering a fine-grained account of the factual degradation reported in previous studies.

6. Limitations and Future Work

Our findings are not without their limitations, which we believe set the grounds for future work.

Model and source diversity. Future work could extend our analysis to closed frontier models (e.g., GPT-5.5, Claude Opus 4.7, Gemini 3.1 Pro) to test observed patterns. Furthermore, our corpus comes from a single news outlet (BBC News). While this provides a consistent register and broad topic coverage, editorial style, hedging conventions, and attribution practices vary substantially across outlets. Extending this analysis across outlets would help separate effects that are properties of iterated translation on news text in general from those that reflect BBC's particular editorial conventions.

Quantifying voice and nominalisation. Our nominalisation metric counts noun tokens whose lemma carries a deverbal suffix (one of –tion, –sion, –ment, –ance, –ence, –al, or gerund –ing). This conflates two distinct observations: lexical formalisation (*jabs* becoming *injection treatments*) and grammatical clause-to-noun-phrase rewriting (*X announced that Y* becomes *the announcement of Y*). It also misses zero-derivation nominalisations like *a build* or *the kill*, which admittedly are rarer than the others. As we note in Section 4.4, the KO-Qwen rise reflects predominantly the former mechanism, with the latter playing a smaller role, but our current metrics cannot distinguish the two systematically. Future work could combine dependency parsing with sentence-level alignment to see where a verb in the source has been rewritten as a noun in the back-translation.

Other evaluations. Due to our budget limit, we used Llama 3.3 70B, one of the considered translation models, as the judge for FActScore, which may introduce a confounder. We also do not have any human validation for more certainty in our findings (e.g., that texts become more confident). Our work can also benefit from more statistical significance testing.

Other settings. Iteratively LLM-processed texts also arise outside of translation, e.g., in agentic pipelines and heterogeneous summarisation chains. Whether the register shifts we report extend to those settings remains an open question for future work.

Impact Statement

Our paper contributes a fine-grained linguistic account of how iterated LLM translation reshapes text, complementing prior work on factual degradation and model collapse. Our methodology (tracking epistemic markers, voice, degree adverbs, and nominalisation across translation chains) provides tools that translation researchers, benchmark designers, and content auditors can extend to other iterative pipelines such as summarisation, paraphrase, and agentic workflows.

Our findings carry implications beyond translation itself. Readers and journalists may encounter LLM-processed prose that strips uncertainty cues and obscures agency without any obvious surface marker, complicating source verification and trust calibration, particularly in coverage of conflict, politics, and other domains where accountability language is of the essence. Multilingual communities whose news reaches global audiences through automated translation may find their distinctive editorial voices flattened into a uniform institutional register, with under-resourced languages and outlets potentially absorbing the largest stylistic costs. Additionally, as LLM outputs increasingly contribute to model training corpora, these systematic shifts may accumulate across model generation iterations, intersecting directly with the model collapse literature.

We therefore urge the machine translation community to move beyond aggregate metrics like FActScore and BLEU to include epistemic and register-based measurements, and we encourage downstream stakeholders (newsrooms, platforms, fact-checkers, and even the average consumer) to treat iterated LLM translation as a distinct source of systematic distortion, with human-moderated fact-checking prioritised for sensitive coverage.

References

Acerbi, A. and Stubbersfield, J. M. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, 2023.

Almahameed, Y. S., Bataineh, K. B. A., and Ammari, R. M. G. The Use of Passive Voice in News Reports for Political Purposes. *Journal of Language Teaching and Research*, 13(6):1196–1202, November 2022. ISSN 2053-0684. doi: 10.17507/jltr.1306.07. URL https://jltr.academypublication.com/index.php/jltr/article/view/4966.

Baker, M. *In other words: A coursebook on translation*. Routledge, 2018.

Biber, D., Johansson, S., Leech, G. N., Conrad, S., and Finegan, E. Grammar of spoken and written english, 2000.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong Model Collapse. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=et5l9qPUhm.

Farghal, M. and Al-Shorafat, M. O. The Translation of English Passives into Arabic: An Empirical Perspective. *Target. International Journal of Translation Studies*, 8(1): 97–118, 1996.

Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In Farkas, R., Vincze, V., Szarvas, G., Móra, G., and Csirik, J. (eds.), *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pp. 1–12, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-3001/.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Hyland, K. Hedging in scientific research articles. 1998.

Karttunen, L. Implicative Verbs. *Language*, 47(2):340–358, 1971. ISSN 0097-8507. doi: 10.2307/412084. URL https://www.jstor.org/stable/412084.

Kim, Y. Korean-English Differences of Communicative Preferences with Focus on the Subject Position. *The Journal of Translation Studies*, 8(2):241–258, 2007. ISSN 1229-795X. doi: 10.15749/jts.2007.8.2.010. URL https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001705273.

Kiparsky, P. and Kiparsky, C. Fact. In Bierwisch, M. and Heidolph, K. E. (eds.), *Progress in Linguistics: A Collection of Papers*, pp. 143–173. De Gruyter Mouton, Berlin, Boston, 1970. ISBN 9783111350219. doi: 10.1515/9783111350219.143. URL https://doi.org/10.1515/9783111350219.143.

Lee, D., Hwang, Y., Kim, Y., Park, J., and Jung, K. Are LLM-Judges Robust to Expressions of Uncertainty? Investigating the effect of Epistemic Markers on LLM-based Evaluation. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8962–8984, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.452. URL https://aclanthology.org/2025.naacl-long.452/.

Li, Y., Guerin, F., and Lin, C. LatestEval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18600–18607, 2024. URL https://dl.acm.org/doi/10.1609/aaai.v38i17.29822.

Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

Luo, J., Cherry, C., and Foster, G. To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. *Transactions of the Association for Computational Linguistics*, 12:355–371, 2024. URL https://aclanthology.org/2024.tacl-1.20/.

Miller, G. A. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1111/.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

Mohamed, A., Geng, M., Vazirgiannis, M., and Shang, G. LLM as a Broken Telephone: Iterative Generation Distorts Information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7493–7509, 2025. URL https://aclanthology.org/2025.acl-long.371/.

Perez, J., Kovač, G., Léger, C., Colas, C., Molinaro, G., Derex, M., Oudeyer, P.-Y., and Moulin-Frier, C. When LLMs play the telephone game: Cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings. In *The Thirteenth International Conference on Learning*

Representations, 2025. URL https://openreview.net/forum?id=fN8yLc3eA7.

Peterson, A. J. AI and the problem of knowledge collapse. *AI & SOCIETY*, 40(5):3249–3269, 2025.

Post, M. A Call for Clarity in Reporting BLEU Scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rathi, N., Jurafsky, D., and Zhou, K. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QsQatTzATT.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024. URL https://www.nature.com/articles/s41586-024-07566-y.

Siewierska, A. *The passive: A comparative linguistic analysis*. Routledge, 2024.

Team, G., Kamath, A., Ferret, J., et al. Gemma 3 Technical Report, March 2025. URL http://arxiv.org/abs/2503.19786. arXiv:2503.19786 [cs].

Wright, D., Masud, S., Moore, J., Yadav, S., Antoniak, M., Christensen, P. E., Park, C. Y., and Augenstein, I. Epistemic Diversity and Knowledge Collapse in Large Language Models. *arXiv preprint arXiv:2510.04226*, 2025. URL https://arxiv.org/abs/2510.04226.

Xiao, R., McEnery, T., and Qian, Y. Passive constructions in english and chinese: A corpus-based contrastive study. *Languages in contrast*, 6(1):109–149, 2006.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT, February 2020. URL http://arxiv.org/abs/1904.09675. arXiv:1904.09675 [cs].

Zhou, K., Jurafsky, D., and Hashimoto, T. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5506–5524, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.335. URL https://aclanthology.org/2023.emnlp-main.335/.

A. Epistemic Markers

The lists below are used by the epistemic-marker evaluator (Section 3.3). Items are stored as lemma forms; for regular verbs the regex appends `(?:s|ed|ing)?` at match time, while irregular forms are listed explicitly. Category overlap is intentional (e.g., *suggest* is both a hedge and a non-factive verb), following the literature.

Hedges (Hyland, 1998; Farkas et al., 2010). **Modals (5):** *may, might, could, would, should*. **Adverbs (41):** *apparently, arguably, approximately, around, broadly, conceivably, fairly, generally, largely, likely, mainly, maybe, mostly, nearly, occasionally, often, ostensibly, partially, partly, perhaps, plausibly, possibly, predominantly, presumably, probably, quite, rarely, rather, relatively, reportedly, reputedly, seemingly, seldom, sometimes, somewhat, supposedly, technically, typically, unlikely, usually, virtually*. **Adjectives (18):** *conceivable, debatable, doubtful, hypothetical, improbable, inconclusive, likely, plausible, possible, probable, questionable, speculative, uncertain, unclear, unlikely, unproven, unsettled, unsure*. **Verbs (24):** *appear, assume, believe, consider, doubt, estimate, expect, feel, hypothesize, imply, indicate, infer, postulate, predict, presume, propose, seem, speculate, suggest, suppose, surmise, suspect, tend, think*. **Phrases (11):** *more or less, in general, on the whole, to some extent, to a certain extent, up to a point, it would seem, it would appear, there is a possibility, raises the question, open to question*.

Boosters (Hyland, 1998). **Adverbs (32):** *absolutely, actually, admittedly, always, assuredly, basically, certainly, clearly, conclusively, decidedly, definitely, doubtless, essentially, evidently, explicitly, frequently, indeed, inevitably, manifestly, necessarily, never, obviously, patently, plainly, surely, truly, unambiguously, undeniably, undoubtedly, unequivocally, unmistakably, unquestionably*. **Verbs (8):** *confirm, demonstrate, determine, discern, establish, know, prove, show*. **Phrases (15):** *of course, no doubt, beyond doubt, without doubt, without question, in fact, as a matter of fact, it is clear, it is evident, it is obvious, it is known, it is a fact, the fact is, well known, well established*.

Factive verbs (Kiparsky & Kiparsky, 1970; Karttunen, 1971). **Truly factive (4):** *regret, resent, grasp, be aware*. **Semi-factive (6):** *discover, notice, realize, recognize, see, understand*. **Extended (10), adopted from corpus-NLP usage:** *acknowledge, admit, confirm, establish, prove, demonstrate, find, reveal, show, verify*. **Irregular forms (19):** *knew, found, showed, shown, proved, proven, saw, understood, realized, recognised, recognized, discovered, confirmed, established, revealed, acknowledged, admitted, verified, demonstrated*.

Non-factive verbs (Kiparsky & Kiparsky, 1970; Karttunen, 1971). **Reporting / attribution verbs (37):** *think, believe, suppose, imagine, hope, expect, claim, report, assert, assume, say, argue, maintain, contend, allege, deny, insist, declare, state, note, add, warn, suggest, indicate, imply, hint, speculate, suspect, announce, stress, emphasise, emphasize, urge, vow, promise, concede, acknowledge*. **Irregular forms (33):** *said, thought, believed, argued, maintained, contended, alleged, denied, insisted, declared, stated, noted, added, warned, suggested, indicated, implied, hinted, speculated, suspected, announced, stressed, urged, vowed, promised, conceded, acknowledged, reported, claimed, asserted, assumed, hoped, expected*.

Evidential markers. **Adverbs (9):** *allegedly, apparently, evidently, ostensibly, purportedly, reportedly, reputedly, seemingly, supposedly*. **Phrases (25):** *according to, as reported by, as stated by, as confirmed by, it is claimed, it has been claimed, it is alleged, it has been alleged, it is reported, it has been reported, it is understood, it has been said, it is believed, sources say, sources said, sources told, experts say, experts suggest, officials say, rumour has it, some believe, some say, some argue, widely believed, widely reported*.

Non-evidential markers. Operationalised as impersonal-source and direct-observation constructions, complementary to the evidential class. **Phrases (39):** *the fact that, it is known, it is established, it is well established, it is documented, data show, data shows, research show, research shows, evidence show, evidence shows, studies show, studies showed, statistics show, statistics shows, figures show, figures showed, analysis show, analysis shows, findings show, findings showed, records show, records showed, documents show, documents showed, footage show, footage shows, was seen, were seen, was spotted, were spotted, was filmed, was recorded, was witnessed, were witnessed, was observed, were observed, witnesses saw, bystanders saw*.

B. Iteration-0 vs. iteration-12 example

In this appendix, we present an article (KO-Qwen self-loop, UK topic, BBC source title *"King Charles not planning visits yet to riot-hit areas"*) which shows a 14→3 drop in epistemic-hedge count between iter-0 and iter-12 with 96% word preservation (691 → 666). Below we excerpt the iter-0 sentences containing hedges and pair each with its iter-12 counterpart. Italicised tokens are hedges in the lexicon of Appendix A.

1. *Iter-0:* "King Charles is not currently *expected* to make visits or official statements about the wave of rioting…"
   Iter-12: "It is extremely difficult for the king to identify areas where protests or unrest may occur, or to issue immediate official statements."

2. *Iter-0:* "…a response to the protests being left to the government, *rather* than a monarch who is *expected* to stay out of politics."
   Iter-12: "…government should lead responses, while the monarchy must maintain political neutrality."

3. *Iter-0:* "In such previous cases, royal visits to trouble spots have *tended* to follow after the wave of unrest and violence has settled."
   Iter-12: "In past similar situations, it was customary for royal family members to visit affected areas only after violence and chaos had significantly diminished."

4. *Iter-0:* "We need to hear his *considered* thoughts then about societal harmony."
   Iter-12: "…efforts to reevaluate social bonds must wait until conditions allow."

5. *Iter-0:* "…I *think* that if I were advising him I would *suggest* making that statement sooner *rather* than later."
   Iter-12: "This is precisely the right moment to discuss values such as cultural pluralism, pluralism, and republicanism…"

6. *Iter-0:* "*In general*, the monarchy does not comment on current political events," he said, *suggesting* any visits should come later.
   Iter-12: "…royal staff *typically* refrain from commenting on current political issues, and … royal visits should only occur after genuine de-escalation has taken place…"

7. *Iter-0:* "…where he is *expected* to spend much of the summer."
   Iter-12: "…plans to spend much of the summer in Scotland."

8. *Iter-0:* "…and it is *likely* that once the disorder has calmed down, there will be visits and attempts to offer reassurance to troubled areas."
   Iter-12: "…it is by no means inappropriate for him to seek out regions experiencing recent instability to offer comfort and support."

C. Degree Adverbs

The **degree adverb** evaluator uses spaCy (`en_core_web_sm`) (Honnibal et al., 2020) to identify adverbial modifier (advmod) dependencies whose lemma falls within a 14-word empirical lexicon, derived from the top-50 adverbial modifier lemmas observed in the BBC corpus and model outputs combined, excluding the comparative *more/most* and directional *far*. The full list includes *almost, clearly, completely, deeply, extremely, fully, much, nearly, particularly, rather, really, simply, truly, very*.

D. Bridge-language attribution lexicons

Table 3 gives the lexicons used in the evidential calque test (Section 4.2). For each language we include the high-frequency news-prose forms that explicitly mark a clause as sourced from a third party (*X said*, *according to Y*, *X reported / indicated / pointed out*); function words with broader uses (e.g., Chinese 从, French *de*, Korean 서) were not included. The lists were fixed before computing the rates in Table 1. As a coverage check, Table 1 reports that English evidentials in the back-translation co-occur with a listed bridge form 86–100% of the time across runs (mean 97%). These words also appear in 73–95% of iteration-12 bridge texts (mean 86%) across the 17 runs. For each iteration 12 article, an entry is considered

Table 3. Bridge-language attribution lexicons. Chinese, French, and Korean lists contain six items each; the Arabic list contains eight, since Modern Standard Arabic spreads attribution across more high-frequency forms.

Bridge	Items
Chinese	据 *jù* (according to), 根据 *gēnjù* (based on), 表示 *biǎoshì* (indicated), 称 *chēng* (said), 报道 *bàodào* (reported), 指出 *zhǐchū* (pointed out).
French	*selon* (according to), *d'après* (according to), *indique* (indicates), *déclar-* (declared / declaration), *rapport* (report), *souligne* (emphasises).
Korean	에따르면 *e ttareumyĕon* (according to), 의하면 *ŭihamyĕon* (according to), 밝혔 *palkyĕ-* (revealed), 전했 *chŏnhae-* (conveyed), 말했 *marhae-* (said), 보도 *podo* (reported).
Arabic	*wafqan*, *wafqā*, *biḥasab*, *ḥasab*, *ṭibqā* (all: according to); *qāla* (said), *afāda* (reported), *dhakara* (mentioned).

present if it appears as a substring of the bridge text. Chinese is romanised in Hanyu Pinyin, Korean in McCune–Reischauer, Arabic in ALA-LC.

The English back-translation side of the calque test matches the patterns *according to*, *based on*, *reportedly*, *source(s) said*, *it was/is reported*, *as reported*, *cited*, *quoted*, *in a/the report*.

E. Voice Statistics

Tables 4 and 5 detail the deltas for voice-related counts referenced in the main text.

Table 4. Net change (Δ) in active, agentless-passive, with-agent-passive, and passive-total density (per 1,000 words; iteration-0 baselines 83.4, 11.2, 1.9, and 13.1 respectively) and in the passive ratio (iteration-0 baseline 0.137), between iteration 0 and iteration 12, per run (600 articles each). Run labels combine the bridge language (CH/FR/KO/AR) with the model (Qwen, Gemma, Llama for self-loops) or setting (2P for bilingual two-player and MM for multilingual multiplayer).

Run	Active	Agentless	With-agent	Passive Total	Ratio
CH-Qwen	−4.11	−2.29	−0.74	−3.03	−0.023
CH-Llama	+0.85	−0.05	+0.12	+0.07	+0.002
CH-Gemma	−0.50	−1.08	−0.44	−1.53	−0.013
FR-Qwen	−5.43	−2.35	−0.41	−2.76	−0.018
FR-Llama	−3.53	+0.59	+0.00	+0.59	+0.012
FR-Gemma	−2.67	−0.89	−0.11	−1.00	−0.005
KO-Qwen	−9.91	−0.97	−0.96	−1.94	−0.005
KO-Llama	−5.43	+2.50	−0.39	+2.11	+0.029
KO-Gemma	−4.37	+0.23	−0.57	−0.34	+0.004
AR-Qwen	−12.48	−1.31	−0.94	−2.26	−0.004
AR-Llama	−5.08	−0.12	−0.23	−0.35	+0.004
AR-Gemma	−1.35	−1.36	−0.63	−1.99	−0.016
CH-2P	−2.79	+0.40	−0.10	+0.30	+0.006
FR-2P	−6.19	−0.92	+0.00	−0.92	+0.001
KO-2P	−6.49	+1.65	−0.57	+1.09	+0.018
AR-2P	−10.26	+1.80	−0.45	+1.35	+0.027
MM	−7.97	−1.10	−0.25	−1.35	−0.002

F. Bootstrap standard errors

For each (run, metric) we collect the per-article rate at each iteration (per 1,000 words, except the passive ratio which is unitless), resample articles within the run with replacement $B{=}2{,}000$ times, and report the SE of the bootstrapped Δ. The Evidential column uses the full evidential lexicon (Appendix A); the calque-test counterpart in Table 1 uses the narrower attribution-pattern subset.

Table 5. Net change (Δ) in nominalisation and total verb/auxiliary density (per 1,000 words) between iteration 0 (baseline 33.0 and 195.1 respectively) and iteration 12, per run (600 articles each). Run labels combine the bridge language (CH/FR/KO/AR) with the model (Qwen, Gemma, Llama for self-loops) or setting (2P for bilingual two-player and MM for multilingual multiplayer).

Run	Nominalisation	Aux/Verb
CH-Qwen	$+9.78$	-6.26
CH-Llama	$+2.40$	$+3.26$
CH-Gemma	$+5.77$	$+0.26$
FR-Qwen	$+8.70$	-7.54
FR-Llama	$+0.90$	-6.93
FR-Gemma	$+3.86$	-5.71
KO-Qwen	$+16.09$	-9.95
KO-Llama	$+4.95$	-7.66
KO-Gemma	$+7.51$	-5.11
AR-Qwen	$+14.79$	-14.59
AR-Llama	$+6.49$	-16.16
AR-Gemma	$+3.25$	-7.46
CH-2P	$+3.32$	-2.34
FR-2P	$+4.35$	-11.36
KO-2P	$+7.34$	-5.86
AR-2P	$+9.69$	-15.44
MM	$+8.94$	-13.78

Table 6. Bootstrap standard errors on iter-0 to iter-12 deltas for epistemic markers and degree adverbs. B=2,000 resamples per run, articles resampled with replacement.

Run	Evidential	Factive	Hedge	Booster	Deg. adv
CH-Qwen	$+0.089_{\pm 0.071}$	$+1.179_{\pm 0.166}$	$-0.629_{\pm 0.274}$	$+2.342_{\pm 0.224}$	$+1.649_{\pm 0.162}$
CH-Gemma	$+0.030_{\pm 0.065}$	$+0.585_{\pm 0.129}$	$+0.410_{\pm 0.237}$	$+0.510_{\pm 0.127}$	$+0.225_{\pm 0.116}$
CH-Llama	$+0.332_{\pm 0.130}$	$+0.578_{\pm 0.234}$	$+2.533_{\pm 1.704}$	$+0.190_{\pm 0.256}$	$-0.657_{\pm 0.163}$
FR-Qwen	$+0.769_{\pm 0.129}$	$+0.600_{\pm 0.151}$	$+0.003_{\pm 0.316}$	$+1.287_{\pm 0.147}$	$+1.317_{\pm 0.136}$
FR-Gemma	$+0.515_{\pm 0.066}$	$+0.096_{\pm 0.115}$	$+0.439_{\pm 0.182}$	$+0.131_{\pm 0.099}$	$+0.178_{\pm 0.087}$
FR-Llama	$+0.356_{\pm 0.107}$	$+0.059_{\pm 0.254}$	$-0.357_{\pm 0.279}$	$-0.267_{\pm 0.143}$	$-0.080_{\pm 0.122}$
KO-Qwen	$+0.794_{\pm 0.091}$	$+1.034_{\pm 0.199}$	$-1.853_{\pm 0.306}$	$+2.297_{\pm 0.180}$	$+1.916_{\pm 0.173}$
KO-Gemma	$+0.133_{\pm 0.069}$	$+0.704_{\pm 0.162}$	$-0.500_{\pm 0.246}$	$+0.050_{\pm 0.148}$	$+0.327_{\pm 0.134}$
KO-Llama	$+1.119_{\pm 0.192}$	$+0.188_{\pm 0.309}$	$-1.598_{\pm 0.426}$	$-0.489_{\pm 0.291}$	$-0.993_{\pm 0.174}$
AR-Qwen	$+0.362_{\pm 0.078}$	$+1.521_{\pm 0.190}$	$-1.526_{\pm 0.282}$	$+1.826_{\pm 0.158}$	$+0.101_{\pm 0.136}$
AR-Gemma	$+0.271_{\pm 0.068}$	$+0.027_{\pm 0.114}$	$-0.626_{\pm 0.195}$	$-0.229_{\pm 0.099}$	$-0.106_{\pm 0.114}$
AR-Llama	$+1.141_{\pm 0.161}$	$-0.148_{\pm 0.217}$	$-0.296_{\pm 0.450}$	$-0.424_{\pm 0.218}$	$-1.215_{\pm 0.169}$
CH-2P	$+0.042_{\pm 0.080}$	$+0.884_{\pm 0.195}$	$-0.187_{\pm 0.355}$	$+0.416_{\pm 0.191}$	$+0.279_{\pm 0.152}$
FR-2P	$+0.698_{\pm 0.103}$	$-0.063_{\pm 0.162}$	$-0.185_{\pm 0.298}$	$-0.101_{\pm 0.146}$	$+0.562_{\pm 0.137}$
KO-2P	$+1.073_{\pm 0.339}$	$+0.276_{\pm 0.219}$	$-1.291_{\pm 0.388}$	$-0.065_{\pm 0.199}$	$+0.599_{\pm 0.197}$
AR-2P	$+0.268_{\pm 0.094}$	$+1.907_{\pm 0.453}$	$-1.139_{\pm 0.342}$	$+2.162_{\pm 0.485}$	$-0.748_{\pm 0.137}$
MM	$+1.171_{\pm 0.853}$	$+0.802_{\pm 0.186}$	$-0.394_{\pm 0.317}$	$+0.806_{\pm 0.177}$	$+1.316_{\pm 0.156}$

Table 7. Bootstrap standard errors for voice and agency deltas (cf. Tables 4 and 5). Active / Agentless / With-agent / Nom. / Aux/Verb are per-1,000-words rates; Ratio is the unitless passive ratio. B=2,000 resamples per run.

Run	Active	Agentless	With-agent	Ratio	Nom.	Aux/Verb
CH-Qwen	$-4.107_{\pm 0.497}$	$-2.286_{\pm 0.251}$	$-0.739_{\pm 0.091}$	$-0.0229_{\pm 0.0028}$	$+9.705_{\pm 0.463}$	$-6.406_{\pm 0.740}$
CH-Gemma	$-0.499_{\pm 0.413}$	$-1.082_{\pm 0.217}$	$-0.445_{\pm 0.090}$	$-0.0129_{\pm 0.0024}$	$+5.746_{\pm 0.393}$	$+0.293_{\pm 0.618}$
CH-Llama	$+0.852_{\pm 0.984}$	$-0.053_{\pm 0.405}$	$+0.123_{\pm 0.192}$	$+0.0028_{\pm 0.0051}$	$+2.470_{\pm 0.871}$	$+3.038_{\pm 2.038}$
FR-Qwen	$-5.433_{\pm 0.389}$	$-2.350_{\pm 0.216}$	$-0.411_{\pm 0.081}$	$-0.0181_{\pm 0.0023}$	$+8.649_{\pm 0.436}$	$-7.683_{\pm 0.741}$
FR-Gemma	$-2.665_{\pm 0.302}$	$-0.886_{\pm 0.161}$	$-0.110_{\pm 0.058}$	$-0.0055_{\pm 0.0018}$	$+3.835_{\pm 0.288}$	$-5.767_{\pm 0.491}$
FR-Llama	$-3.532_{\pm 0.599}$	$+0.586_{\pm 0.277}$	$+0.002_{\pm 0.080}$	$+0.0120_{\pm 0.0036}$	$+1.015_{\pm 0.508}$	$-6.866_{\pm 0.963}$
KO-Qwen	$-9.914_{\pm 0.580}$	$-0.972_{\pm 0.325}$	$-0.963_{\pm 0.094}$	$-0.0047_{\pm 0.0035}$	$+16.161_{\pm 0.558}$	$-10.044_{\pm 1.025}$
KO-Gemma	$-4.372_{\pm 0.454}$	$+0.231_{\pm 0.245}$	$-0.569_{\pm 0.083}$	$+0.0035_{\pm 0.0026}$	$+7.496_{\pm 0.410}$	$-5.231_{\pm 0.664}$
KO-Llama	$-5.429_{\pm 1.102}$	$+2.497_{\pm 0.527}$	$-0.389_{\pm 0.146}$	$+0.0328_{\pm 0.0063}$	$+4.977_{\pm 0.843}$	$-7.810_{\pm 1.719}$
AR-Qwen	$-12.482_{\pm 0.456}$	$-1.312_{\pm 0.249}$	$-0.943_{\pm 0.091}$	$-0.0036_{\pm 0.0030}$	$+14.723_{\pm 0.480}$	$-14.641_{\pm 0.741}$
AR-Gemma	$-1.350_{\pm 0.370}$	$-1.358_{\pm 0.183}$	$-0.633_{\pm 0.087}$	$-0.0161_{\pm 0.0021}$	$+3.307_{\pm 0.338}$	$-7.478_{\pm 0.523}$
AR-Llama	$-5.078_{\pm 0.839}$	$-0.120_{\pm 0.410}$	$-0.227_{\pm 0.274}$	$+0.0044_{\pm 0.0051}$	$+6.302_{\pm 0.783}$	$-16.102_{\pm 1.444}$
CH-2P	$-2.787_{\pm 0.724}$	$+0.403_{\pm 0.416}$	$-0.104_{\pm 0.117}$	$+0.0070_{\pm 0.0039}$	$+3.325_{\pm 0.558}$	$-2.337_{\pm 1.168}$
FR-2P	$-6.185_{\pm 0.577}$	$-0.919_{\pm 0.297}$	$-0.001_{\pm 0.101}$	$+0.0015_{\pm 0.0035}$	$+4.336_{\pm 0.482}$	$-11.444_{\pm 0.881}$
KO-2P	$-6.485_{\pm 0.779}$	$+1.655_{\pm 0.442}$	$-0.566_{\pm 0.130}$	$+0.0201_{\pm 0.0042}$	$+7.355_{\pm 0.629}$	$-5.849_{\pm 1.559}$
AR-2P	$-10.257_{\pm 0.675}$	$+1.802_{\pm 0.429}$	$-0.450_{\pm 0.118}$	$+0.0276_{\pm 0.0043}$	$+9.725_{\pm 0.655}$	$-15.492_{\pm 1.042}$
MM	$-7.974_{\pm 0.615}$	$-1.099_{\pm 0.291}$	$-0.253_{\pm 0.120}$	$-0.0002_{\pm 0.0038}$	$+8.877_{\pm 0.574}$	$-13.902_{\pm 1.249}$