LEARNING NOT TO LEARN: WHEN EVOLVING NORMS MEET THOSE THAT OUGHT TO BE PRESERVED

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding social norms helps large language models (LLMs) provide culturally appropriate responses that can influence human judgments and values. As norms change, LLMs must revise outdated judgments while retaining knowledge that remains valid. Existing continual learning methods do not explicitly distinguish norms that require revision from those that remain valid, risking outdated judgments or unnecessary forgetting. We propose a continual learning framework guided by normative validity, including 1) a mechanism that *distinguishes norms requiring revision from those remaining valid*, and 2) an update constraint that *protects the latter*. We construct a dataset spanning cultures and three temporal stages from the International Social Survey Programme (ISSP) and General Social Survey (GSS). Each scenario includes its country and temporal context, together with a reference judgment and explanation. Comparisons across stages identify stable norms for protection while evolving norms and novel norms guide learning. When these objectives conflict, our update rule adjusts the learning direction to reduce interference. Across three stages, relative improvements in final stage macro F1 reach 13.29% on Qwen and 12.94% on Llama. Improvements in retaining previously learned norms that remain valid exceed those of the strongest baselines by 8.16% on Qwen and 23.02% on Llama. Ablations support the contributions of norm status distinctions and update constraints, while gains on unseen scenarios support generalization. By tying knowledge preservation to continued normative validity, our framework supports adaptation to social change while maintaining cultural continuity.

1 INTRODUCTION

Cultural understanding and social adaptation are essential for large language models to make appropriate judgments in social interactions (Fung et al., 2023). Their importance grows as migration and digital platforms bring cultural expectations into contact and sometimes conflict (Erten et al., 2018; Brady & Crockett, 2024). Despite technical advances in code generation and productivity (Wang et al., 2021; Noy & Zhang, 2023), the soft power of LLMs in understanding cultural expectations and adapting to changing norms remains underinvestigated (Li et al., 2024). Social norms govern judgments of appropriate conduct, yet their applicability evolves as generational shifts interact with broader social and policy changes (Gelfand et al., 2024; González et al., 2025). In applications involving social judgments, outdated normative knowledge can impose obsolete expectations, whereas indiscriminate updates can erase norms that communities still uphold. The central challenge is to preserve previously learned norms when they remain valid in the relevant cultural context, while revising their current application when expectations change. This demands continual adaptation that distinguishes social change from knowledge that remains valid.

Our evaluation of 46 LLMs across ten providers reveals difficulties in judging social norms across time (Appendix A, Figures 5–7). They fail to consistently judge evolving norms in current contexts and recover appropriate judgments for historical contexts. Addressing this gap requires resolving two competing risks during learning, as preserving obsolete judgments can obstruct valid change, whereas learning new norms can disrupt knowledge that remains applicable (Jang et al., 2022). Cultural adaptation methods establish the importance of social context (AlKhamissi et al., 2024), and continual learning methods protect previous knowledge through replay or update constraints (De Lange et al., 2022). However, protection based on prior learning alone does not deter-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mine whether a historical social judgment should still govern the present. The unresolved question is how to *update evolving normative knowledge* while explicitly *protecting norms whose validity persists*.

To address these problems, we propose a constrained continual learning framework for knowledge updates under dynamic social norms, as illustrated in Figure 1.

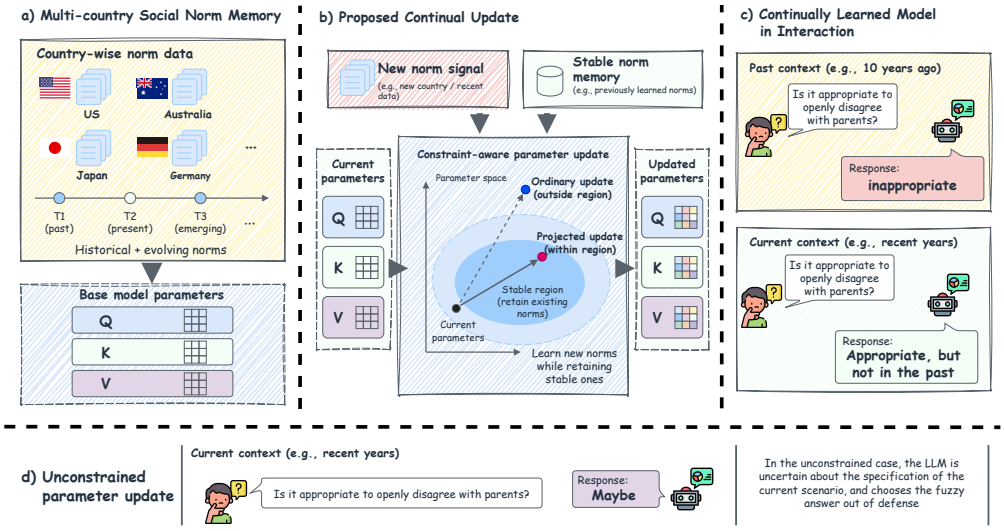


Figure 1. Overview of the proposed framework. **a)** Normative judgments are associated with scenarios situated in specific countries and temporal stages. **b)** Previously learned norms that remain valid supply the memory constraint, while evolving and novel norms guide adaptation of the model parameters. **c)** Illustrative interactions show judgments under past and current time contexts. **d)** The corresponding illustration with unconstrained updating shows an uncertain response in the current context.

Our framework represents normative knowledge through judgments and explanations attached to scenarios within country and temporal contexts. Comparisons of survey attitudes across stages determine whether previously encountered norms remain stable or have evolved. Stable examples define the protected memory, whereas evolving norms and those first encountered in the learning stream supply adaptation examples. We formulate this process as minimizing adaptation loss subject to a bound on the increase in stable-norm loss, and derive a closed-form update using Lagrangian methods and Karush–Kuhn–Tucker (KKT) conditions. When an update conflicts with the local stability constraint, the rule corrects its direction to protect still-valid knowledge. This makes *continued validity*, rather than prior acquisition alone, the basis for protecting normative knowledge during model updates. The resulting framework advances the capacity of language models to respond to changing expectations while respecting the cultural contexts in which existing norms remain applicable.

Our contributions are threefold.

- **Data.** We construct a three-stage cross-cultural temporal social-norm dataset grounded in longitudinal survey evidence from ISSP and GSS, explicitly characterizing initial, stable, evolving, and novel norms and supporting unseen-scenario evaluation. We will publicly release the processed dataset upon acceptance.
- **Method.** We propose a constrained continual learning framework for dynamic social norms that directs knowledge updates toward evolving and novel norms while protecting stable norms, with a closed-form update derived from the local constrained objective.
- **Results.** Across Qwen and Llama, our method outperforms the evaluated continual learning baselines in current norm judgments and stable-norm retention; unseen-scenario and ablation results further establish its ability to combine adaptation, preservation, and transfer to new social situations.

2 Related Work

Cultural norm reasoning and adaptation ground model judgments in sociocultural context, recognizing that an action's acceptability depends on the community and situation. Rao et al. (2025) evaluate cultural adaptability through judgments conditioned on explicit norms within national and value contexts, revealing persistent difficulties even when the relevant norms are supplied. Liu et al. (2025) move from evaluation to adaptation by using simulated social interactions to generate culturally situated training conversations. These approaches establish the importance of contextual knowledge and cultural learning, but their reported evaluations do not jointly test adaptation to successive normative changes and retention of norms that remain valid. Our work extends this contextual perspective to successive stages by distinguishing stable norms from those that evolve or enter the learning stream.

Continual learning addresses sequential adaptation by replaying earlier examples or constraining updates to reduce interference with previous tasks (Rolnick et al., 2019). Wang et al. (2024b) select replay data based on instruction similarity while accounting for complexity and diversity, improving retention during continual instruction tuning. At the optimization level, Lopez-Paz & Ranzato (2017) constrain remembered-task losses, whereas Chaudhry et al. (2019) use an averaged reference gradient to reduce computational cost. These mechanisms provide foundations for preserving prior capabilities, but their standard formulations do not distinguish historical normative judgments that remain applicable from those superseded by change. Applied without this distinction, retention can oppose desirable adaptation. We build on constrained-update methods by selecting protected examples according to continued normative validity. Previously learned stable norms supply the constraint, whereas current examples of evolving and novel norms supply the adaptation objective. The contribution therefore lies in selective protection and its temporal formulation, rather than gradient projection alone.

Knowledge editing offers a complementary route to updating outdated information while limiting collateral changes (Meng et al., 2022; 2023). Wang et al. (2024a) separate pretrained and edited knowledge into routed parametric memories to support lifelong editing, while Fang et al. (2025) project parameter perturbations into a null space associated with preserved knowledge. Both address the tension between update success and preservation, but their evaluated settings do not establish how to maintain social judgments whose validity depends jointly on culture and time. Our framework shares their selective-update objective, while expressing preservation as a first-order bound on stable-norm loss during continual adaptation. Across these directions, our central distinction is to condition protection on whether a norm remains applicable, coupling adaptation to evolving norms with explicit protection of stable norms and excluding obsolete judgments from that protection.

3 Problem Formulation

We study how language models continually adapt to changing social norms. Using existing social survey results, we distinguish new or evolving norms to be learned from stable norms to be preserved.

3.1 Task Definition

Each record $s = (t, u, v, w, x, y, z)$ links a scenario w to country context u and temporal stage t. The topic identifier v links the same normative issue across stages. Cultural context enters the model through country u and time t, supplied alongside scenario w. Given (u, t, w), the model predicts judgment x with explanation y. The judgment is $x \in \{\text{appropriate}, \text{context-dependent}, \text{inappropriate}\}$.

Survey attitudes for each country and topic pair (u, v) determine whether previously learned norms remain valid across stages. Norm status z is training metadata that assigns examples to adaptation or protection, rather than a prediction target. Stable norms retain their validity, whereas evolving norms permit revised judgments in the current context. Novel norms mark first exposure in the learning stream rather than historical emergence, while D_1 contains initial norms.

At each stage $t > 1$, adaptation data D_t^A contain current examples of evolving norms together with novel norms. Memory M_t^S contains previously learned examples whose norms remain stable. Training reduces $L_{\text{new}}(\theta; D_t^A)$ while limiting increases in $L_{\text{stable}}(\theta; M_t^S)$. Thus, prior knowledge

is protected when it remains valid, while its current application can change when norms evolve. Detailed status criteria and examples across time are provided in Appendix C.

3.2 Datasets

We use the 1994, 2002, and 2022 waves of the International Social Survey Programme's Family and Changing Gender Roles Cumulation (GESIS, ZA8795 v1.0.0) and the US General Social Survey (NORC, 1972–2024 cumulative release 3a) (ISSP Research Group, 2026; NORC at the University of Chicago, 2026). Based on these surveys, we construct datasets covering 18 country contexts and 17 norm topics. Each stage includes training data with separate validation and test sets, while an additional set supports evaluation on unseen scenarios. Table 1 reports their sizes.

To derive labels from the surveys, we group responses to each question by country and year. Answers are coded from -1 (opposition) to 1 (support) and averaged using the survey weights. We exclude groups with fewer than 100 valid responses. Under our predefined labeling rules, mean scores of at least 0.20 indicate support, whereas scores of at most -0.20 indicate opposition. Scores between these thresholds indicate a neutral attitude.

We use GPT-5.4 to construct scenarios that either support or oppose the surveyed positions, as well as scenarios expressing mixed positions. A normative judgment expresses how a scenario relates to the survey attitude for its country and year. The scenario is labeled appropriate when its position agrees with that attitude and inappropriate when it disagrees. Mixed-position scenarios and those associated with neutral attitudes are labeled context-dependent. We manually verified a randomly selected 30% of the constructed data. Figure 2 illustrates the construction and verification process. Due to space constraints, detailed data fields and dataset analyses are provided in Appendix D.

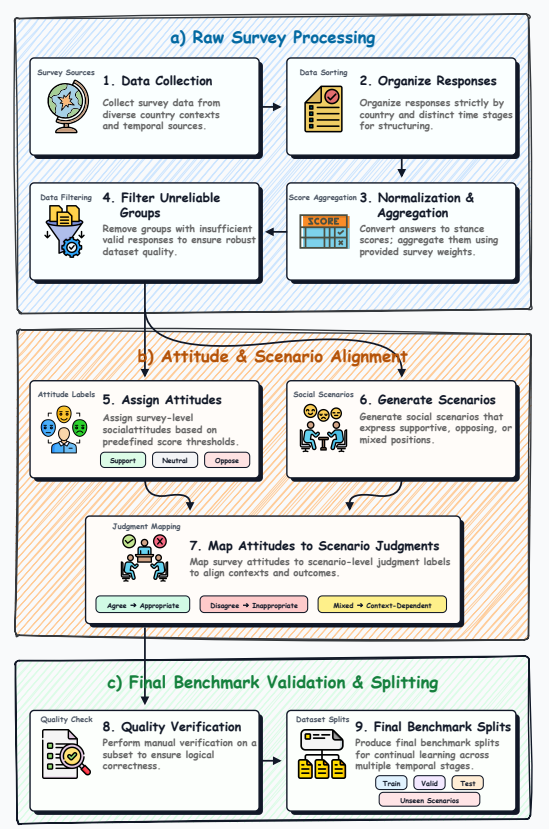


Figure 2. Data construction and verification pipeline. **a)** Survey responses are processed to derive attitudes for each country and year. **b)** Scenarios are aligned with these attitudes to assign normative judgments. **c)** Manual verification precedes the final dataset splits.

4 Method

Our framework incorporates cultural context through the conditional mapping from (u, t, w) to (x, y) defined in Section 3. Figure 1 summarizes how examples participate

Table 1. Dataset sizes across three stages. The unseen set contains held-out social scenarios evaluated at T3.

Stage	Training	Validation	Test	Unseen
T1	1,500	250	250	–
T2	1,500	250	250	–
T3	1,500	250	250	–
Total	4,500	750	750	150

in learning according to norm status z. Current examples of evolving and novel norms form *adaptation examples* D_t^A, while previously learned norms that remain valid form *stable examples* M_t^S. Only the latter constrain parameter updates, leaving the current application of evolving norms open to revision.

4.1 Constrained Learning

Let θ denote the model's trainable parameters and $\bar{\theta}$ their values before an update. The learning objective is

$$\begin{aligned} \min_{\theta} \quad & L_{\text{new}}(\theta; D_t^A) \\ \text{s.t.} \quad & L_{\text{stable}}(\theta; M_t^S) \leq L_{\text{stable}}(\bar{\theta}; M_t^S) + \epsilon, \end{aligned} \tag{1}$$

Here $\epsilon \geq 0$ is the permitted increase in loss on M_t^S. Protection is expressed through predictive performance on stable examples rather than by freezing all parameters associated with earlier learning.

For $d = \theta - \bar{\theta}$, define $g_n = \nabla_\theta L_{\text{new}}(\bar{\theta})$ and $g_s = \nabla_\theta L_{\text{stable}}(\bar{\theta})$. The gradient g_n directs adaptation to current examples, while g_s defines the local constraint protecting stable norms. First-order Taylor expansion approximates the two loss changes by $g_n^\top d$ and $g_s^\top d$. Adding a quadratic term to limit the update magnitude gives

$$\min_{d} \quad g_n^\top d + \frac{1}{2\eta}\|d\|^2 \qquad \text{s.t.} \quad g_s^\top d \leq \epsilon, \tag{2}$$

Here $\eta > 0$ is the learning rate. The objective permits revision of evolving norms, while the constraint limits interference with previously learned norms that remain valid.

4.2 Closed-Form Update

The Lagrangian of Equation 2 is

$$\mathcal{L}(d, \lambda) = g_n^\top d + \frac{1}{2\eta}\|d\|^2 + \lambda(g_s^\top d - \epsilon), \qquad \lambda \geq 0. \tag{3}$$

Stationarity gives $d = -\eta(g_n + \lambda g_s)$. Combining this relation with the KKT feasibility and complementary slackness conditions yields

$$d^* = -\eta g_n - \frac{\max(0, -\eta g_s^\top g_n - \epsilon)}{\|g_s\|^2}\, g_s. \tag{4}$$

When the ordinary update $-\eta g_n$ satisfies the constraint, the correction is zero. Otherwise, it projects the update onto the boundary $g_s^\top d = \epsilon$, limiting the predicted increase in stable-norm loss. If $g_s = 0$, we use the ordinary update without division.

We update the model's trainable parameters using $\theta \leftarrow \bar{\theta} + d^*$ throughout the learning stream, recomputing both gradients at each step. The constraint bounds the stable-norm loss change to first order at each update. For $\epsilon = 0$, the update also satisfies $g_n^\top d^* \leq 0$, preserving a non-increasing first-order adaptation loss. Appendix B gives the complete derivation and its conditions.

5 Experiments

We evaluated adaptation to evolving norms alongside retention of norms that remain valid. Further experiments tested transfer to unseen scenarios and the contributions of both framework components.

5.1 Experimental Setup

Data and models. We trained sequentially on $D_1 \rightarrow D_2 \rightarrow D_3$ (Section 3.2), using 1,500 training examples per stage. Each stage had 250 validation examples and 250 test examples. The 150 unseen scenarios were used only for evaluation at T3. The model received country and temporal context alongside each scenario to predict a judgment with an explanation. Qwen3-8B (Qwen Team, 2025) was the primary model, with Llama-3.1-8B-Instruct (Llama Team, 2024) also evaluated in the main experiment. We report means and standard deviations over ten random seeds numbered 1 through 10.

Baselines and ablations. Sequential LoRA (Hu et al., 2022) learns successive stages, whereas Replay LoRA (Hu et al., 2022; Rolnick et al., 2019) also reuses earlier examples. GEM (Lopez-Paz & Ranzato, 2017) constrains updates using earlier task memories, whereas A-GEM (Chaudhry et al., 2019) uses an averaged reference gradient. The remaining baselines were InsCL (Wang et al., 2024b) and GainLoRA (Liang et al., 2025). Ablations removed either the update constraint (w/o Constraint) or the norm status distinction (w/o Stable/Evolving Split).

5.2 Evaluation Metrics

We evaluated all final metrics after T3 training, with higher values indicating better performance. Accuracy and F1 are percentages, whereas Stable-BWT is measured in percentage points.

Current Macro-F1 and Accuracy. Both metrics measure judgment quality on the complete D_3 test set.

$$\text{Macro-F1} = \frac{1}{3}\sum_{c=1}^{3} F1_c, \qquad F1_c = \frac{2P_cR_c}{P_c + R_c}, \tag{5}$$

Here P_c and R_c denote precision and recall for judgment class c (Opitz, 2024). Accuracy measures the fraction of correct predictions across all test examples.

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{test}}}. \tag{6}$$

Evolving Macro-F1. Macro-F1 on the evolving norm subset of D_3 measures adaptation to changed norms.

Stable-BWT. This metric measures accuracy changes on earlier norms that remain valid.

$$\text{Stable-BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(R^{\text{stable}}_{T,i} - R^{\text{stable}}_{i,i}\right), \qquad T = 3. \tag{7}$$

Here $R^{\text{stable}}_{t,i}$ denotes accuracy on the same stable norm test subset of D_i after stage t. Positive values indicate improvement, whereas negative values indicate forgetting (Lopez-Paz & Ranzato, 2017).

Unseen Macro-F1. We apply Equation 5 to all 150 unseen T3 scenarios and separately to their evolving norm subset. Unseen Macro-F1 and Unseen Evolving Macro-F1 measure transfer to new social situations.

Historical evolving-norm F1 (HF1). After T3, we average Macro-F1 on earlier evolving norms under their original temporal contexts and labels.

$$\text{HF1} = \frac{F1(D_1^{\text{evolving}}) + F1(D_2^{\text{evolving}})}{2}. \tag{8}$$

This measures retention of historical judgments for norms whose current application has changed.

5.3 Main Results

Our method achieved the highest mean scores on all four metrics for both models (Table 2). InsCL was the strongest baseline on every metric for both models. On Qwen3-8B, Current Macro-F1 reached 83.29%, exceeding InsCL by 1.10 percentage points. Evolving Macro-F1 reached 83.54%,

Table 2. Results after T3 training, reported as Mean ± standard deviation over 10 seeds numbered 1 through 10. Bold indicates the best result for each metric within each model.

Method	Accuracy (%) ↑	Current Macro-F1 (%) ↑	Evolving Macro-F1 (%) ↑	Stable-BWT (pp) ↑
Qwen3-8B				
Sequential LoRA	73.56 ± 2.13	73.52 ± 2.21	73.41 ± 2.84	-3.06 ± 0.99
Replay LoRA	76.52 ± 4.84	76.51 ± 4.80	76.32 ± 3.42	0.75 ± 1.84
GEM	78.84 ± 4.25	78.81 ± 4.22	79.48 ± 2.18	3.63 ± 1.49
A-GEM	77.76 ± 3.23	77.68 ± 3.28	79.56 ± 2.97	2.35 ± 1.07
InsCL	82.20 ± 1.00	82.19 ± 1.01	82.51 ± 0.54	4.78 ± 0.47
GainLoRA	80.00 ± 1.03	79.97 ± 1.08	80.43 ± 1.09	4.48 ± 0.55
Ours	$\mathbf{83.28 \pm 2.82}$	$\mathbf{83.29 \pm 2.81}$	$\mathbf{83.54 \pm 3.89}$	$\mathbf{5.17 \pm 0.67}$
Llama-3.1-8B-Instruct				
Sequential LoRA	71.36 ± 1.81	71.11 ± 1.83	70.10 ± 1.46	-1.43 ± 1.17
Replay LoRA	74.68 ± 4.54	74.51 ± 4.56	73.11 ± 5.98	0.70 ± 2.61
GEM	76.64 ± 3.05	76.55 ± 3.08	75.74 ± 6.50	0.86 ± 1.32
A-GEM	78.28 ± 1.60	78.18 ± 1.65	80.24 ± 1.32	2.36 ± 3.09
InsCL	79.16 ± 0.74	79.09 ± 0.81	80.51 ± 2.02	2.65 ± 0.82
GainLoRA	78.64 ± 1.04	78.69 ± 1.04	79.00 ± 2.18	2.28 ± 0.27
Ours	$\mathbf{80.44 \pm 1.73}$	$\mathbf{80.31 \pm 1.70}$	$\mathbf{82.06 \pm 1.43}$	$\mathbf{3.26 \pm 1.09}$

compared with 82.51% for InsCL. Stable-BWT reached 5.17 percentage points, compared with 4.78 for InsCL.

On Llama-3.1-8B-Instruct, Current Macro-F1 reached 80.31%, exceeding InsCL by 1.22 percentage points. Evolving Macro-F1 reached 82.06%, exceeding InsCL by 1.55 percentage points. Stable-BWT reached 3.26 percentage points, compared with 2.65 for InsCL. These results show improved judgments under current cultural contexts while strengthening knowledge of earlier norms that remain valid.

5.4 Unseen Scenarios

After T3, we evaluated the model on 150 unseen scenarios (Figure 3). Unseen Macro-F1 reached 76.16%, exceeding the strongest baseline InsCL at 75.55% by 0.61 percentage points. Unseen Evolv-

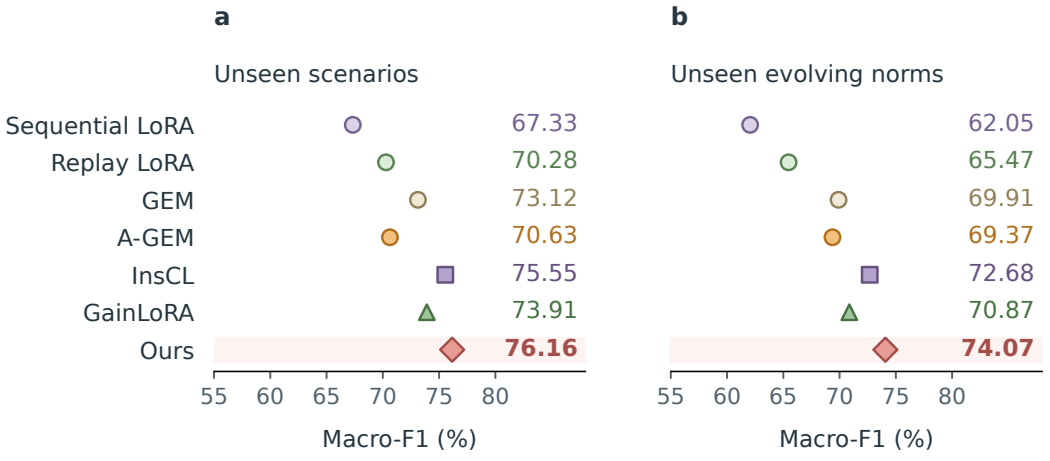


Figure 3. Generalization to unseen scenarios after T3, measured by Macro-F1 on all 150 scenarios in panel a and their evolving norm subset in panel b. Both metrics are percentages, with higher values indicating better performance.

ing Macro-F1 reached 74.07%, exceeding InsCL at 72.68% by 1.39 percentage points. The larger gain on evolving norms shows that updated judgments transferred to new scenarios within the supplied cultural contexts.

5.5 Ablation Study

Removing either component reduced all four ablation metrics (Figure 4). Without the constraint, Current Macro-F1 fell by 4.18 percentage points and Evolving Macro-F1 by 4.50. Stable-BWT fell from 5.17 to -0.70 percentage points, indicating forgetting of earlier norms that remained valid. Without the norm status distinction, Current Macro-F1 fell from 83.29% to 81.29%. Evolving Macro-F1 fell from 83.54% to 81.79%, while Stable-BWT decreased from 5.17 to 3.83 percentage points. Norm status identifies valid prior knowledge for protection, and the constraint protects it during adaptation. The adaptation objective allows revised judgments when norms evolve within a cultural context.

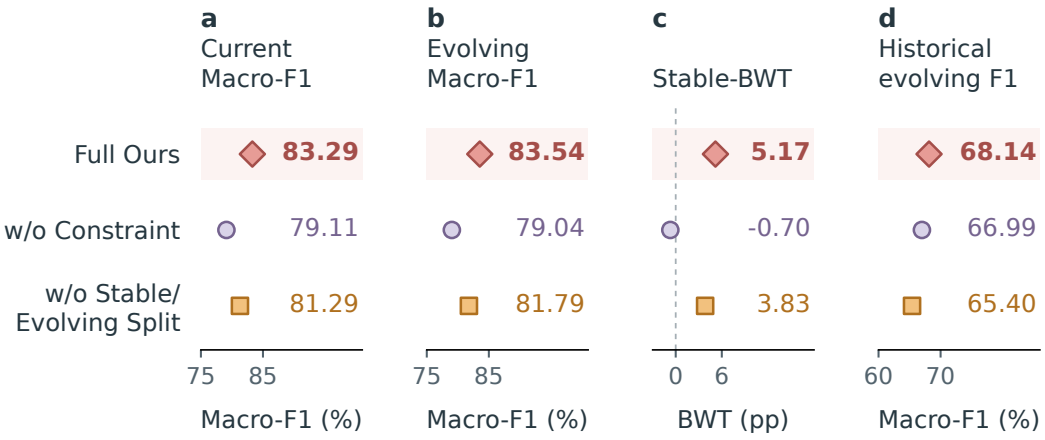


Figure 4. Ablation of the update constraint and norm status distinction after T3 training. Panels a and b report Current Macro-F1 and Evolving Macro-F1, respectively. Panel c reports Stable-BWT, while panel d reports HF1 under the original T1 and T2 contexts and labels. F1 values are percentages, whereas Stable-BWT is measured in percentage points.

HF1 reached 68.14%, compared with 66.99% without the constraint and 65.40% without the norm status distinction. The full method better retained historical judgments under their original contexts while learning the current application of evolving norms.

6 Discussion

Our findings show that learning evolving norms and preserving stable norms can improve together. The gains across both models identify selective protection as a practical approach to adapting social judgments over time. Generalization tests and component ablations provide further support.

Stable norms can improve during adaptation. After training through T3, Sequential LoRA has a Stable-BWT of -3.06 percentage points on Qwen, indicating reduced accuracy on earlier stable norms (Table 2). The replay baseline and both methods using gradient constraints achieve positive backward transfer, while our method reaches $+5.17$ percentage points compared with $+3.63$ for GEM. Thus, preserving valid norms need not mean merely limiting their degradation, because their accuracy can improve as the model learns new and evolving norms. Together with the gains in evolving-norm F1, this result demonstrates that selective protection can support both retention and adaptation.

Distinguishing stable from evolving norms is critical. The ablations show that deciding what to preserve matters alongside how updates are constrained. Removing either component reduced current adaptation and stable-norm retention; removing the constraint changed positive backward transfer into forgetting. Our framework builds on the update constraints introduced by GEM and A-GEM, while making the validity of a norm the basis for protection. The contribution is the coupling

of contextually represented normative knowledge with updates conditioned on continued validity. Previously learned norms receive protection while they remain stable, whereas evolving norms remain open to revision in their current context. This connects the stability–plasticity problem to the changing applicability of social knowledge.

Learned norms transfer to unseen scenarios. The improvement on unseen evolving norms shows that learned changes extend to social situations absent from training. Together with stronger historical evolving-norm F1, this indicates that updating current judgments can coexist with retaining earlier judgments in their original time contexts. This extends the contextual adaptation perspective of NormAd and cultural learning approaches by evaluating adaptation and retention across successive stages. The framework couples the application of current norms with the preservation of stable norms. The evaluation also examines whether models recall historical judgments in their original contexts.

These findings motivate studying how selective protection can operate when norm states must be inferred from incoming observations. A central next question is whether uncertainty about normative change can guide protection without preventing valid updates. Longer learning sequences and tests of conflicting norms within communities would further reveal how this balance changes as social evidence accumulates.

7 Conclusion

We studied continual learning of normative judgments conditioned on country and temporal context. The central principle is to protect previously learned norms while they remain valid and allow their current application to change as norms evolve. Our constrained learning framework improves adaptation and stable-norm retention across two models, with additional gains on unseen scenarios. Its theoretical contribution is to formalize norm validity within the learning objective, addressing the distinction between knowledge that requires revision and knowledge that warrants protection. This provides a basis for updating socially aware language models as expectations change. As migration and digital platforms bring cultures into closer contact, our work advances a central capability for socially responsive AI by adapting to changing expectations while preserving still-valid norms within their cultural contexts. Future work should test this approach in real interactions, using feedback from affected communities to assess whether improved norm judgments translate into appropriate model behavior.

8 Limitations

Limited cultural coverage and data realism. Our dataset covers 18 country contexts and 17 norm topics, representing only part of the cultural diversity encountered by language models used worldwide. Although grounded in surveys and partially verified by humans, the scenarios are generated with AI assistance and may not fully reflect how people interpret norms in everyday situations. Evaluation covers three stages; adaptation and retention over longer sequences of normative change remain untested.

Dependence on predefined norm states and local constraints. The method relies on predefined survey thresholds to distinguish stable from evolving norms and a fixed schedule for introducing novel norms, rather than inferring norm states during learning. Incorrect assignments could protect outdated judgments or leave still applicable norms unprotected. Moreover, the update constraint controls only a local linear approximation to stable norm loss, without guaranteeing preservation of the nonlinear loss over an entire learning stage.

Interactions among models remain untested. Our evaluation focuses on social norm judgments by individual models; we have not tested the framework in settings involving interactions among multiple models.

Reproducibility Statement

To support reproducibility, Section 3.2 describes the data construction and labeling procedure, and Section 5 specifies the evaluated models, data splits, random seeds, baselines, and evaluation metrics.

Appendix B provides the detailed derivation of the update rule. Upon acceptance, we will publicly release the processed dataset on Hugging Face.

Ethics Statement

Judgments derived from surveys describe attitudes in specific cultural and temporal contexts, not universal moral standards. They may underrepresent minority perspectives and should not be used to justify discrimination or impose a community's majority views on individuals. Our framework supports contextual norm modeling rather than prescriptive moral decision making.

AI Use Statement

We used GPT-5.4 to assist in generating synthetic social scenarios and AI tools to polish the language of an author-written initial draft. The data construction and manual verification procedure is described in Section 3.2.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12404–12422, 2024. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

William J. Brady and M. J. Crockett. Norm psychology in the digital age: How social media shapes the cultural evolution of normativity. *Perspectives on Psychological Science*, 19(1): 62–64, 2024. doi: 10.1177/17456916231187395. URL https://doi.org/10.1177/17456916231187395.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

E Yagmur Erten, Pieter van den Berg, and Franz J Weissing. Acculturation orientations affect the evolution of a multicultural society. *Nature communications*, 9(1):58, 2018.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29c8c615b3187ee995029284702d3f43-Paper-Conference.pdf.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15217–15230, 2023. doi: 10.18653/v1/2023.emnlp-main.941. URL https://aclanthology.org/2023.emnlp-main.941/.

Michele J Gelfand, Sergey Gavrilets, and Nathan Nunn. Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual review of psychology*, 75(1): 341–378, 2024.

Roberto González, Héctor Carvacho, and Nicole Tausch. Psychological processes underlying normative transformation and social change. *Nature Reviews Psychology*, 4(6):404–416, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

ISSP Research Group. International social survey programme: Family and changing gender roles I–V cumulation. Data set, 2026. URL https://doi.org/10.4232/1.14750. ZA8795, Version 1.0.0.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vfsRB5MImo9.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65183–65216, 2024. doi: 10.52202/079017-2082. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/77f089cd16dbc36ddd1caeb18446fbdd-Paper-Conference.pdf.

Yan-Shuo Liang, Jia-Rui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2310. URL https://papers.neurips.cc/paper_files/paper/2025/hash/63692d8d567db671c700df5df912204a-Abstract-Conference.html.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3114–3134, 2025. doi: 10.18653/v1/2025.acl-long.156. URL https://aclanthology.org/2025.acl-long.156/.

Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass editing memory in a transformer. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://memit.baulab.info/.

NORC at the University of Chicago. GSS 1972–2024 cross-sectional cumulative data. Data set, 2026. URL https://gss.norc.org/get-the-data.html. Release 3a, July 2026.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. URL https://www.science.org/doi/abs/10.1126/science.adh2586.

Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836, 2024. doi: 10.1162/tacl_a_00675. URL https://aclanthology.org/2024.tacl-1.46/.

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

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In *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. 2373–2403, 2025. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53764–53797, 2024a. doi: 10.52202/079017-1703. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/60960ad78868fce5c165295fbd895060-Paper-Conference.pdf.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 663–677, 2024b. doi: 10.18653/v1/2024.naacl-long.37. URL https://aclanthology.org/2024.naacl-long.37/.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8696–8708, 2021. doi: 10.18653/v1/2021.emnlp-main.685. URL https://aclanthology.org/2021.emnlp-main.685/.

Appendix

A Model Evaluation

We evaluate 46 models from ten providers on the three temporal test sets, using the model identifiers recorded in the evaluation results. Table 3 reports their exact macro F1 scores, and Figures 5 through 7 visualize performance across models and temporal contexts. Accuracy and macro F1 across three classes measure performance on the complete test sets from D1 through D3. E2 and E3 denote the subsets of evolving norms of D2 and D3, and HF1 averages macro F1 on historical evolving norms in their original D1 and D2 contexts. The table orders models by D3 macro F1 to show how current performance relates to judgments of evolving norms and historical judgments. Section 5 evaluates the complementary training objective of continual adaptation while preserving stable norms.

Temporal norm reasoning remains a substantial capability gap. Table 3 reveals a broad challenge across the evaluated models, with D3 macro F1 averaging 54.67% and ranging from 22.54% to 71.30%. DeepSeek V3.2 achieves the highest D3 score, followed by GPT 5.6 Terra at 69.90% and GPT 5.6 Sol at 69.76%. Even the strongest evaluated model leaves considerable room for improvement on judgments grounded in everyday social expectations. This gap makes temporal social norm reasoning a concrete target for model development, with consequences for assistance in culturally situated decisions. Figure 5 places these results across all three time contexts, showing the breadth of the challenge beyond a single model family.

Current and historical judgments expose distinct demands. Historical judgments are consistently more difficult in this evaluation, as 45 of 46 models have lower HF1 than E3 macro F1. Mean HF1 is 41.44%, compared with 53.75% on E3, a difference of 12.31 percentage points. DeepSeek V3.2 reaches 72.57% on E3 but 53.45% on HF1, while GPT 5.6 Sol scores 70.93% and 54.97%, respectively. The model ranked first on D3 therefore differs from the model ranked first on HF1, reinforcing the need to evaluate both abilities. These results identify temporal context as an essential part of normative competence because a model must distinguish what applies now from what applied in an earlier setting. Preserving historically situated knowledge and updating current judgments are complementary requirements for socially responsive models.

Aggregate performance conceals weaknesses on changing norms. Across the 46 models, mean macro F1 on the complete test sets rises from 50.37% on D2 to 54.67% on D3, whereas mean F1 on evolving norms falls from 54.28% to 53.75%. At the model level, 42 have higher D3 than D2 macro F1, yet 20 of these simultaneously have lower E3 than E2 macro F1. In total, 23 of the 46 models score lower on E3 than E2. DeepSeek V4 Pro illustrates this divergence, as F1 on the complete test set rises by 2.04 percentage points while F1 on evolving norms falls by 8.08 points. Claude Sonnet 4.5 shows the same pattern, with F1 increasing by 0.50 percentage points on the complete test set and decreasing by 6.00 points on evolving norms. These contrasts across datasets, visualized in Figure 7, show why aggregate evaluation must be complemented by explicit measurements of changing norms.

Implications for continual normative learning. Together, the table and figures establish the need for accurate judgments across current and historical contexts. Explicit evaluation of normative change reveals where these abilities diverge. Their combination makes temporal social norm adaptation a distinct learning problem that requires deciding which knowledge to revise and which knowledge remains valid. Our framework makes this decision using survey comparisons within each country context. Previously learned stable norms constrain updates, while evolving norms and norms first encountered in the learning stream guide adaptation. The evaluation thus gives the research problem a concrete empirical basis and identifies the capabilities that continual normative learning must develop.

Table 3. Temporal social-norm evaluation of all 46 models, ordered by D3 Macro-F1. All scores are percentages. D1–D3 report full-test Macro-F1; E2 and E3 report Macro-F1 on evolving-norm subsets. HF1 averages the historical evolving-norm Macro-F1 scores under the original D1 and D2 time contexts. Bold marks the highest score in each column. The final row is the unweighted mean across models.

Model	D1 F1	D2 F1	D3 F1	E2 F1	E3 F1	HF1
deepseek-v3.2	57.36	64.86	**71.30**	65.87	**72.57**	53.45
gpt-5.6-terra	54.35	66.12	69.90	73.18	70.56	54.01
gpt-5.6-sol	**57.85**	66.72	69.76	70.65	70.93	**54.97**
deepseek-v4-pro	56.76	**67.42**	69.46	**76.21**	68.13	54.93
kimi-k2.6	51.32	61.76	68.68	68.15	71.56	46.93
kimi-k2.5	53.80	60.05	67.49	63.71	68.93	49.09
kimi-k2.7-code	50.46	62.71	67.01	68.39	68.10	48.19
claude-sonnet-4-6	55.03	63.30	66.93	69.96	70.28	53.01
grok-4.5	44.31	56.90	65.72	56.69	64.23	45.31
claude-sonnet-4-5	49.44	65.20	65.70	72.68	66.68	46.74
grok-4	51.30	62.82	65.67	69.65	64.33	52.86
qwen3.5-plus	57.64	60.30	64.45	65.44	62.75	50.94
claude-opus-4-7	52.92	56.72	63.77	62.63	61.07	49.40
qwen3.5-flash	55.14	60.56	63.46	65.92	61.84	52.39
mimo-v2.5-pro	47.32	55.60	62.93	63.58	61.57	46.66
deepseek-v4-flash	51.17	56.10	62.57	64.07	60.42	46.19
claude-opus-4-6	48.97	55.06	62.10	59.73	61.30	44.30
qwen3.6-max-preview	48.23	51.90	61.36	56.21	59.60	44.33
qwen3.6-flash	43.75	52.14	60.72	57.90	58.00	44.67
qwen3.6-plus	54.86	58.15	60.51	63.97	57.43	49.83
qwen3.7-plus	51.77	56.63	60.26	61.41	58.76	48.79
claude-opus-4-5	46.81	57.02	59.72	60.97	58.10	46.56
claude-haiku-4-5	44.88	55.69	59.36	58.25	62.70	46.82
qwen3.7-max	45.19	54.27	58.66	60.33	57.18	44.36
mimo-v2.5	47.15	56.46	57.79	64.63	57.27	44.92
grok-4.3	51.27	52.35	57.39	57.26	58.18	45.99
grok-4-20-non-reasoning	50.72	51.84	57.01	52.14	60.12	42.19
glm-5.1	50.07	51.41	56.14	57.10	55.33	43.90
gemini-3.1-flash-lite	47.83	50.38	55.95	51.65	52.72	44.56
gemini-3.1-pro-preview	52.30	52.30	55.51	59.09	54.13	47.91
gemini-3-pro-preview	52.40	49.95	54.87	56.91	53.82	49.97
MiniMax-M3	40.04	46.72	53.62	53.95	54.27	36.14
claude-opus-4-8	43.60	47.81	52.87	51.92	50.21	42.59
glm-5	38.07	43.65	52.34	51.06	52.25	33.02
glm-5.2	47.81	54.70	51.33	58.73	49.47	41.39
gemini-3.5-flash	41.90	42.90	49.21	47.71	44.78	36.57
gemini-3-flash-preview	40.98	48.89	48.01	52.27	45.55	37.97
MiniMax-M2.7	29.45	33.73	38.58	37.55	39.28	28.62
gpt-5.6-luna	20.63	24.86	31.75	23.26	29.57	13.48
MiniMax-M2.5	26.77	24.93	29.23	26.15	29.60	23.53
gpt-5.5	22.34	30.89	28.98	27.04	27.07	20.97
gpt-5.4	20.85	27.00	28.81	23.12	27.63	16.62
MiniMax-M2.1	24.64	24.52	28.69	22.44	27.16	23.09
gpt-5.4-mini	21.10	23.84	28.35	22.82	23.62	15.91
MiniMax-M2	23.80	24.58	28.29	23.68	24.93	21.73
gpt-5.4-nano	22.97	25.29	22.54	21.03	18.33	20.35
Mean across 46 models	44.51	50.37	54.67	54.28	53.75	41.44

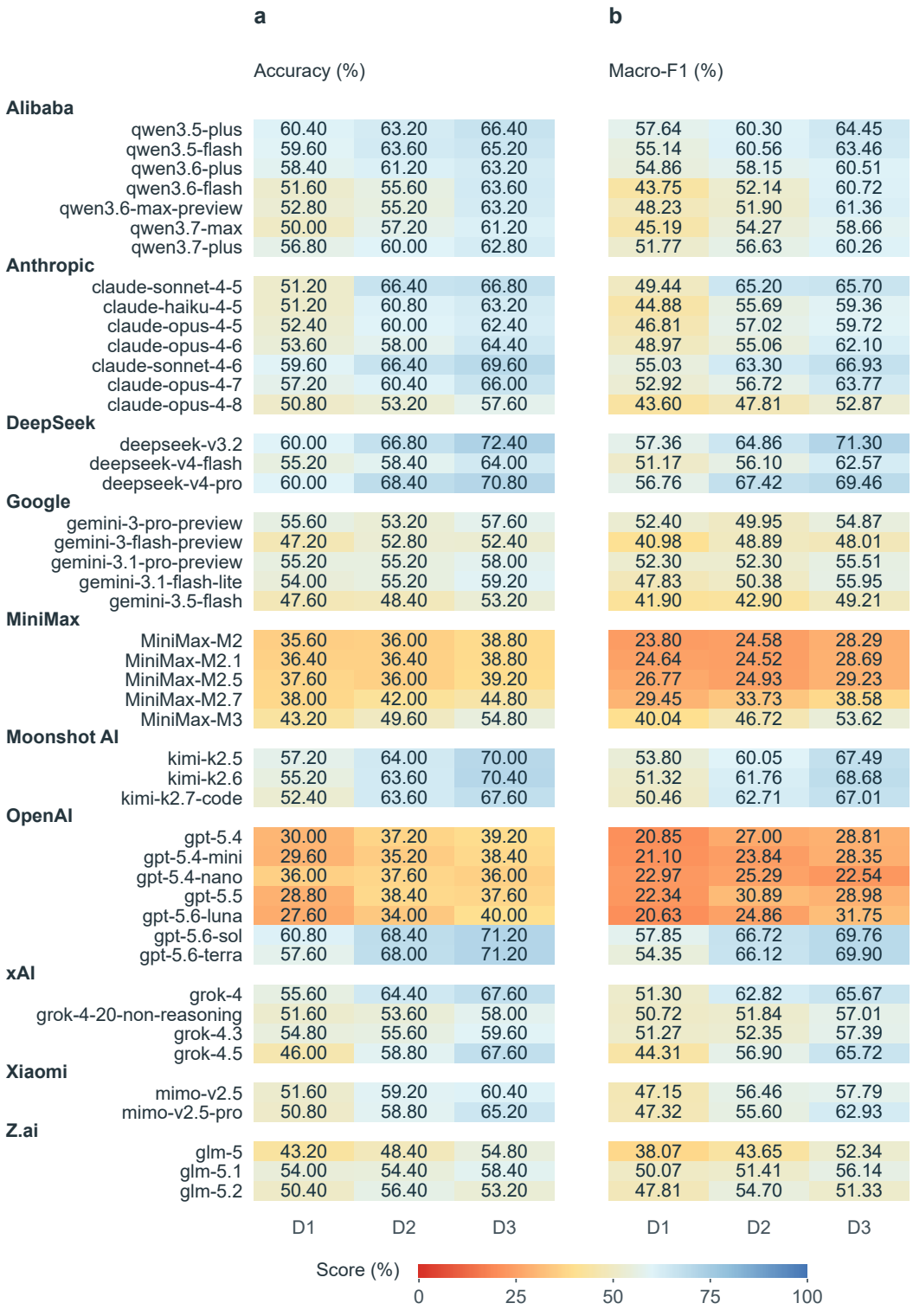


Figure 5. Performance of all 46 models on the complete temporal test sets. **a)** Accuracy on D1 through D3. **b)** Macro F1 on the same test sets. All scores are percentages. Models are grouped by provider, with the same order used across the three appendix figures.

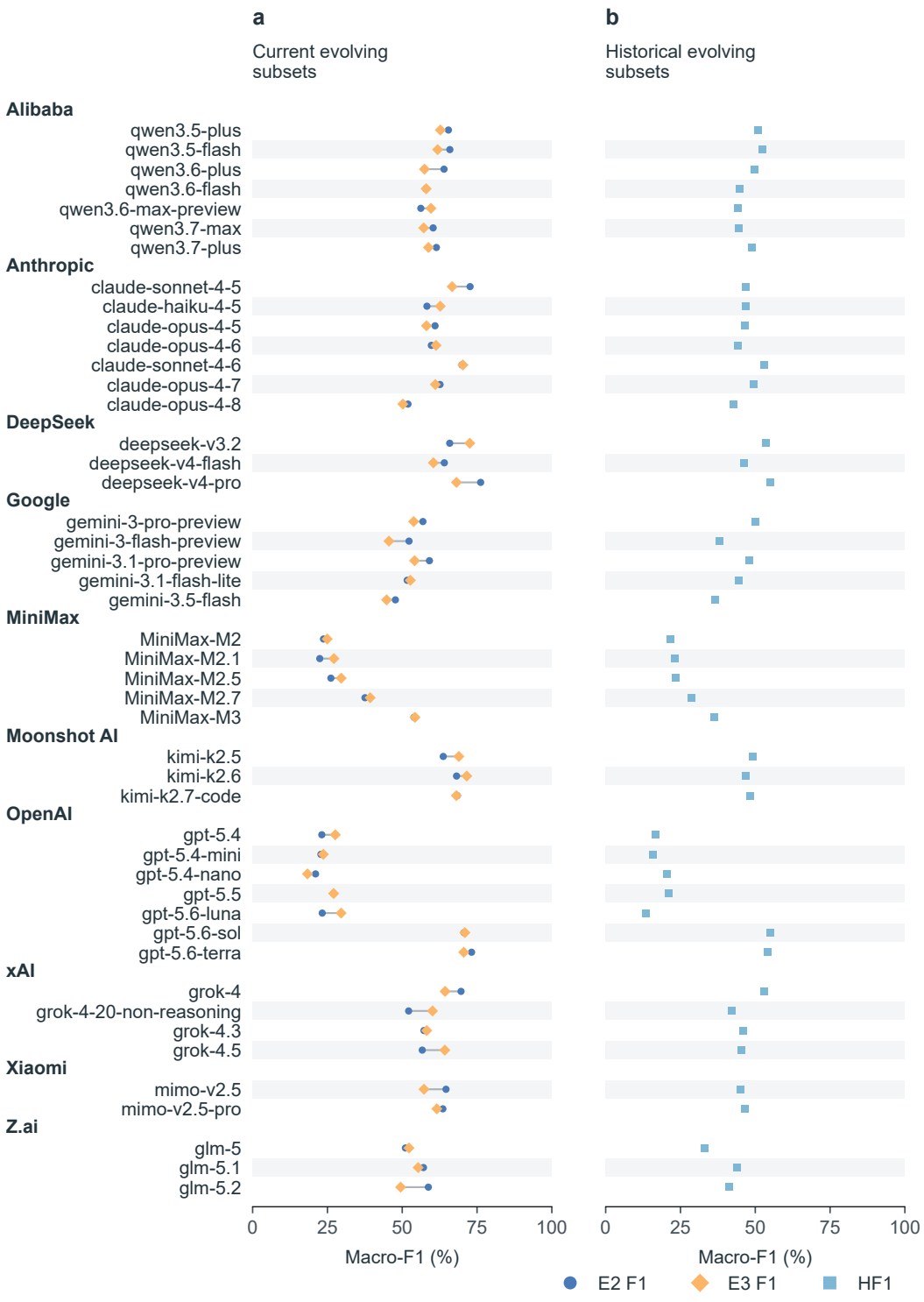


Figure 6. Judgments of evolving norms across temporal contexts. **a)** Macro F1 on E2 and E3, with lines connecting the scores of each model. **b)** HF1 on historical evolving norms under their original time contexts. All scores are percentages.

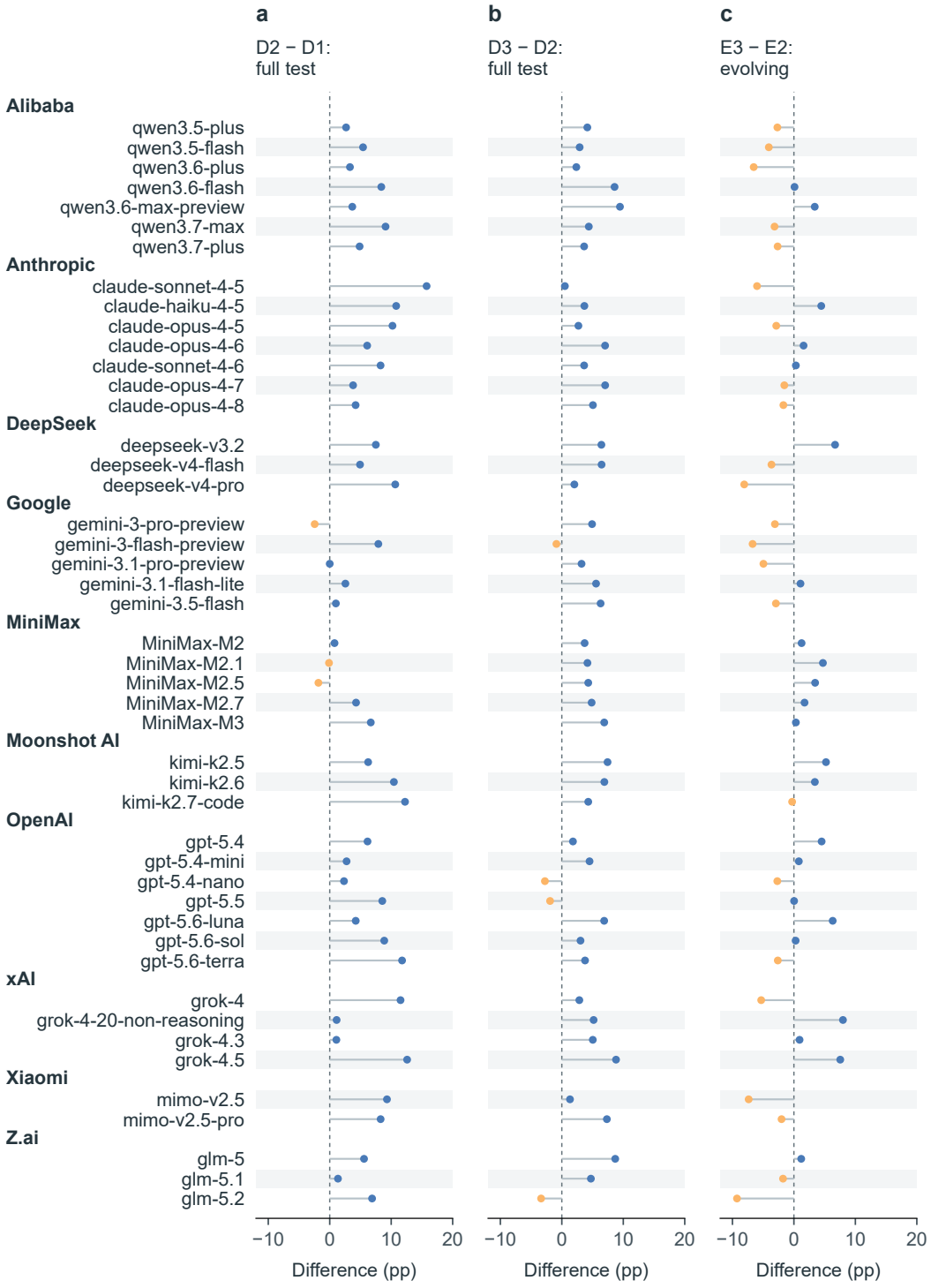


Figure 7. Differences in macro F1 for each model, reported in percentage points. **a)** D2 minus D1 compares the complete test sets at the first two stages. **b)** D3 minus D2 compares the complete test sets at the last two stages. **c)** E3 minus E2 compares the corresponding subsets of evolving norms. Blue denotes values at or above zero, while orange denotes values below zero.

B Detailed Derivation of KKT Update Rule

B.1 From the Loss Constraint to a Local Problem

Let $\bar{\theta}$ denote the model's trainable parameters before an update and $d = \theta - \bar{\theta}$ the displacement. At the start of stage t, $\bar{\theta} = \theta_{t-1}$; subsequent steps use the updated parameters as their expansion point. With data arguments suppressed, define

$$g_n = \nabla_\theta L_{\text{new}}(\bar{\theta}), \qquad g_s = \nabla_\theta L_{\text{stable}}(\bar{\theta}). \tag{9}$$

First-order expansions give

$$L_{\text{new}}(\bar{\theta} + d) \approx L_{\text{new}}(\bar{\theta}) + g_n^\top d, \tag{10}$$

$$L_{\text{stable}}(\bar{\theta} + d) \approx L_{\text{stable}}(\bar{\theta}) + g_s^\top d. \tag{11}$$

Substituting the second expression into the stable-loss constraint in Equation 1 and canceling the reference loss gives $g_s^\top d \le \epsilon$. The constant adaptation loss can also be dropped. A linear objective alone need not have a finite minimizer, so we add a quadratic proximal term to obtain

$$\min_d \quad g_n^\top d + \frac{1}{2\eta}\|d\|^2 \qquad \text{s.t.} \quad g_s^\top d \le \epsilon. \tag{12}$$

For $\eta > 0$, the objective is strictly convex and the feasible set is nonempty because $d = 0$ is feasible for $\epsilon \ge 0$. The local problem therefore has a unique minimizer. Without the constraint, setting its gradient to zero gives $d_0 = -\eta g_n$.

B.2 Lagrangian and Optimality Conditions

For $g_s \neq 0$, the Lagrangian is

$$\mathcal{L}(d, \lambda) = g_n^\top d + \frac{1}{2\eta}\|d\|^2 + \lambda(g_s^\top d - \epsilon). \tag{13}$$

The four KKT conditions comprise 1) stationarity; 2) primal feasibility; 3) dual feasibility; and 4) complementary slackness, as written below.

$$g_n + \frac{1}{\eta} d + \lambda g_s = 0 \qquad \text{(stationarity)}, \tag{14}$$

$$g_s^\top d \le \epsilon \qquad \text{(primal feasibility)}, \tag{15}$$

$$\lambda \ge 0 \qquad \text{(dual feasibility)}, \tag{16}$$

$$\lambda(g_s^\top d - \epsilon) = 0 \qquad \text{(complementary slackness)}. \tag{17}$$

Stationarity implies

$$d = -\eta(g_n + \lambda g_s). \tag{18}$$

Inactive constraint. If $g_s^\top d_0 = -\eta g_s^\top g_n \le \epsilon$, choosing $\lambda = 0$ satisfies all four conditions. The optimal update is

$$d^* = -\eta g_n. \tag{19}$$

Active constraint. If $-\eta g_s^\top g_n > \epsilon$, the ordinary update is infeasible. The multiplier is positive and complementary slackness requires $g_s^\top d^* = \epsilon$. Substitution gives

$$g_s^\top[-\eta(g_n + \lambda g_s)] = \epsilon, \tag{20}$$

$$-\eta g_s^\top g_n - \eta\lambda\|g_s\|^2 = \epsilon, \tag{21}$$

$$\lambda = \frac{-\eta g_s^\top g_n - \epsilon}{\eta\|g_s\|^2} > 0. \tag{22}$$

Hence,

$$d^* = -\eta g_n - \frac{-\eta g_s^\top g_n - \epsilon}{\|g_s\|^2}\, g_s. \tag{23}$$

Combining the two cases gives Equation 4. If $g_s = 0$, the linearized constraint is $0 \le \epsilon$ and $d^* = -\eta g_n$; no division is needed.

B.3 Projection Interpretation

Completing the square gives

$$g_n^\top d + \frac{1}{2\eta}\|d\|^2 = \frac{1}{2\eta}\|d + \eta g_n\|^2 - \frac{\eta}{2}\|g_n\|^2. \tag{24}$$

The final term does not depend on d. Thus, the solution is the Euclidean projection of $-\eta g_n$ onto $g_s^\top d \le \epsilon$. It leaves feasible updates unchanged and makes the smallest Euclidean correction needed when an update violates the constraint. Because g_s is computed only from stable examples, the projection protects previously learned norms whose validity persists. Current examples of evolving norms contribute through g_n, allowing their judgments to be revised for the current context.

B.4 Local Stability

The solution satisfies $g_s^\top d^* \le \epsilon$. Therefore,

$$L_{\text{stable}}(\bar{\theta} + d^*) - L_{\text{stable}}(\bar{\theta}) \approx g_s^\top d^* \le \epsilon. \tag{25}$$

For $\epsilon = 0$, the first-order stable-loss change is nonpositive. The inequality is exact for the linearized constraint, but the nonlinear loss can differ because of higher-order terms. Repeated local updates do not by themselves establish a stage-wide bound relative to θ_{t-1}.

B.5 Preserving an Adaptation Direction

For $\epsilon = 0$, an inactive constraint gives $g_n^\top d^* = -\eta\|g_n\|^2 \le 0$. When the constraint is active,

$$d^* = -\eta g_n + \eta \frac{g_s^\top g_n}{\|g_s\|^2}\, g_s. \tag{26}$$

Taking the inner product with g_n gives

$$g_n^\top d^* = -\eta\|g_n\|^2 + \eta \frac{(g_s^\top g_n)^2}{\|g_s\|^2} \le 0, \tag{27}$$

by the Cauchy–Schwarz inequality $(g_s^\top g_n)^2 \le \|g_s\|^2\|g_n\|^2$. Thus, the correction preserves a nonincreasing first-order adaptation loss. Strict improvement requires a compatible descent direction; the projected update can be zero when the two gradients are exactly opposed. This result is local and does not imply an expected or unconditional decrease in the nonlinear training loss.

C Basic Concepts and Illustrative Examples

This section clarifies the learning problem in Section 3 through two examples traced across the three temporal test sets. The example about marriage draws on the ISSP Family and Changing Gender Roles Cumulation, while the example about suicide draws on the General Social Survey. The survey years are 1994 for T1 and 2002 for T2, followed by 2022 for T3. The task is to judge a scenario within its specified country and time context while learning norms across successive stages. Learning a changed judgment and retaining a judgment that remains applicable serve different purposes.

C.1 Basic Concepts

Contextual social norm. In this study, a social norm is operationalized through attitudes derived from surveys about a normative issue in a specified country and period. The norm identifier v names the issue, while country u and stage t determine the attitude used to judge a scenario. These aggregate attitudes provide a contextual reference for the benchmark without establishing universal moral correctness or implying agreement among all members of a country.

Normative issue and scenario instance. A normative issue, such as acceptance of unmarried cohabitation, can be expressed through many concrete scenarios w. A scenario describes an action or position whose relation to the attitude derived from surveys determines label x. Agreement is labeled

appropriate, whereas disagreement is labeled *inappropriate*. Mixed positions or neutral survey attitudes receive the label *context-dependent* under the construction rules in Section 3.2. Thus, a norm identifier alone does not determine a scenario's label. Explanation y connects the judgment to the scenario and its context. The labels express alignment with the benchmark's reference attitude rather than prescribing personal choices.

Norm state and norm status. The *attitude state* indicates whether an issue receives support or opposition, with neutrality representing the intermediate category. By contrast, *norm status* z describes its role in the learning stream. For an issue recurring within the same country, *stable* means that the attitude state is unchanged and the absolute change in its weighted mean is less than 0.10. A norm is *evolving* if the attitude state changes or if the absolute mean change is at least 0.20 with an approximate 95% change interval excluding zero. Ambiguous transitions are excluded. A norm can therefore evolve without switching from opposition to support. The status *novel* denotes scheduled first exposure in the learning stream rather than the historical emergence of an issue or attitude. Records in the first stage are *initial*. Status depends on the learning stage rather than constituting a permanent property of an issue.

Adaptation and retention across time. Current adaptation involves learning judgments associated with evolving or novel norms at the incoming stage. Stable retention maintains performance on previously learned norms whose validity persists. Historical judgment instead concerns a question posed under its original time context, even when current attitudes differ. Revising a current judgment does not itself constitute catastrophic forgetting. However, updating to current attitudes does not justify returning the latest judgment for every historical question. The constraint on stable memory directly targets stable retention, while historical judgment remains a separate evaluation question beyond its guarantee.

Novel norms and unseen scenarios. A novel norm introduces an issue within a country according to the exposure schedule. An unseen scenario provides a situation withheld from learning through which a norm is evaluated. A norm first introduced at T3 may therefore have training instances in D_3 alongside separate unseen instances evaluated after training. An unseen scenario need not introduce a new country or normative issue.

Information available to the learner. The prediction input is (u, t, w) and the target is judgment x with explanation y. Norm status z supplies construction metadata through which the training procedure assigns examples of evolving or novel norms to D_t^A, while examples of stable norms enter M_t^S. Status is neither an additional prediction input nor a prediction target. The method uses these supplied distinctions without autonomously inferring social change from raw survey responses.

C.2 Marriage Before Parenthood Across Time

This example follows the Australian attitude toward marriage before parenthood. The norm is initial at T1 and stable at T2 before becoming evolving at T3. Each exchange below adapts a test scenario and its reference explanation from the corresponding stage. The scenarios share the position that marriage should precede parenthood but describe different couples rather than repeated observations of one event.

T1 in 1994
User
In Perth, Callum and Renee schedule a small wedding before trying for a baby. They spend their savings on the ceremony rather than an immediate fertility consultation because they want to marry before becoming parents.
Reference response
The judgment is *appropriate* under the Australian reference attitude at T1. The couple makes marriage the first step toward parenthood, which aligns with the expectation that people intending to have children should marry.

T2 in 2002
User
In Perth, Noah and Amelia bring their wedding forward before beginning fertility treatment. They accept an additional treatment expense because they want to be married before becoming parents.
Reference response
The judgment remains *appropriate* under the Australian reference attitude at T2. Marriage continues to be regarded as the expected step before parenthood, so the couple's stated position aligns with the norm that remains stable.

T3 in 2022
User
In Australia, Liam and Zoe schedule a wedding before starting fertility treatment because they want to be married before becoming parents. They postpone the first treatment cycle to pay for the ceremony.
Reference response
The dataset assigns *inappropriate* under the Australian reference attitude at T3. Its explanation interprets the couple's position as making marriage a prerequisite for parenthood, whereas the reference attitude no longer requires prospective parents to marry.

What changes across stages. The reference attitude favors marriage before parenthood at T1 and retains that expectation at T2. At T3, it no longer treats marriage as necessary for people who intend to have children. The contrast concerns alignment with this expectation rather than the moral acceptability of choosing to marry. Because the scenarios also differ in their circumstances, their labels do not constitute a controlled comparison of an identical input across time.

Role in the learning problem. The stable norm at T2 supplies knowledge to preserve, whereas its evolving status at T3 calls for adaptation. This trajectory illustrates why a norm protected at one stage need not remain protected at the next. Questions posed under the original T1 or T2 context still require the corresponding historical judgment rather than an unconditional substitution of the T3 attitude.

C.3 Attitudes Toward Suicide in Response to Weariness with Life

This example follows United States attitudes toward whether being tired of living provides grounds for suicide. The norm is initial at T1 and stable at T2 before becoming evolving at T3. The exchanges adapt test scenarios in which someone responds to another person's distress by seeking support. They illustrate a change in the strength of a survey attitude while its direction remains unchanged.

T1 in 1994
User
In a college dorm, Jamal tells his roommate that he is exhausted by life and has been thinking about dying. His roommate stays with him and contacts the campus crisis service while asking the resident adviser to help keep him safe.
Reference response
The judgment is *appropriate* because the roommate treats Jamal's distress as a reason to provide immediate support. The reference attitude rejects weariness with life as grounds for suicide, with a weighted opposition share of 83.15%.

T2 in 2002
User
During a call in Chicago, Marcus tells Elena that he is exhausted by life and has been thinking about dying. Elena stays on the phone and contacts a crisis service with him while arranging for his brother to remain with him.

Reference response
The judgment remains *appropriate* because Elena responds with support rather than endorsing suicide. The reference attitude remains opposed, with a weighted opposition share of 85.55%. The small change in the survey mean meets the criterion for a stable norm.

T3 in 2022
User
In a Chicago apartment, Lena tells Marcus that she is exhausted by life and does not expect to keep going. Marcus stays with her and calls a crisis service with her while asking her sister to come over.

Reference response
The judgment is still *appropriate* because Marcus responds to distress with immediate support. The reference attitude remains opposed, although the weighted opposition share falls to 71.41%. The decline in opposition is sufficient for the evolving designation without reversing the judgment of this supportive response.

What changes across stages. The weighted survey mean is -0.6630 at T1 and -0.7109 at T2 before rising to -0.4283 at T3. All three values remain below the opposition threshold of -0.20. The absolute change from T1 to T2 is 0.0479, satisfying the stability criterion. From T2 to T3, the increase is 0.2827 with an approximate 95% change interval from 0.1829 to 0.3824. This transition meets the evolving criterion even though opposition remains the attitude state. Percentages are rounded to two decimals, while means and changes are rounded to four decimals.

Role in the learning problem. Unlike the marriage example, this trajectory does not reverse the direction of the reference attitude. It shows why evolving status cannot be equated with a change in every scenario label. The survey describes public attitudes rather than establishing clinical guidance, and the displayed responses remain supportive throughout. Both examples present dataset reference judgments rather than predictions produced by the trained models.

D Dataset Documentation

D.1 Data Fields and Task Inputs/Outputs

Country and time define the context for judging a scenario. The model receives these inputs and produces a judgment with an explanation. The topic identifier links scenarios about the same issue, while norm status guides adaptation or protection during training. Table 4 assigns each field its role, followed by a complete example.

Table 4. Core fields shared by the training and evaluation records. The model receives (u, t, w) and predicts (x, y). Metadata fields are excluded from model inputs and prediction targets.

Field	Role	Meaning
`id`	Record metadata	Unique identifier for tracing a sample.
`time`	Input t	Temporal stage specifying the period in which the scenario is judged.
`culture`	Input u	Country context in which the reference attitude applies.
`norm_id`	Norm metadata v	Topic identifier linking scenarios about the same issue across countries and stages.
`scenario`	Input w	Description of the situation whose normative judgment is requested.
`norm_status`	Training metadata z	Status assigning examples to adaptation or stable protection during training.
`judgment`	Target x	Reference judgment of the scenario under the specified country and time.
`explanation`	Target y	Reason connecting the judgment to the scenario and its normative context.

Complete example from the T3 test set

Record metadata

`id` NG2_D3_TEST_000002
`norm_id` marriage_before_parenthood
`norm_status` evolving

Model input

`time` T3
`culture` Australia
`scenario`
In Australia at T3, Liam and Zoe schedule a wedding before starting fertility treatment because both say they want to be married before becoming parents. They postpone the first treatment cycle to pay for the ceremony.

Reference output

`judgment` inappropriate
`explanation`
Liam and Zoe make marriage a condition of beginning parenthood, conflicting with the Australian T3 view that people who want children do not necessarily need to marry.

D.2 Data Coverage and Composition

Overall coverage. Figure 8 summarizes 6,150 scenarios across 18 countries and 17 topics. Each main stage contains 2,000 scenarios. T1 covers 18 countries and 11 topics. T2 retains all 18 countries and expands to 15 topics. T3 covers 17 countries and 15 topics. Together, these stages represent 159 distinct combinations of country and topic, or 271 contexts when stage is included. The additional 150 unseen scenarios cover 44 combinations across 17 countries and 15 topics at T3.

Composition by data split. Figure 9 separates training coverage from validation and test coverage. Each stage provides 1,500 training scenarios. The validation and test sets each contain 250 scenarios. Training and validation cover every country and topic represented in their stage. The T2 test set covers 17 countries and 13 topics. The other test sets retain the full country and topic coverage of their stages. The unseen set provides a separate evaluation on 150 additional scenarios. These fixed split sizes support comparisons across stages under equal sample budgets.

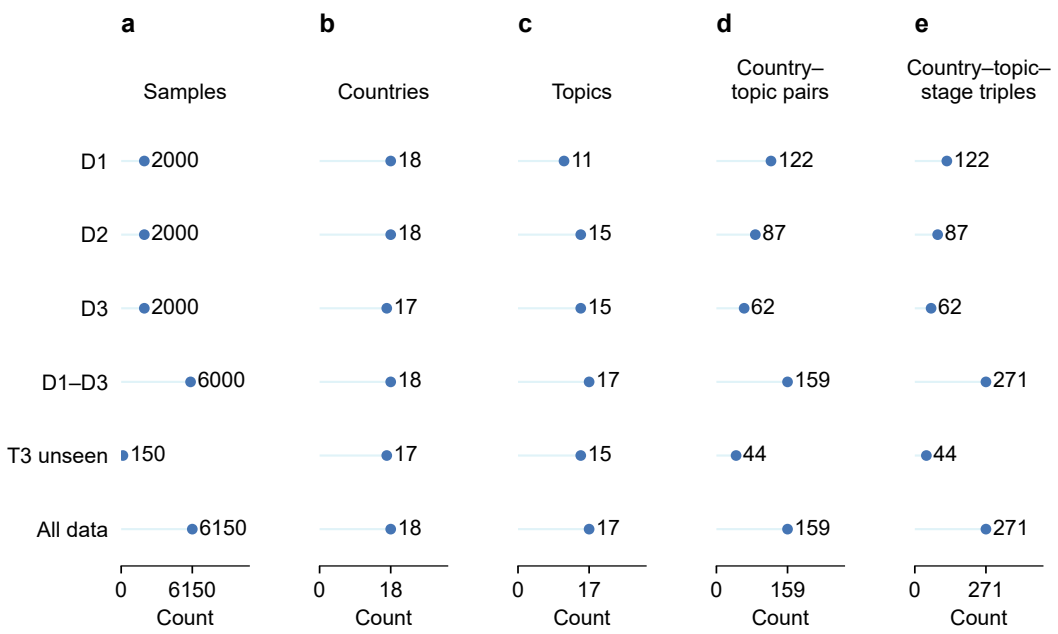


Figure 8. Overall dataset coverage. **a)** Number of scenarios. **b)** Number of countries. **c)** Number of topics. **d)** Number of distinct combinations of country and topic. **e)** Number of distinct combinations when stage is also included. The combined stage counts exclude unseen scenarios, whereas the final row includes them.

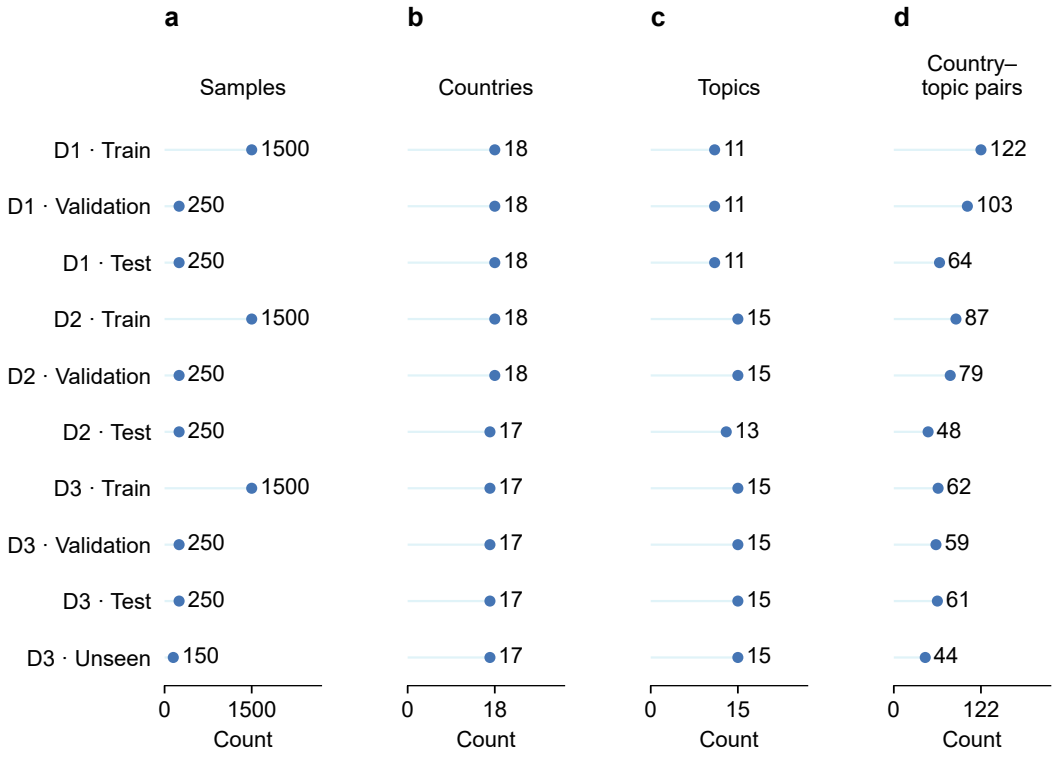


Figure 9. Coverage within each data split. **a)** Number of scenarios. **b)** Number of countries. **c)** Number of topics. **d)** Number of distinct combinations of country and topic. The unseen set is reported separately from the T3 test set.

Topic coverage. Figure 10 shows nine topics represented at every stage. Unmarried cohabitation enters the stream at T2. Parenting by couples of the same sex enters at T3. The largest topic concerns the bond between working mothers and their children, with 809 scenarios including unseen examples. This combination of recurring and newly introduced topics supports evaluation of retention alongside adaptation.

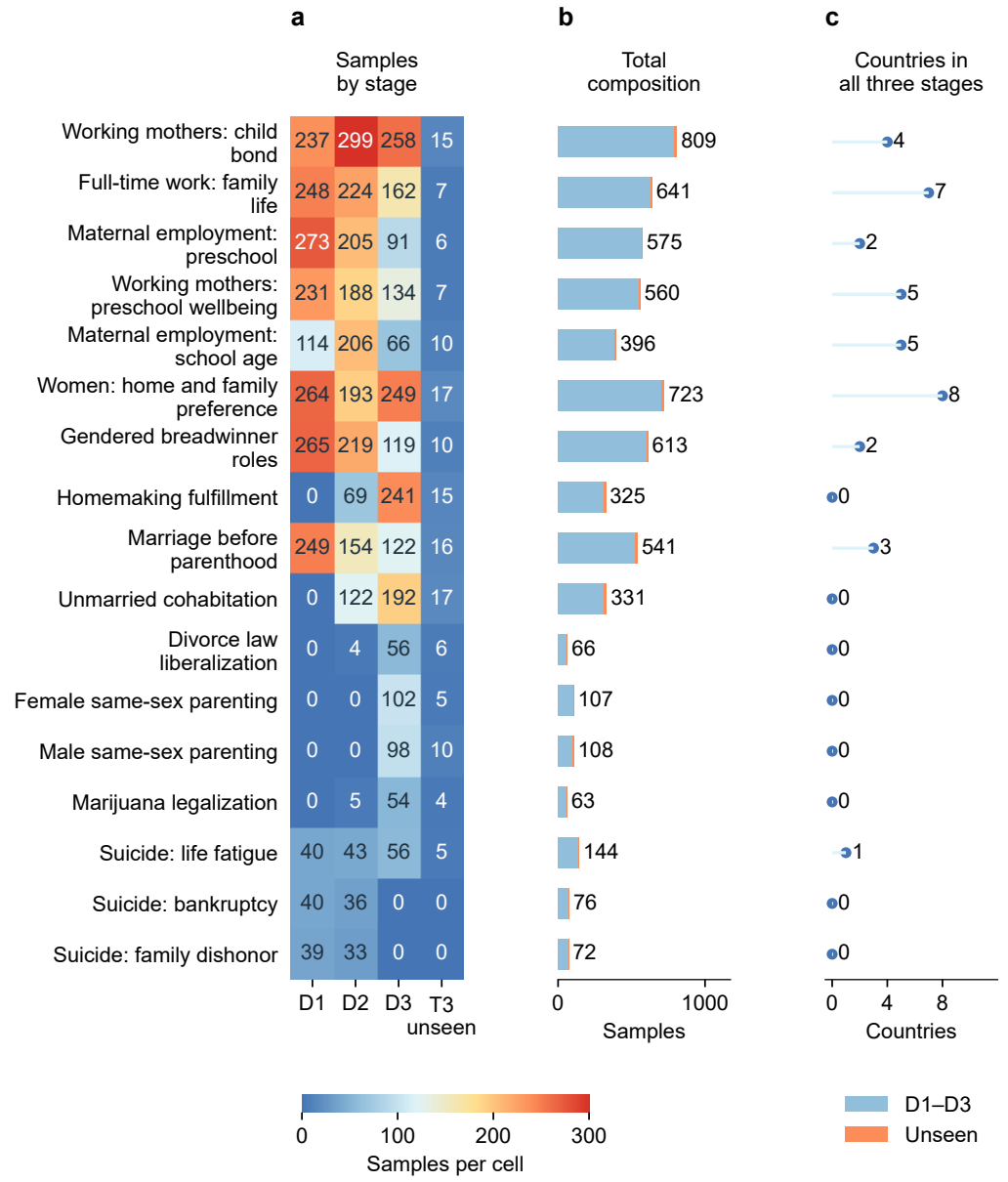


Figure 10. Topic coverage and sample composition. **a)** Scenario counts for each topic within each stage, with unseen scenarios shown separately. **b)** Total scenarios per topic, separating the three main stages from unseen instances. **c)** Number of countries in which the same topic is represented in all three stages. Panel c counts combinations represented at every stage.

Country coverage. Figure 11 reports the sample distribution across countries. T1 and T2 cover all 18 countries. T3 and its unseen set cover 17 countries, with Hungary absent from both. Country totals range from 212 scenarios for Germany to 473 for the Netherlands. This geographic breadth supports comparisons of normative judgments across cultural contexts.

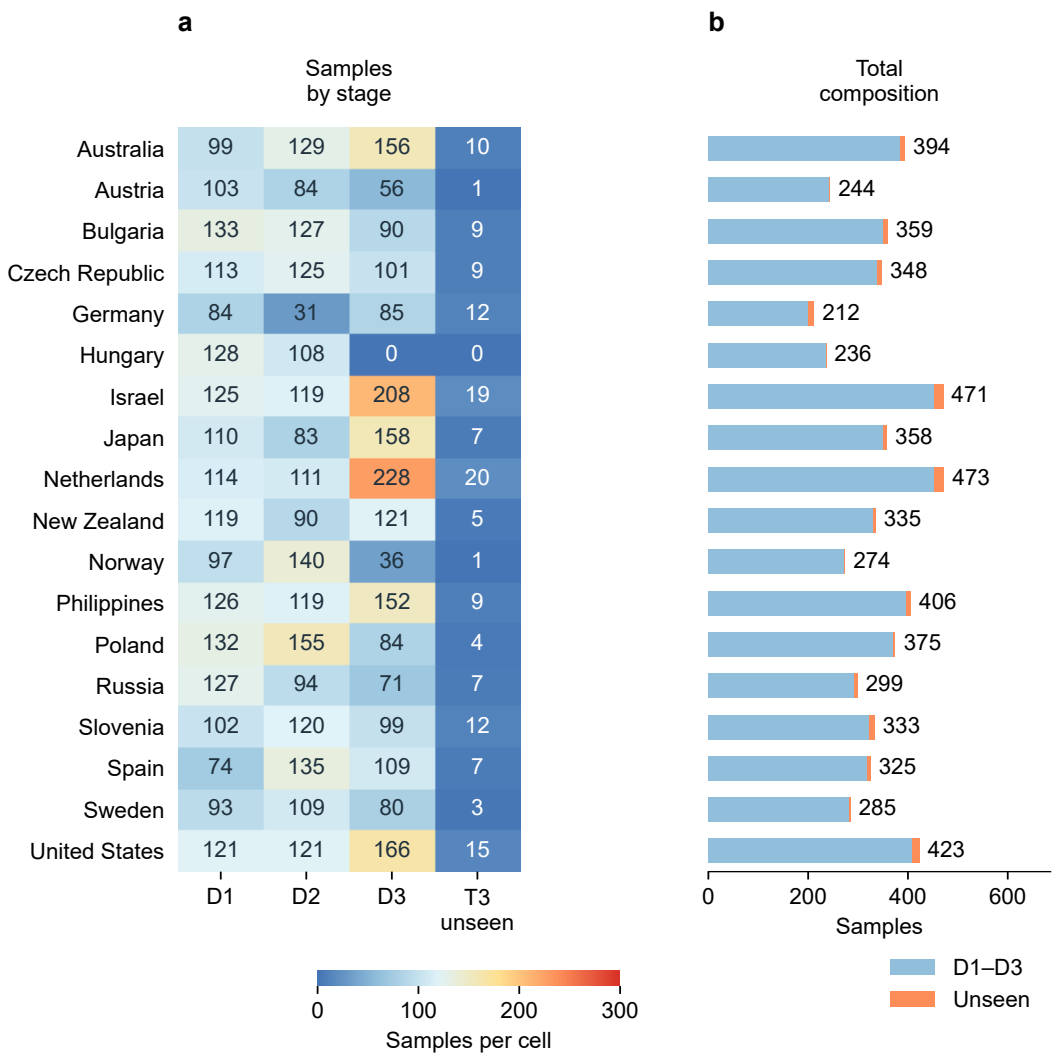


Figure 11. Country coverage and sample composition. **a)** Scenario counts by country within each stage, with unseen scenarios shown separately. **b)** Total scenarios per country, separating the three main stages from unseen instances. Counts refer to scenario samples.

D.3 Distribution of Norm Status

Distribution across stages. Figure 12 reports counts and proportions for each norm status. All 2,000 T1 scenarios have initial status, accounting for 100% of that stage. T2 and T3 each contain 700 stable scenarios, or 35% of their samples. Each stage also contains 1,100 evolving scenarios, accounting for 55%. The remaining 200 scenarios have novel status and contribute 10%. The unseen set contains 45 stable scenarios, accounting for 30%. Its 90 evolving scenarios contribute 60%, while 15 novel scenarios contribute 10%. This composition emphasizes adaptation while retaining a substantial set for assessing stable norm preservation.

Distribution across distinct normative contexts. The same figure also counts distinct combinations of country and topic within each stage. Stable norms cover 34 of 87 contexts at T2, or 39.08%. At T3, they cover 12 of 62 contexts, or 19.35%. The scenario share remains fixed at 35% in both stages. The dataset thus varies contextual coverage while maintaining the same sample allocation for stable norms. Aggregate context counts include stage identity and total 271.

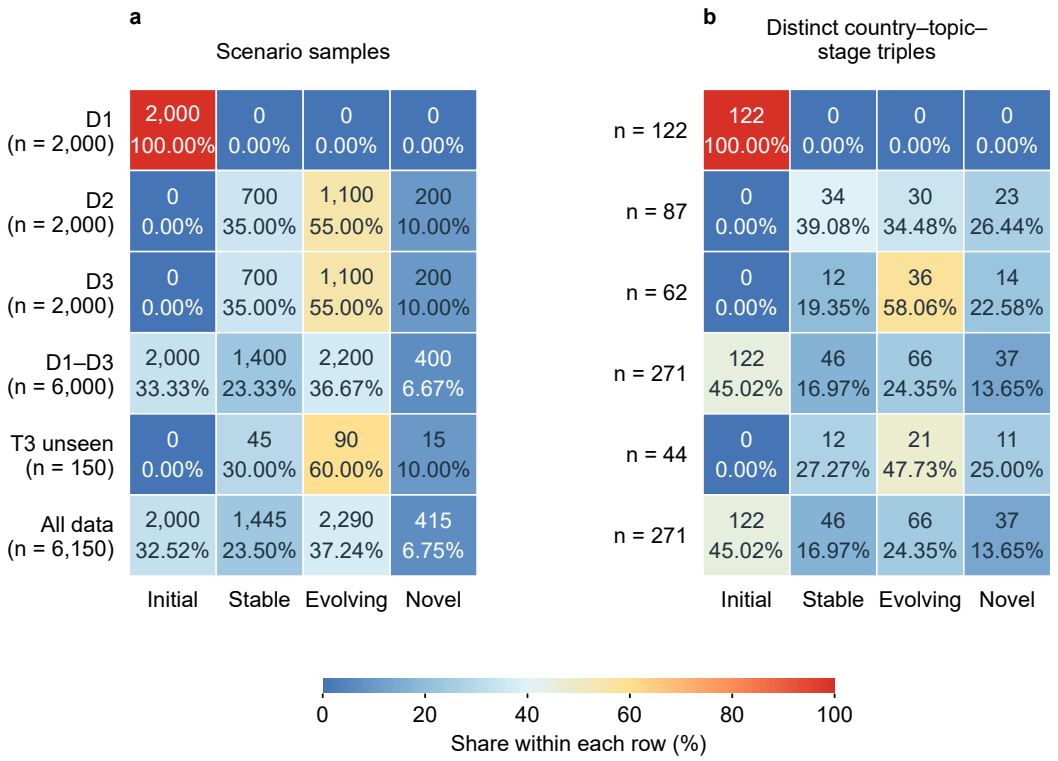


Figure 12. Norm status distribution across stages and in the complete dataset. **a)** Counts and proportions based on individual scenarios. **b)** Counts and proportions after retaining one record for each distinct combination of country and topic within a stage. Each cell reports its count above its percentage, while color indicates the percentage within the row. The combined stage totals exclude unseen scenarios, whereas the final row includes them. Repeated contexts are counted once within each aggregate in panel b.

Distribution within data splits. Figure 13 reports both counting schemes within each split. At T2 and T3, each training set contains 525 stable scenarios and 825 evolving scenarios. A further 150 training scenarios have novel status. Stable scenarios account for 34.80% of validation samples and 35.20% of test samples. The corresponding evolving shares are 55.20% and 54.80%. Novel scenarios account for 10% in each split. This consistent allocation supports comparison of adaptation and retention across training and evaluation.

D.4 Norm Status Transitions Across Stages

Tracking the same normative context. We track each combination of country and topic after pooling the splits within each stage. This analysis covers 159 contexts and excludes the unseen set. The figures use D1 through D3 for the datasets at T1 through T3. The indicator not yet observed identifies contexts awaiting their first appearance in the stream. The indicator no samples in this stage identifies previously observed contexts absent from the current stage.

Transitions from T1 to T2. Figure 14 follows the 122 initial contexts at T1. At T2, 34 become stable and 30 become evolving. The remaining 58 have no samples at T2. Another 23 contexts enter with novel status, while 14 await introduction at T3. For example, unmarried cohabitation in Germany enters at T2 and becomes evolving at T3.

Transitions from T2 to T3. Figure 15 shows 17 contexts moving from stable to evolving, while two move from evolving to stable. Four contexts remain stable and 14 remain evolving. Among contexts introduced at T2, six become stable and five become evolving at T3. Another 14 contexts enter the stream at T3. These include male same sex parenting in Australia. The United States topic

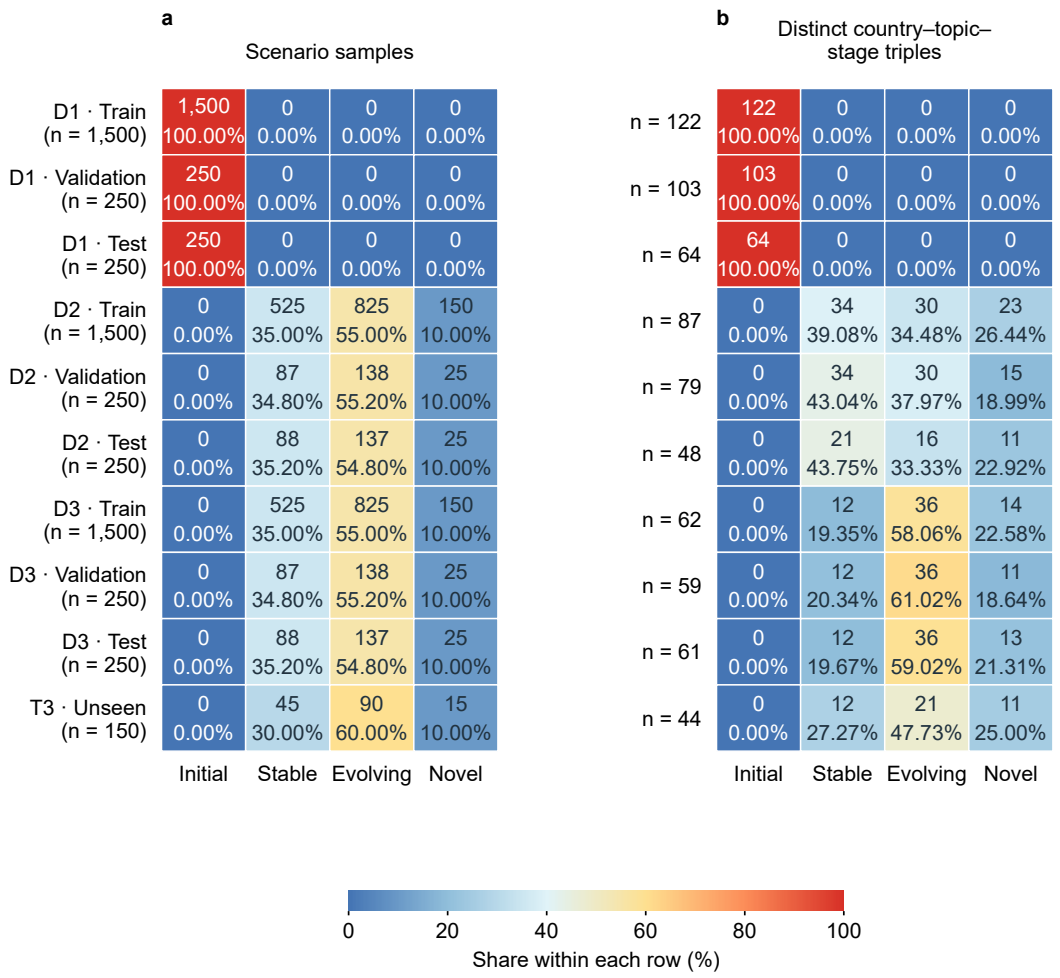


Figure 13. Norm status distribution within each data split. **a)** Counts and proportions based on individual scenarios. **b)** Counts and proportions based on distinct combinations of country and topic within a stage, deduplicated separately for each split. Each cell shows the count and its percentage within the row, with color encoding that percentage. The unseen set is shown separately from the T3 test set. Percentages are rounded to two decimals.

concerning suicide after bankruptcy appears at T1 and T2 but has no samples at T3. These transitions provide direct cases for selecting which learned norms to preserve or update.

Complete trajectories across three stages. Figure 16 groups the contexts into 11 trajectories. 37 contexts have samples at every stage. Within this group, 19 switch between stable and evolving status from T2 to T3. Marriage before parenthood in Australia is initial at T1 and stable at T2 before becoming evolving at T3. This concrete trajectory connects norm preservation to later adaptation. The largest coverage group contains 58 contexts observed only at T1, accounting for 36.48% of all contexts.

D.5 Coverage of Country and Topic Combinations

Coverage shared across stages. Figure 17 summarizes coverage after pooling the splits within each stage and excluding the unseen set. The 18 countries and 17 topics define 306 possible combinations. The dataset covers 159 combinations, or 52.0%, with 147 combinations unrepresented. 37 combinations appear at every stage. Another 27 appear only at T1 and T2, while 11 appear only at T2 and T3. This shared coverage provides recurring contexts for tracking normative change.

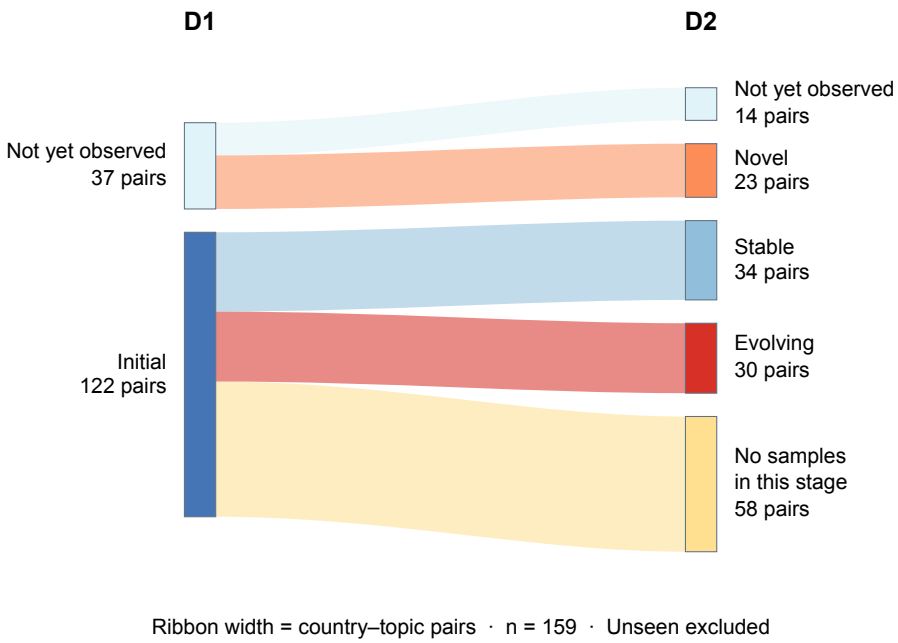


Figure 14. Norm status transitions from T1 to T2. Flow widths represent counts of distinct combinations of country and topic, with 159 contexts tracked in total. Node labels report the number of contexts in each state. Coverage indicators retain contexts that are absent from a stage without assigning them a norm status.

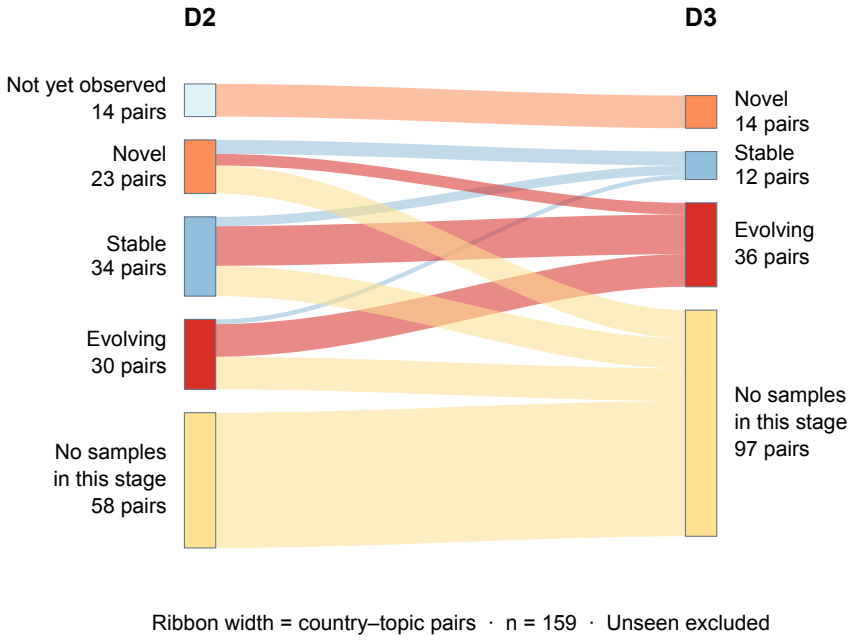


Figure 15. Norm status transitions from T2 to T3. Flow widths count distinct combinations of country and topic using the same union of 159 contexts as in Figure 14. At T3, stable contexts come from both previously observed norms and norms introduced at T2. Coverage indicators identify contexts without samples at a given stage.

D1	D2	D3	Count	Share
Not yet observed	Not yet observed	Novel	14	8.81%
Not yet observed	Novel	Stable	6	3.77%
Not yet observed	Novel	Evolving	5	3.14%
Not yet observed	Novel	No samples in this stage	12	7.55%
Initial	Stable	Stable	4	2.52%
Initial	Stable	Evolving	17	10.69%
Initial	Stable	No samples in this stage	13	8.18%
Initial	Evolving	Stable	2	1.26%
Initial	Evolving	Evolving	14	8.81%
Initial	Evolving	No samples in this stage	14	8.81%
Initial	No samples in this stage	No samples in this stage	58	36.48%

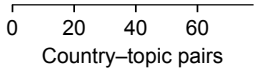
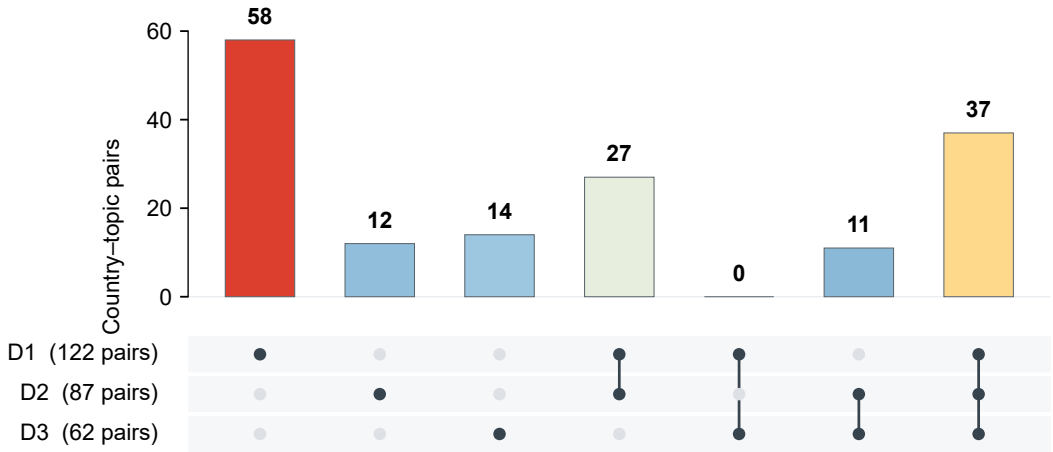


Figure 16. Complete norm status trajectories across the three stages. Each row represents one observed sequence of norm statuses or coverage indicators for a distinct combination of country and topic. Bar lengths show context counts, and percentages use all 159 contexts as the denominator. The unseen set is excluded throughout. Percentages are rounded to two decimals.



Filled dots identify the exact combination of covered stages.

159 / 306 pairs covered (52.0%); 147 pairs absent from all three stages.

Figure 17. Intersections of country and topic coverage across stages. Dark dots identify an exact combination of covered stages. The bar above counts contexts observed only in those stages. Stage labels report the total number of distinct contexts within each stage. D1 through D3 correspond to T1 through T3. The unseen set is excluded, and contexts absent from all stages are reported separately below the plot.

Coverage within countries and topics. Figure 18 maps each topic to its covered countries. Austria and Norway each cover 11 topics, as does Spain. The United States covers six topics. Maternal employment with preschool children appears in 16 countries. Women's full time work and family life also appears in 16 countries. Divorce law liberalization and marijuana legalization appear only in the United States. The three suicide topics are also specific to the United States. Germany has no samples for marriage before parenthood, whereas Australia includes this topic at all three stages. The grid combines broadly shared topics with topics specific to individual countries.



Figure 18. Coverage of individual country and topic combinations. Columns represent countries and rows represent topics. Circle area encodes the number of covered stages among T1 through T3. Color encodes the total scenario count across those stages and the T3 unseen set. Crosses mark combinations with no samples. The grid contains 306 possible combinations, with 159 observed combinations and 147 uncovered combinations.

D.6 Textual Characteristics of Scenarios and Explanations

Distributions of word counts. We count words by splitting each text on whitespace. Figure 19 pools the splits within each stage and reports unseen examples separately. Across the 6,000 main scenarios, the median scenario length is 47 words, with an interquartile range from 42 to 51. The median explanation length is 34 words, with an interquartile range from 30 to 38. Mean scenario length is 46.39 words at T1 and 46.58 at T2. It reaches 47.34 words at T3. Mean explanation length stays between 33.62 and 34.10 words across stages. These narrow ranges show consistent average text lengths throughout the learning stream.

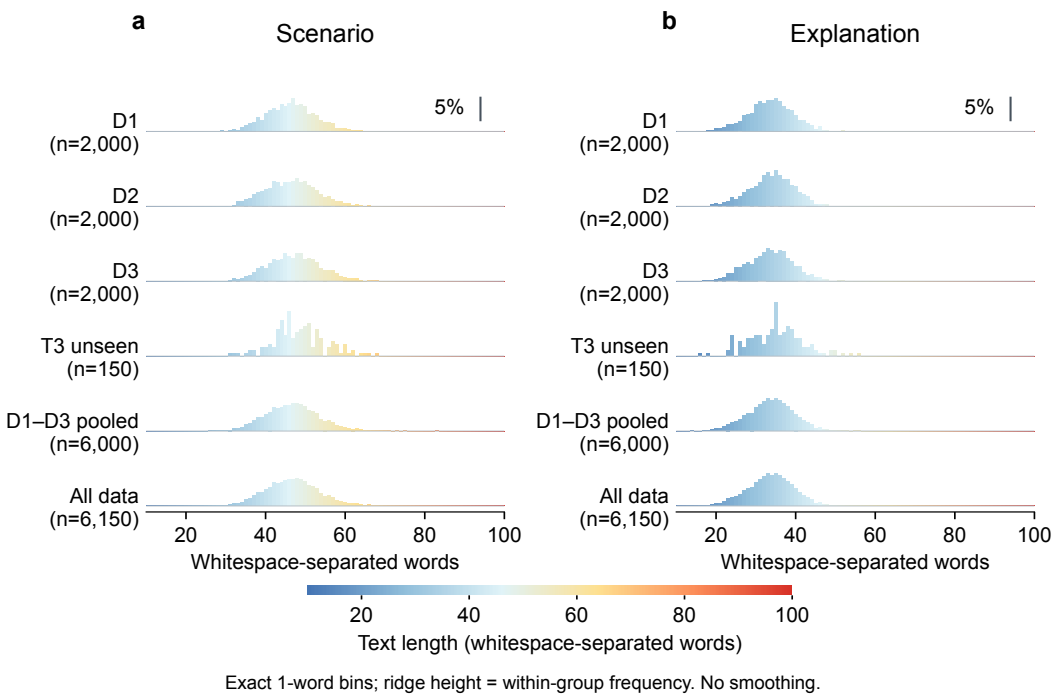


Figure 19. Distributions of text lengths measured by splitting on whitespace. **a)** Scenario lengths. **b)** Reference explanation lengths. Each ridge uses bins of one word and shows relative frequencies within its group without smoothing. Color encodes text length, and the vertical scale marker represents a frequency of 5%. D1 through D3 correspond to T1 through T3. The pooled group excludes unseen examples, while the complete dataset includes them.

Character counts and distribution tails. Figure 20 counts Unicode characters, including spaces and punctuation. Across all 6,150 samples, scenarios have a median of 284 characters. Their central half spans 258 to 311 characters. Explanations have a median of 222 characters, with the central half spanning 197 to 247. The 95th percentiles reach 355 characters for scenarios and 284 for explanations. The unseen medians are 294.50 and 233.50 characters, respectively.

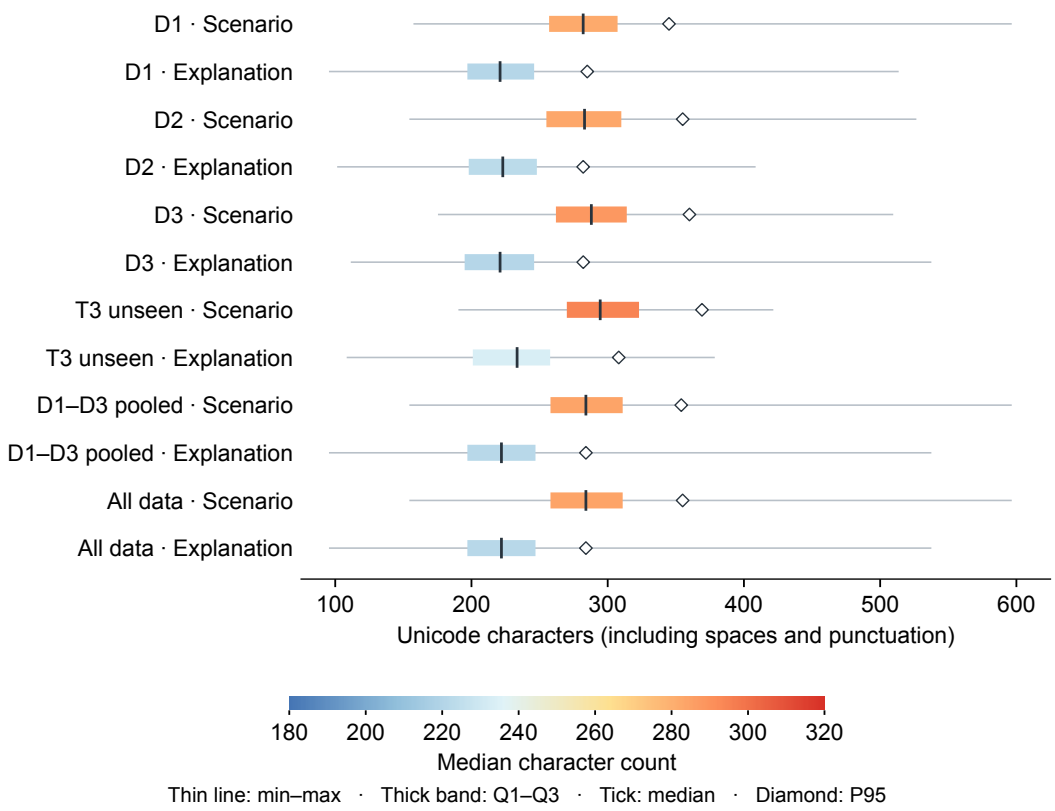


Figure 20. Character lengths of scenarios and reference explanations. Thin lines span the observed minimum and maximum, while thick bands extend from the first to the third quartile. Vertical ticks mark medians and diamonds mark the 95th percentile. Color encodes the median character count. All counts include spaces and punctuation. The pooled group excludes the unseen set, which is included in the complete dataset.

Lengths within data splits. Figure 21 compares mean lengths within each split. Across the nine main splits, mean scenario lengths range from 45.92 to 47.94 words. Mean explanation lengths range from 33.55 to 34.18 words. The unseen set averages 48.54 words per scenario and 35.08 words per explanation. The main splits maintain closely matched average lengths, while unseen examples extend evaluation to somewhat longer texts.

Lengths by judgment label. Figure 22 shows that context dependent examples have the longest scenarios and explanations on average in every stage. The same pattern holds in the unseen set. Across the three main stages, context dependent scenarios average 50.52 words. Appropriate scenarios average 44.04 words, while inappropriate scenarios average 45.75. Context dependent explanations average 36.70 words. The corresponding means are 31.80 for appropriate explanations and 32.90 for inappropriate explanations. These label groups provide distinct text length profiles for analyzing normative judgments.

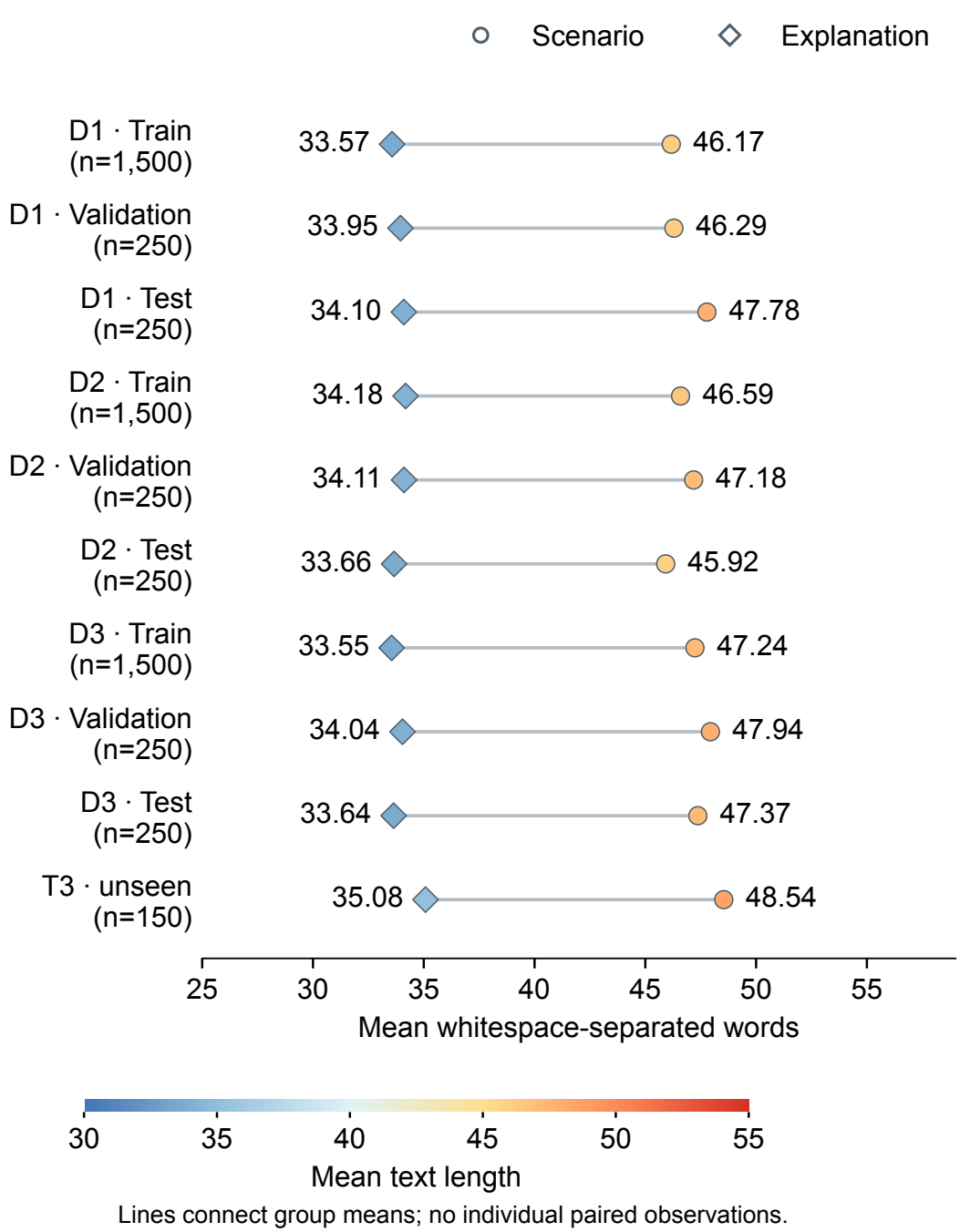


Figure 21. Mean text lengths within each data split, measured by splitting on whitespace. Circles represent scenarios and diamonds represent reference explanations. Each connector joins the scenario mean and explanation mean for the same group. Row labels give the sample size, while marker annotations report means rounded to two decimals. Color also encodes mean length. The unseen set is reported separately from the T3 test set.

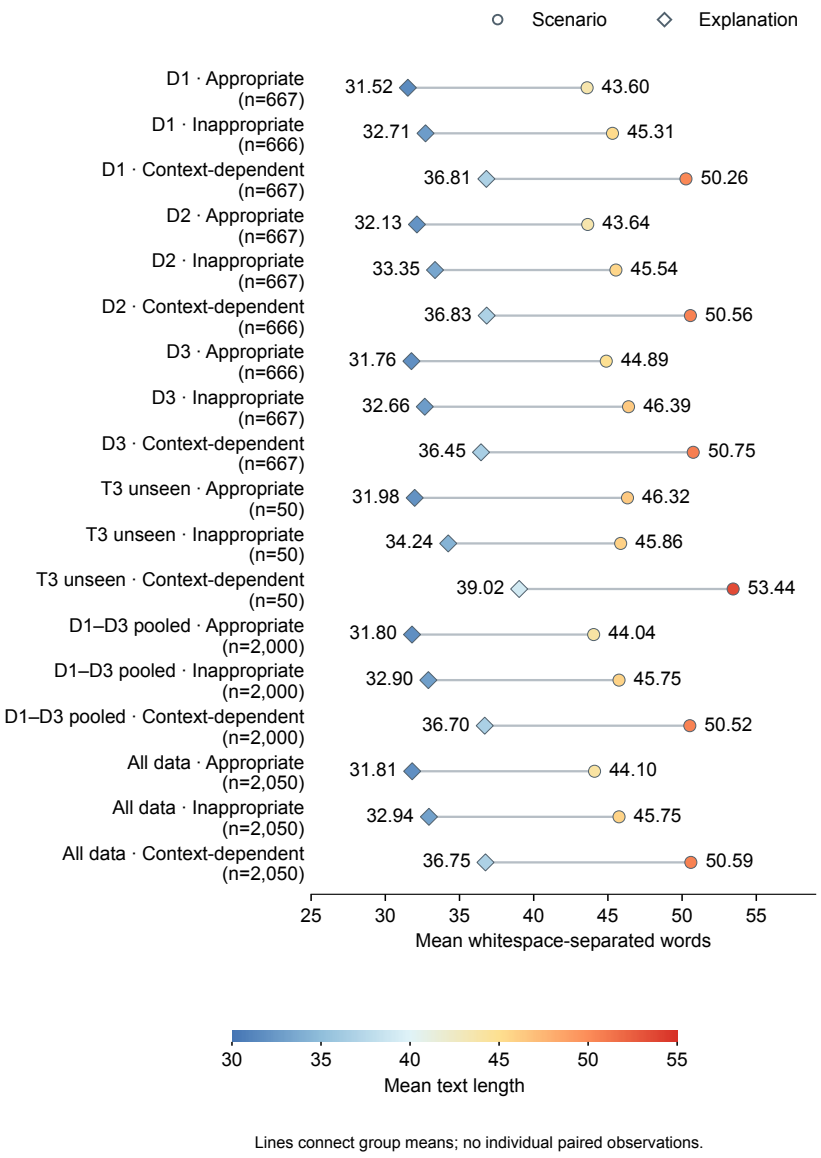


Figure 22. Mean text lengths grouped by judgment label and data scope. Lengths are measured by splitting on whitespace. Circles represent scenarios and diamonds represent reference explanations, with connectors joining group means. Row labels report group sizes, and annotations give mean lengths rounded to two decimals. Color encodes mean length. The pooled group excludes unseen examples, while the complete dataset includes them.