

When Accuracy Aligns and Fails: Diagnosing Human–LLM Response Alignment at Population Scale

Seoyoon Park^{1*}, Hyeji Choi^{2*}, Subin An^{3*},
Xiaonan Wang¹, Heejae Kim¹, Joeun Kang¹, Hansaem Kim^{1†}

¹Yonsei University, ²DATUMO.INC, ³Diquest
{seoyoon.park,nan,kimmejay,j0eun,khss}@yonsei.ac.kr
hyeji.choi@selectstar.ai aaasb9946@gmail.com

Abstract

We propose CAPE, a Choice-Distribution Alignment and Performance Evaluation framework for diagnosing human–LLM response alignment using population-scale human choice distributions. CAPE separates outcome-level difficulty alignment from response-distribution alignment with four metrics: OA, MAE, PA, and DA. We apply CAPE to 609 Korean CSAT items with item-level response distributions from approximately 300K examinees per item, and evaluate 16 LLMs. Results show that within the evaluated model set, model capability is strongly associated with alignment, with accuracy closely tied to both full-choice and distractor-only alignment. However, group-level scores are affected by aggregation size and model composition, and neither reasoning augmentation nor Korean adaptation yields consistent residual advantages. Bloom-level analyses further suggest that capability-linked alignment varies across item types. CAPE thus enables evaluation beyond accuracy by diagnosing how model success and failure patterns relate to human population responses.

1 Introduction

Large language models (LLMs) increasingly achieve high accuracy on benchmark tasks, but high accuracy does not necessarily imply human-like response behavior (Zhou et al., 2024; Hu and Levy, 2023). Models with similar accuracy may still differ in which items they find difficult and which incorrect options they select when they fail (Li et al., 2025). This is especially important for language comprehension, a core target of LLM evaluation where a single accuracy score can obscure how models understand and misunderstand text relative to humans. Existing evaluations largely rely on scalar outcomes such as accuracy and F1, while

reasoning-trace analyses do not show whether models reproduce human-like difficulty patterns or distractor preferences at scale (Huber and Niklaus, 2025). Prior work on human–LLM alignment has also often relied on limited human samples or summary statistics (Santurkar et al., 2023; Meister et al., 2025), leaving choice-level response patterns underexplored (Liu et al., 2025).

To address this gap, we propose CAPE, a Choice-Distribution Alignment and Performance Evaluation framework for diagnosing human–LLM response alignment using population-scale human choice distributions. CAPE separates outcome-level difficulty alignment from response-distribution alignment through four metrics: Outcome Alignment (OA), Mean Absolute Error (MAE), Process Alignment (PA), and Distractor Alignment (DA). In particular, DA removes the correct option and compares distractor-only distributions, distinguishing correctness-driven alignment from human-like distractor preference. We apply CAPE to 609 Korean CSAT items with item-level choice distributions from approximately 300K examinees per item, and evaluate 16 LLMs. Results show that task-specific accuracy is strongly related to both full-choice and distractor-only alignment, while group-level scores are sensitive to aggregation size and model composition. Bloom-level analyses further suggest that capability-linked alignment varies across item types.

This study makes three contributions. First, we propose a domain-transferable methodology for diagnosing human–LLM response alignment from item-level human choice distributions, applicable whenever multiple-choice assessments provide response statistics. Second, we introduce CAPE to separate outcome-level difficulty alignment from response-distribution alignment, including distractor-only alignment. Third, we use Bloom’s taxonomy (Bloom, 1956) as an item-level probe of when capability-linked alignment holds or varies.

*Equal contribution.

†Corresponding author.

2 Related Works

Human–LLM response alignment. Recent work compares LLM outputs with human response *distributions* rather than gold answers alone, spanning opinion and survey simulation (Argyle et al., 2023; Santurkar et al., 2023; Meister et al., 2025), psycholinguistic and grammaticality alignment (Hu et al., 2023, 2024), and confidence or calibration relative to human judgments (Kadavath et al., 2022; Chhikara, 2025; Steyvers et al., 2025). Recent work has also treated fuller human response distributions as a modeling target. For example, Nighojkar et al. (2025) use personality prompting and optimization to improve the match between LLM and human response distributions. A parallel line treats LLMs as cognitive subjects and finds that matching human outcomes need not imply matching human processes (Binz and Schulz, 2023; Dillion et al., 2023; Hu and Levy, 2023; Jones et al., 2024; Binz et al., 2025; Song et al., 2025; Strachan et al., 2024). Closest to us, distractor- and choice-level comparisons on multiple-choice items relate model behavior to the options humans select (Liu et al., 2025; Tam et al., 2025). We differ by using a population-scale human baseline (about 300K examinees per item), by separating full-choice from distractor-only alignment, and by reporting both alongside outcome-level difficulty alignment rather than a single score.

Cognitive structure and capability. A related strand probes *where* capability matches human cognition, often via Bloom’s taxonomy or psychometric structure (Zhang et al., 2023; Huber and Niklaus, 2025; Raimondi and Gabbrielli, 2026), and through the broader proposal to decompose competence rather than report a single score, e.g. formal versus functional linguistic ability (Mahowald et al., 2024; AlKhamissi et al., 2025; Hanna et al., 2026). Yet greater capability need not yield human-like behavior: models can form a *machine consensus* on difficulty that diverges from human perception (Li et al., 2025), become *less* reliable with scale and instruction tuning (Zhou et al., 2024), and fail to recall knowledge they encode (Allen-Zhu and Li, 2023; Berglund et al., 2024). Unlike this prior work, we use Bloom’s taxonomy not to rank cognitive levels but to probe *when* capability-linked alignment holds across both correctness and distractor structure, grounding difficulty in observed examinee behavior; this motivates our analysis of low-cognitive-level items, where human-like dis-

tractor preference may coexist with failed retrieval.

Standardized exams and student simulation. Standardized human exams are widely used as LLM testbeds (Hou et al., 2024; Kang et al., 2024; Tan et al., 2024; Zhong et al., 2024), and a growing body uses LLMs to simulate student responses or predict item difficulty for educational assessment (Bitew et al., 2023; Benedetto et al., 2024; Lu and Wang, 2024; Feng et al., 2025; Wang et al., 2025). Our cognitive-level annotation follows the CSAT’s official objective structure and revised Bloom’s taxonomy (Anderson and Krathwohl, 2001; Moon et al., 2004; Cho, 2007; Chandio et al., 2016).

3 Dataset

3.1 The CSAT Korean Language Section

The College Scholastic Ability Test (CSAT) is Korea’s standardized college entrance examination. Its Korean language section consists of approximately 45 five-choice items each year, covering passage-based reading comprehension and inference across six subdomains: Literature, Reading, Speaking, Writing, Grammar, and Classical Grammar.

Sub-domain	Instances
Literature	248
Reading	219
Speaking	53
Writing	42
Grammar	38
Classical Grammar	9

Table 1: Distribution of CSAT Korean items by subdomain

This section provides a suitable testbed for human–LLM alignment analysis because it spans diverse cognitive processes, shares a common five-choice answer space across items, and includes distractors designed to attract actual examinee errors. These properties allow human and LLM responses to be compared over the same choice distribution and support analysis of human-like distractor preferences. We analyze 609 items from the official CSAT and KICE (Korean Institute for Curriculum and Evaluation)-administered mock CSAT exams between 2020 and 2024, with subdomain counts reported in Table 1. Examples for CSAT are shown in Appendix A.

3.2 Population-Scale Human Response Distribution

The CSAT provides a rare setting for population-scale response alignment: about 500K students take it each year, and its central role in Korean college admissions has led major exam-preparation platforms to release item-level response statistics. These statistics allow us to construct empirical human choice distributions over all five options, rather than compare models only with gold answers or small human samples.

For human response data, we use item-level response statistics (Table 2) from the official CSAT and mock CSAT exams provided by Megastudy¹, one of Korea’s major college entrance exam preparation platforms.

Item	Ans.	Acc.	Choice Dist. (%)				
			1	2	3	4	5
1	3	90	2	2	90	2	4
2	5	89	2	4	2	3	89
3	4	69	3	6	17	69	5

Table 2: Example of Megastudy item-level response statistics used to construct human choice distributions. **Ans.** is the correct answer, **Acc.** the overall accuracy (%), and columns 1–5 the percentage of examinees selecting each choice.

Each item includes a five-choice response distribution from approximately 300K examinees, allowing us to estimate examinees’ uncertainty over answer choices and their preferences for plausible distractors. This enables us to analyze both outcome-level performance and process-level response alignment.

Although these data are closely connected to the target CSAT-taking population, they are based on platform users and voluntary submissions rather than the official full examinee population. We therefore treat them as a large-scale behavioral baseline, rather than a complete census of examinee responses, and discuss possible platform-based self-selection as a limitation.

3.3 Cognitive Level Annotation

We use Bloom’s taxonomy as a cognitive probe for examining when model capability does or does not translate into human-like response distributions. Each item was annotated with one of the six levels

¹<https://www.megastudy.net/Entinfo/correctRate/main.asp>

Bloom’s (N)	CSAT
Remembering (76)	Vocabulary & Concepts
Understanding (396)	Factual Understanding Inferential Understanding
Applying (42)	Inferential Understanding
Analyzing (61)	Inferential Understanding
Evaluating (23)	Critical Understanding
Creating (11)	Application/Creativity

Table 3: Mapping between CSAT Korean understanding layers and Bloom’s taxonomy used for cognitive-level annotation. Numbers in parentheses denote the number of items mapped to each Bloom’s level after annotation. See Appendix B for the full breakdown and Appendix C for mapping criteria.

of the revised Bloom’s taxonomy: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating (Bloom, 1956; Anderson and Krathwohl, 2001).

This annotation is grounded in the hierarchy of behavioral objectives in the Korean national curriculum (Moon et al., 2004; Cho, 2007). Because this hierarchy reflects differences in cognitive processing depth, we map these CSAT Korean understanding layers onto Bloom’s taxonomy, as summarized in Table 3. Two certified in-service Korean language teachers independently annotated the items. Inter-annotator agreement was high (Cohen’s $\kappa = 0.84$), and disagreements were resolved through adjudication. Because Creating contains only 11 items, analyses involving this category are interpreted as exploratory.

4 CAPE Framework

4.1 Models

We evaluated 16 LLMs in this study. The models are grouped by accessibility, reasoning augmentation, open-source model scale, and Korean adaptation. The model selection is designed to cover four comparison axes: proprietary versus open-source models, reasoning versus non-reasoning models, model scale, and Korean fine-tuning. Open-source models are grouped by scale into light ($\leq 3B$), small (4–8B), middle (9–32B), and big (33–72B) categories. Where possible, we include both general-purpose multilingual models and Korean fine-tuned models. This setup is intended not to rank individual models, but to examine whether CAPE can diagnose differences in alignment associated with model characteristics. The evaluated models and

their CSAT Korean accuracy scores are reported in Table 4. Exact model versions and experiment prompts are provided in Appendix D and E.

4.2 Alignment Metrics

This study compares human examinee responses and LLM responses along two dimensions: an outcome dimension and a process dimension. The outcome dimension captures shared item-level difficulty structure and performance gaps, whereas the process dimension captures distributional similarity within the same five-choice answer space. CAPE defines four metrics: Outcome Alignment (OA), Mean Absolute Error (MAE), Process Alignment (PA), and Distractor Alignment (DA).

$$\begin{aligned} \text{OA} &= \rho_s(\{h_i\}_{i=1}^N, \{l_i\}_{i=1}^N), \\ \text{MAE} &= \frac{1}{N} \sum_{i=1}^N |h_i - l_i|, \\ \text{PA} &= \frac{1}{N} \sum_{i=1}^N \left(1 - \text{JSD}(p_h^{(i)}, p_l^{(i)})\right), \\ \text{DA} &= \frac{1}{|\mathcal{I}_{\text{DA}}|} \sum_{i \in \mathcal{I}_{\text{DA}}} \left(1 - \text{JSD}(\hat{p}_h^{(i)}, \hat{p}_l^{(i)})\right). \end{aligned} \quad (1)$$

For each item i , let h_i and l_i denote human and LLM accuracy, and let $p_h^{(i)}$ and $p_l^{(i)}$ denote their response distributions over the five answer choices. The human distribution is obtained from the proportion of approximately 300K examinees selecting each option. For a model group, $p_l^{(i)}$ is constructed by aggregating the selected choices across models within the group. Unless otherwise specified, $p_l^{(i)}$ refers to the distribution of G_{all} , which consists of all 16 models. Subgroup distributions are used for analyses by model characteristics. For an individual model, $p_l^{(i)}$ is represented as a one-hot distribution over the option selected by that model.

For distractor-only analysis, we remove the correct option and renormalize the remaining four choices, yielding $\hat{p}_h^{(i)}$ and $\hat{p}_l^{(i)}$. For an individual model, $\hat{p}_l^{(i)}$ is defined only for incorrectly answered items. When the model selects the correct answer, removing the correct option leaves zero distractor mass, so the item is excluded from that model’s mean DA. For a model group, $\hat{p}_l^{(i)}$ is defined whenever at least one model in the group selects a distractor. Thus, individual-model DA is an error-conditional measure and may be computed over

different failure sets across models. We denote the set of items for which $\hat{p}_l^{(i)}$ is defined by $\mathcal{I}_{\text{DA}}$.

OA captures rank-order similarity in item difficulty using Spearman’s correlation coefficient, whereas MAE measures the absolute human–LLM performance gap. We compute PA and DA as one minus the Jensen–Shannon divergence (JSD) between the corresponding response distributions, using base-2 logarithms. Thus, higher PA and DA values indicate greater distributional similarity. PA measures full-choice response alignment, reflecting both correctness and distractor preference, whereas DA measures distractor-level alignment conditional on the presence of LLM distractor mass. For G_{all} , DA is defined for 598 of the 609 items (98.2%); the remaining 11 items were answered correctly by all models. Together, PA and DA distinguish correctness-driven alignment from similarity in distractor preferences. We compute all metrics at the item, Bloom-level, and model-group levels. Code for computing the CAPE metrics is available at <https://github.com/beammeup1229/CAPE>.

5 Results

5.1 Capability Baseline

In this paper, we use accuracy on the CSAT Korean items as a task-specific measure of model capability. We therefore begin by examining accuracy across the 16 LLMs before turning to CAPE-based alignment analysis.

5.1.1 Accuracy by Model

As shown in Table 4, model accuracy ranges widely from 0.20 to 0.88, providing substantial capability variation across the 16 LLMs. Proprietary and reasoning-augmented models generally achieve higher performance. Among open-source models, larger models tend to perform better overall. However, performance does not improve monotonically with parameter size or Korean fine-tuning. We treat these accuracy differences not as a model ranking, but as capability variation for analyzing human–LLM alignment in Section 5.2.

5.1.2 Accuracy by Bloom Level

Table 5 compares the accuracy of the full AI group (G_{all}) and human examinees across Bloom levels. Overall, the full AI group underperforms humans, but the performance gap does not increase monotonically with cognitive level. Instead, the largest gap appears at the Remembering level: human accuracy is 71.5%, whereas the full AI group achieves

Model Group	Model	Acc.	Model Group	Model	Acc.
Proprietary	Claude 3.5 Sonnet	0.88	Small	Qwen 2.5 7B	0.46
	GPT-4o	0.80		Llama 3.1 8B	0.27
	Gemini 2.0 Flash	0.78		Llama 3.1 8B Korean*	0.23
Proprietary reasoning	CLOVA HCX-DASH-001*	0.37	Middle	Qwen 2.5 14B	0.62
	GPT-o1	0.86		Qwen 2.5 32B	0.75
	Gemini 2.0 Flash Thinking	0.85		Qwen 2.5 72B	0.70
Light	Llama 3.2 3B	0.20	Big	Llama 3.1 70B	0.47
	Llama 3.2 3B Korean*	0.23		Llama 3.1 70B Korean*	0.55

Table 4: Evaluated LLMs and their accuracy on the CSAT Korean items. Asterisks indicate Korean fine-tuned models.

Bloom’s Level	Human	Full AI	Gap
Remembering	71.5	37.5	−34.0
Understanding	74.2	53.2	−21.0
Applying	67.1	51.1	−16.0
Analyzing	66.5	50.5	−16.0
Evaluating	63.8	47.8	−16.0
Creating	79.3	79.3	0.0
Overall	72.3	51.3	−21.0

Table 5: Accuracy of human examinees and the full AI group by Bloom’s taxonomy level. The gap denotes (Full AI – human) accuracy.

only 37.5%, yielding a 34.0 percentage-point gap. This unexpected gap at the lowest Bloom level motivates our later analysis of retrieval-like failure and distractor alignment in Section 5.3.

5.2 Capability and Alignment Robustness

Section 5.1 showed substantial accuracy variation across models. We now examine whether human–LLM alignment is primarily explained by these capability differences. Table 6 reports OA, MAE, PA, and DA by model group. In the raw scores, the full AI group achieves the highest PA and DA, while the Korean fine-tuned group also shows relatively high DA.

However, raw group-level alignment should not be interpreted as an intrinsic property of each group. Higher alignment may reflect model capability, since more accurate models are more likely to place response mass on the correct option, where human responses are typically concentrated. It may also reflect aggregation size, as larger groups produce smoother choice distributions that can mechanically increase similarity to the human distribution.

We therefore separate the analysis into three steps. First, we test the relationship between accuracy and PA/DA at the individual-model level. Second, we use group-size-matched permutation tests to control for aggregation effects. Third, we

Group	OA	MAE	PA	DA
Full AI	0.40***	0.21	0.89	0.81
Proprietary	0.34***	0.19	0.85	0.58
Reasoning	0.29***	0.26	0.81	0.56
Light	0.02	0.56	0.41	0.51
Small	0.18***	0.45	0.60	0.60
Middle	0.32***	0.31	0.73	0.55
Big	0.31***	0.31	0.72	0.57
Korean	0.25***	0.41	0.67	0.65

Table 6: CAPE alignment metrics by model group. OA measures item-level difficulty alignment (*** $p < .001$), MAE measures the absolute performance gap, PA measures full-choice response alignment, and DA measures distractor-only alignment. See Appendix F for per-group results by Bloom level.

inspect residual DA patterns after accounting for accuracy, interpreting them cautiously as model-specific deviations rather than stable group-level effects.

5.2.1 Capability and Response Alignment

To examine the relationship between capability and alignment, we compute accuracy, PA, and DA for each of the 16 models and correlate each alignment metric with accuracy (Figure 1). Accuracy is almost perfectly correlated with PA ($r = 1.00$, $p < .001$) and strongly correlated with DA ($r = 0.90$, $p < .001$). We then conduct a leave-one-model-out analysis to assess whether these relationships are driven by any single influential model. After excluding each of the 16 models in turn, the Accuracy–DA correlation remains between $r = .879$ and $r = .926$, while the Accuracy–PA correlation remains between $r = .997$ and $r = .998$. Thus, the overall relationships are not driven by any single model.

The near-perfect PA–accuracy relationship is expected because PA compares the full answer-choice distribution, including the correct option. As models answer more items correctly, their distributions

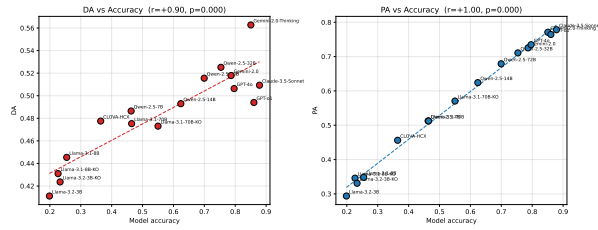


Figure 1: Relationship between model accuracy and response alignment metrics. The left panel shows the relationship between accuracy and distractor-only alignment (DA), while the right panel shows the relationship between accuracy and full-choice response alignment (PA). Accuracy is strongly correlated with both PA and DA, indicating that model capability is closely associated with correct-answer selection and human-like distractor preference.

move closer to the largest mass of the human distribution, which is typically concentrated on the correct answer. Thus, PA captures process-level similarity, but is also partly structurally coupled with correctness.

By contrast, the DA result is more informative. Since DA removes the correct option, its strong correlation with accuracy suggests that higher-capability models do not merely select correct answers more often. When they fail, they also tend to choose distractors that humans find plausible. Thus, capability is linked to both human-like success and human-like failure. Because individual-model DA is computed over each model’s observed failures, however, models may be compared over different failure sets. Cross-model DA comparisons should therefore be interpreted together with model accuracy.

5.2.2 Controlling for Group-Size Effects

Group-level alignment scores can be confounded by group size. As more models are aggregated, the resulting choice distribution becomes smoother, which can mechanically increase its similarity to the human response distribution. To control for this effect, we randomly sampled model sets of the same size k as each target group from the full set of 16 models 1,000 times. We then computed PA and DA for each sampled set to construct group-size-matched null distributions (Table 7).

The permutation results show that the expected DA tends to increase with group size. This indicates that the high DA of the full AI group may partly reflect a smoothing effect induced by model aggregation. Therefore, the raw group-level alignment

scores in Table 6 should be interpreted together with z -scores computed against random groups of the same size. Some groups show nominal positive deviations even after group size is controlled. The Korean fine-tuned group shows higher DA than expected under the group-size-matched null distribution ($z = +2.03$, uncorrected one-sided $p = .020$), and the small group shows a similar nominal deviation ($z = +1.58$, $p = .034$).

However, these nominal deviations do not survive correction for multiple comparisons across the seven non-independent group comparisons in Table 7. We therefore do not interpret them as stable group-specific alignment effects. Rather, they indicate that raw group-level DA remains sensitive to model composition even after group size is controlled.

5.2.3 Residual Alignment Beyond Capability

The permutation analysis controls for group size, but not for model capability. We therefore examine model-level DA residuals, defined as observed DA minus the DA predicted by the accuracy regression in Figure 1. These residuals indicate whether each model lies above or below the capability–DA trend, and are distinct from the group-size-matched deviations in Table 7.

The residuals are not consistent within model groups. The two reasoning-augmented models fall on opposite sides of the regression line, with Gemini-2.0-Thinking above the trend and GPT-o1 below it. Likewise, the nominal Korean-group deviation in Table 7 is not mirrored by consistently positive residuals across Korean-adapted models; only CLOVA-HCX clearly exceeds its accuracy-predicted DA. Thus, neither reasoning augmentation nor Korean adaptation provides consistent evidence of stronger human-like distractor alignment after accounting for capability.

Overall, Section 5.2 shows a strong capability–alignment relationship within the evaluated model set, while raw group-level scores may reflect aggregation size, model composition, and model-specific deviations rather than stable group-level effects. Section 5.3 next examines whether this relationship holds across cognitive levels.

5.3 Cognitive Level Modulates Capability-Linked Alignment

Section 5.2 showed a strong relationship between model capability and human–LLM alignment within the evaluated model set. In this section,

Group	k	Obs. DA	Null DA	DA z	DA p	PA z (p)
Korean	4	0.65	0.59	+2.03	0.020*	-2.46 (.972)
Small	3	0.60	0.55	+1.58	0.034*	-2.41 (.974)
Reasoning	2	0.56	0.52	+1.26	0.082	+1.28 (.023*)
Middle	2	0.55	0.52	+0.95	0.154	+0.44 (.401)
Big	3	0.57	0.55	+0.61	0.286	-0.31 (.701)
Light	2	0.51	0.52	-0.44	0.721	-2.89 (1.000)
Proprietary	4	0.58	0.61	-0.82	0.785	+1.06 (.102)

Table 7: Group-size-matched permutation results for PA and DA. Observed scores are the alignment values computed from the actual model groups, whereas null scores are the average alignments obtained from 1,000 randomly sampled model subsets with the same group size k . Positive z -scores indicate that the observed group alignment is higher than expected under the size-matched null distribution; p values are one-sided permutation tests ($p < .05$).

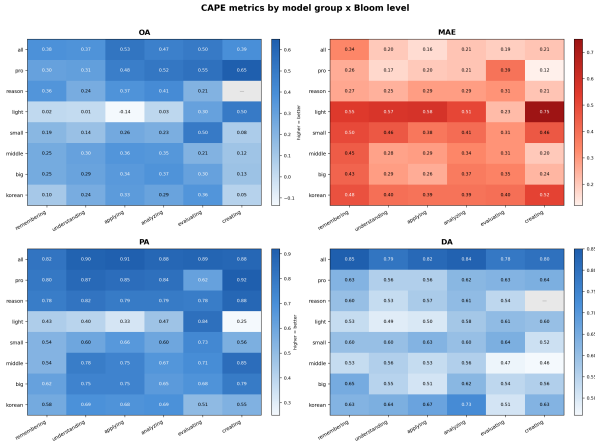


Figure 2: Heatmaps for CAPE metrics by model group and Bloom's taxonomy level.

we examine whether this capability–alignment relationship holds consistently across cognitive levels. To this end, we use Bloom's taxonomy not as a difficulty label, but as a cognitive probe for identifying when model capability translates into human-like response distributions and when it fails to reproduce human-like distractor structures.

Figure 2 presents the overall landscape of OA, MAE, PA, and DA across model groups and Bloom levels. Overall, PA and MAE show a relatively clear capability gradient along the model-group axis. However, this heatmap alone does not reveal whether capability predicts full-choice alignment and distractor-only alignment in the same way across cognitive levels.

We therefore conduct two analyses in this section. First, we compare the distribution of DA – PA using item-level bootstrapping within each Bloom level to examine whether distractor-only alignment exceeds full-choice alignment. Second, we compute the correlations between accuracy and PA/DA across the evaluated models within each Bloom level to analyze whether capability-linked align-

ment remains consistent across metrics.

5.3.1 Remembering: Human-like distractor alignment without successful retrieval

Table 8 reports the bootstrap distribution of DA – PA by Bloom level. Remembering is the only Bloom level with a positive DA – PA value (+0.025), and DA exceeds PA in 93.5% of the 1,000 bootstrap samples. This contrasts with the other Bloom levels, where PA generally exceeds DA, and motivates a closer examination of retrieval-like failure.

This pattern suggests that human-like distractor alignment and reliable retrieval may be dissociated at the Remembering level, whereas PA consistently exceeds DA at the other Bloom levels. Models may exhibit distractor preferences similar to those of human examinees while failing to translate them into correct-answer retrieval. Thus, the key point for Remembering is not simply that model accuracy is low, but that even at the lowest cognitive level, human-like failure patterns can be separated from successful retrieval.

One illustrative Remembering item required examinees to identify an inappropriate example of Middle Korean grammatical particle usage (Appendix G). Human accuracy was 84.0%, whereas model accuracy was only 18.8%. Nevertheless, a half of incorrect human responses and 53.8% of model errors were concentrated on the same distractor, resulting in very high distractor alignment (DA = .989). This suggests that, at least for some items, models may fail to identify the correct grammatical usage while still being attracted to the same distractor that human examinees find plausible.

5.3.2 Evaluating: An Exploratory Weakening of the Capability–DA Association

At the Evaluating level, we observe a weaker relationship between capability and distractor align-

Bloom's Level	DA - PA	95% CI	Pr(DA > PA)
Remembering	+0.025	[-0.007, +0.057]	0.935
Understanding	-0.103	[-0.117, -0.088]	0.000
Applying	-0.096	[-0.139, -0.058]	0.000
Analyzing	-0.047	[-0.082, -0.012]	0.006
Evaluating	-0.108	[-0.181, -0.046]	0.000
Creating	-0.077	[-0.186, +0.012]	0.050

Table 8: Bootstrap comparison of distractor-only alignment and full-choice alignment by Bloom level. DA-PA indicates the alignment difference; positive values mean that DA exceeds PA. Confidence intervals quantify uncertainty in the estimated difference, and Pr(DA > PA) indicates the proportion of bootstrap samples in which DA is greater than PA.

Bloom's Level	DA-Acc. (<i>r</i>)	95% CI	PA-Acc. (<i>r</i>)
Remembering	+0.84***	[0.587, 0.942]	+0.99***
Understanding	+0.84***	[0.593, 0.944]	+1.00***
Applying	+0.69**	[0.302, 0.885]	+0.99***
Analyzing	+0.76***	[0.416, 0.910]	+0.99***
Evaluating	+0.39	[-0.134, 0.741]	+0.99***
Creating	+0.27	[-0.333, 0.713]	+0.98***

Table 9: Bloom-level correlations between model accuracy and response alignment metrics (** $p < .01$; *** $p < .001$). DA-Acc. and PA-Acc. denote Pearson correlations between accuracy and DA/PA, respectively, consistent with the regression-based capability analysis in Figure 1. The 95% confidence intervals are reported for DA-Acc. Correlations are computed across the 16 LLMs except for Creating DA-Acc., which is based on the 13 models for which DA is defined and is interpreted as exploratory.

ment than at the preceding Bloom levels. As shown in Table 9, PA is almost perfectly correlated with accuracy across all Bloom levels ($r = 0.98$ – 1.00). This indicates that the full-choice distribution, which includes the correct option, becomes closer to the human response distribution as model capability increases.

However, DA shows a different pattern at the Evaluating level. From Remembering through Analyzing, accuracy remains strongly correlated with DA ($r = 0.69$ – 0.84). At Evaluating, by contrast, this association is much weaker ($r = .387$, 95% CI $[-.134, .741]$), despite the near-perfect PA-accuracy correlation reported above. Because the confidence interval includes zero, we interpret this pattern as an exploratory weakening of the capability-DA association rather than a confirmed decoupling.

One illustrative Evaluating item required examinees to compare an initial draft, its revision, and five suggestions to determine which suggestion had not been incorporated into the revised text (Ap-

pendix G). Human and model accuracy were 90.1% and 75.0%, respectively, and full-choice alignment was also high (PA = .900). However, humans and models most frequently selected different distractors, resulting in relatively low distractor alignment (DA = .576). Thus, even when models reach the correct answer relatively often, their errors may not reflect the same evaluative distinctions that confuse human examinees. This example is intended to aid interpretation of the aggregate result rather than provide additional statistical evidence.

5.3.3 Mid-level tasks as the stable region of capability-linked alignment

Understanding, Applying, and Analyzing levels do not show either of the contrasting patterns observed at Remembering and Evaluating. Across these levels, PA consistently exceeds DA (Table 8), and accuracy is significantly and positively correlated with both PA and DA (Table 9). This stable pattern is especially important because Understanding accounts for the largest share of the dataset. At these levels, greater model capability coincides with closer human-like alignment in both the full-choice and distractor-only distributions, without the retrieval-side inversion suggested at Remembering or the weaker capability-DA association observed at Evaluating.

Taken together, the Bloom-level analysis shows that the capability-alignment relationship varies across cognitive levels. At the Remembering level, human-like distractor alignment may not translate into reliable retrieval. At the Evaluating level, the capability-DA association is weaker and more uncertain. In contrast, Understanding, Applying, and Analyzing show more stable capability-linked alignment without either of these patterns. These findings demonstrate that Bloom's taxonomy functions as a cognitive probe for revealing the scope and limits of human-LLM alignment.

6 Discussion

This study reframes human-LLM alignment as how models succeed and fail, rather than how often they answer correctly. Across the evaluated models, more accurate models not only select more correct answers, but also tend to choose human-plausible distractors when they fail. This pattern links performance to human population behavior in both success and failure patterns, which accuracy alone cannot reveal.

At the same time, response alignment requires cautious interpretation. Full-choice alignment is tightly coupled with correctness, and raw group-level scores can be inflated by aggregation size or shaped by model composition. After controlling for group size and residual variation beyond accuracy, we find no consistent evidence that reasoning augmentation or Korean adaptation independently improves human-like distractor alignment.

Bloom-level analysis further shows that capability-linked alignment is not uniform across item types. Remembering suggests that human-like distractor preference can coexist with failed retrieval, whereas Evaluating suggests that improved correctness does not guarantee human-like distractor alignment. Understanding, Applying, and Analyzing show more stable capability-linked alignment. These patterns should be interpreted distributionally: high distractor alignment does not imply human-like reasoning, nor is it inherently desirable.

7 Conclusion

This paper introduced CAPE, a Choice-Distribution Alignment and Performance Evaluation framework for diagnosing human-LLM response alignment with population-scale human choice distributions. Across 609 CSAT Korean items and 16 LLMs, CAPE shows that alignment varies systematically with accuracy but is not reducible to it. More accurate models align better with humans in both correct-answer selection and plausible distractor choice. However, this alignment is not a stable property of model groups. It can reflect aggregation size, model composition, and variation across item types.

Taken together, CAPE shifts LLM evaluation from measuring correctness alone to diagnosing response patterns of success and failure. Rather than ranking models by apparent human-likeness, it identifies when model responses converge with or diverge from human population behavior, and whether such alignment is capability-driven, aggregation-driven, or model-specific. Because CAPE requires only multiple-choice items, model answers, and item-level human choice statistics, it can be extended to standardized assessments across languages and domains.

Limitations

This study analyzes human-LLM alignment in terms of correctness, difficulty perception, choice-distribution alignment, and distractor structure, but it has several limitations. First, our analysis does not directly measure the internal cognitive processes of individual models. Instead, it compares population-scale human choice distributions with model-group response distributions. The findings should therefore be interpreted as evidence of behavioral alignment at the response-distribution level, not as evidence that humans and LLMs share the same reasoning mechanisms.

Second, this study is based on the Korean language section of the CSAT. Although this setting provides large-scale human response distributions and Bloom-level annotations, it remains unclear whether the observed alignment patterns generalize to other languages, cultures, and task environments. Tasks such as mathematical reasoning, multimodal reasoning, and conversational interaction may involve different distractor structures and misconception patterns. Future work should extend CAPE to broader domains and combine population-level choice distributions with individual-model probabilities or reasoning traces to better examine the relationship between behavioral alignment and underlying mechanisms.

Ethics Statement

The human response data are publicly available, item-level aggregate choice-distribution statistics reported by a commercial exam-preparation platform. They contain no personally identifiable information or individual-level records, and we use them solely for non-commercial research purposes. Also, we used model outputs generated under the respective API or model-access terms. AI assistants were used only for language polishing and drafting support. All substantive research decisions, analyses, and final manuscript content were reviewed and controlled by the authors.

Acknowledgments

This research was supported by the Culture, Sports and Tourism R&D Program, funded by Ministry of Culture, Sports and Tourism, through the Korea Culture Technology Planning and Evaluation Institute, an affiliated institute of the Korea Creative Content Agency in 2026 (Project Name: Development of an AI Agent Integrating Korean Language

Knowledge for Personalized Language Consultation Services, Project Number: RS-2026-25506607, Contribution Rate: 100%). Also, We thank the anonymous reviewers for their thoughtful comments and constructive feedback. Their suggestions helped us improve the clarity, methodological presentation, and overall quality of the paper.

References

- Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A. Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. From language to cognition: How LLMs outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24321–24339.
- Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.
- Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Addison Wesley Longman.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. In *Political Analysis*, volume 31, pages 337–351.
- Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations (ICLR)*, pages 18623–18642.
- Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, and 1 others. 2025. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009.
- Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Distractor generation for multiple-choice questions with predictive prompting and large language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 48–63. Springer.
- Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. Longmans, London.
- Muhammad Tufail Chandio, Saima Murtaza Pandhiani, and Rabia Iqbal. 2016. Bloom's taxonomy: Improving assessment and teaching-learning process. *Journal of Education and Educational Development*, 3(2):203–221.
- Prateek Chhikara. 2025. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*.
- Youn-Hyeong Cho. 2007. A study on the delineation of the objective of the evaluation of language abilities in the entrance examinations. *Korean Language Education Research*, 10(1):75–102.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.
- Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew Lan. 2025. Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 31–45. Springer.
- Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2026. Are formal and functional linguistic mechanisms dissociated in language models? *Computational Linguistics*, pages 1–41.
- Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-EVAL: A comprehensive chinese K-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4194–4213.
- Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5040–5060.
- Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
- Thomas Huber and Christina Niklaus. 2025. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 5211–5246.

- Cameron R. Jones, Sean Trott, and Benjamin Bergen. 2024. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, 12:803–819.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). Preprint, arXiv:2207.05221.
- Ye-jee Kang, Hansaem Kim, Seoyoon Park, Joeun Kang, Yujin Kim, Jaewon Lee, Gayeon Jung, Gyuri Choi, Changsu Choi, Inho Won, Minjun Kim, Hyeonseok Lim, Kyungtae Lim, and Younggyun Hahm. 2024. Comprehensive evaluation of LLMs’ korean language proficiency. *HANGEUL*, 85(3):679–715.
- Ming Li, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. 2025. Can LLMs estimate student struggles? human-AI difficulty alignment with proficiency simulation for item difficulty prediction. *arXiv preprint arXiv:2512.18880*.
- Naiming Liu, Shashank Sonkar, and Richard G. Baraniuk. 2025. Do LLMs make mistakes like students? exploring natural alignment between language models and human error patterns. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 364–377. Springer.
- Xinyi Lu and Xu Wang. 2024. Generative students: Using LLM-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 16–27.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.
- Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 24–49.
- Youngjin Moon, Byunggon Min, and Younghee Yoo. 2004. Manual for developing test items for the language arts section of the college scholastic ability test. Technical report, Korea Institute for Curriculum and Evaluation.
- Animesh Nigohjkar, Bekhzodbek Moydinboyev, My Duong, and John Licato. 2025. [Giving ai personalities leads to more human-like reasoning](#). Preprint, arXiv:2502.14155.
- Bianca Raimondi and Maurizio Gabbriellini. 2026. Mechanistic interpretability of cognitive complexity in LLMs via linear probing using bloom’s taxonomy. *arXiv preprint arXiv:2602.17229*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 29971–30004. PMLR.
- Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*.
- Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.
- James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.
- Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, and Yun-Nung Chen. 2025. None of the above, less of the right: Parallel patterns between humans and LLMs on multi-choice question answering. *arXiv preprint arXiv:2503.01550*.
- Eliza Janel L. Tan, Keith Alexandre L. Ramos, Maria Elisha Kaye B. Nazario, Steven Vincent D. Lim, and Shirley B. Chu. 2024. AI to the test: Measuring ChatGPT’s objective accuracy in the SATs in comparison to human performance. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 153–158. IEEE.
- Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025. What limits LLM-based human simulation: LLMs or our design? *arXiv preprint arXiv:2501.08579*.
- Zheyuan Zhang, Jifan Yu, Juanzi Li, and Lei Hou. 2023. Exploring the cognitive knowledge structure of large language models: An educational diagnostic assessment approach. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1643–1650.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.
- Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José

Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

A CSAT-Bloom Example Items

Grammar × Remembering

[Korean original]

15. <보기>의 [A]에 들어갈 말로 적절한 것만을 있는 대로 고른 것은?

<보기>

학생: 선생님, 자기 소개서를 써 봤는데, 띄어쓰기가 맞는지 가르쳐 주시겠어요? 헛갈리는 부분을 표시해 왔어요.

양로원에 가서 봉사 활동을 했습니다. 사실 그 시간에 ㉠봉사 보다는 게임을 하고 싶었습니다. 그저 작은 일을 ㉡도울 뿐이었는데 ㉢너 밖에 없다며 행복해하시는 어르신들의 말씀을 들을 ㉣때 만큼은 마음이 뿌듯해했습니다.

선생님: 한글 맞춤법에 따르면, 문장의 각 단어는 띄어 써야 하지만, 조사는 예외적으로 그 앞말에 붙여 쓴다.

학생: 아, 그럼 [A]은/는 앞말에 붙여 써야 하는군요.

- ① ㉠'보다', ㉢'밖에' ② ㉡'뿐', ㉢'밖에' ③ ㉡'뿐', ㉣'만큼' ④ ㉠'보다', ㉡'뿐', ㉣'만큼' ⑤ ㉠'보다', ㉢'밖에', ㉣'만큼'

[English translation]

15. Choose all of the items that can appropriately fill [A] in the box below.

<Box>

Student: Teacher, I wrote my personal statement—could you check my word spacing? I marked the parts I'm unsure about.

“I went to a nursing home and did volunteer work. Honestly, at that time I wanted to play games (㉠**boda**/than) volunteering. I was only (㉡**ppun**/just) helping with small tasks, yet at the moments (㉢**mankeum**/at least) when I heard the elders—who were happy saying there was (㉣**bakke**/no one but) you—my heart felt full.”

Teacher: Under Korean orthography, each word in a sentence is written separately, but *josa* (postpositional particles) are exceptionally attached to the preceding word.

Student: Ah, so [A] should be attached to the preceding word.

- ① ㉠**boda**, ㉢**bakke** ② ㉡**ppun**, ㉢**bakke** ③ ㉡**ppun**, ㉣**mankeum** ④ ㉠**boda**, ㉡**ppun**, ㉣**mankeum** ⑤ ㉠**boda**, ㉢**bakke**, ㉣**mankeum**

Rationale. The item asks which underlined forms are *josa* that must attach to the preceding word—i.e., recall and application of an explicit spacing rule rather than inference. We therefore map it to *Remembering*.

Speaking × Understanding

[Korean original]

(가)는 텔레비전 방송의 인터뷰이고, (나)는 (가)를 시청하고 산림 치유 프로그램에 참여한 학생이 쓴 수기이다. 물음에 답하십시오.

(가)

진행자: 산림 치유에 대해 알아보고자 ◇◇ 국립 산림 치유원의 산림 치유 지도사 이○○ 님을 모셨습니다. 안녕하세요.

지도사: 안녕하세요.

진행자: 시청자 분들께 산림 치유와 산림 치유 프로그램에 대해 간단히 소개해 주시겠어요?

지도사: 산림 치유란 피톤치드, 나뭇잎의 초록색 등과 같은 숲의 환경 요소로 심신의 건강을 회복시키는 것입니다. 산림욕, 숲 치료라고들 하시는데요, 공식 명칭은 산림 치유입니다. (중략)

진행자: 끝으로 시청자 분들께 한 말씀 해 주시죠.

지도사: 숲은 마음을 토닥여 주는 친구입니다. 숲으로 오세요.

진행자: 오늘 좋은 말씀 감사합니다.

4. (가)에 나타난 의사소통 방식으로 적절하지 않은 것은?

- ① '진행자'는 '지도사'의 답변에 자신의 의견을 덧붙이고 있다.
 ② '지도사'는 '진행자'가 잘못 이해하고 질문한 내용을 바로잡아 주고 있다.
 ③ '진행자'는 '지도사'의 답변에 대한 추가 정보를 요청하는 질문을 하고 있다.
 ④ '진행자'는 자신의 경험을 언급하며 '지도사'의 질문에 대해 답변하고 있다.
 ⑤ '지도사'는 기대되는 긍정적인 결과를 언급하며 '진행자'의 참여를 권유하고 있다.

[English translation]

(A) is a television interview; (B) is a reflective essay written by a student who watched (A) and joined a forest-therapy program. Answer the question.

(A)

Host: To learn about forest therapy, we have invited Mr./Ms. Lee, a forest-therapy instructor at the ◇◇ National Forest Therapy Center. Hello.

Instructor: Hello.

Host: Could you briefly introduce forest therapy and its programs to our viewers?

Instructor: Forest therapy restores physical and mental health using environmental elements of the forest, such as phytoncides and the green of leaves. People call it forest bathing or forest healing, but the official term is *forest therapy*. (omitted)

Host: Finally, please share a few words with our viewers.

Instructor: The forest is a friend that comforts your heart. Come to the forest.

Host: Thank you for your insightful words today.

Which of the following is not an appropriate description of the communication in (A)?

- ① The host adds their own opinion to the instructor's answers.
- ② **The instructor corrects content that the host misunderstood and asked about.**
- ③ The host asks questions requesting additional information about the instructor's answers.
- ④ The host answers the instructor's questions by referring to their own experience.
- ⑤ The instructor invites the host's participation by mentioning expected positive outcomes.

Rationale. The item requires comprehending the discourse roles and intentions exchanged in the interview and judging which described communicative behavior did not actually occur. This tests comprehension of the passage's meaning rather than factual recall or inference across passages → *Understanding*.

Writing × Applying

[Korean original]

(가)

부원 1: 도서관에 전파 식별 시스템을 설치하고 새로 단장한 지 일주일이지났네! 학생들 반응이 좋은 것 같아. 그렇지 않나?

부원 2: 맞아. 반 친구들이 도서관 이용이 편리해졌다는 말을 많이 하더라. 그런 말을 들으니, 전파 식별 시스템을 도입하자고 건의하길 잘했다는 생각이 들었어.

부원 3: 전에는 대출이나 반납을 하려면 줄을 길게 서야 했는데, 이젠 안 그래도 돼서 좋다고 하더라. (후략)

(나)

우리 학교 도서관이 새 단장을 마쳤습니다. 전파 식별 시스템 도입으로 달라진 우리 학교 도서관을 여러분에게 소개합니다.

[수월해진 대출, 반납] 길게 줄 서는 일 없이 대출·반납이 가능합니다. 전파 식별 시스템 덕분에 바코드를 일일이 찍지 않고, 한 번에 여러 권을 동시에 처리할 수 있기 때문입니다. (중략) 더욱 편리해진 도서관, 앞으로도 많은 이용 바랍니다. — ○○고등학교 도서관

40. 다음은 (가)의 ㉠~㉣을 바탕으로 '부원 1'이 작성한 메모의 일부이다. 메모의 내용이 (나)에 반영된 양상으로 적절하지 않은 것은?

- ① ㉠을 바탕으로 작성된 메모의 '여러 권 동시 처리 가능'은, 대출에 소요되는 시간이 줄었다는 내용으로 (나)에 반영되었다.
- ② ㉡을 바탕으로 작성된 메모의 '장서 점검기 활용의 이점'은, 이용 편의를 고려해서 도서를 정리했다는 내용으로 (나)에 반영되었다.
- ③ ㉢을 바탕으로 작성된 메모의 '도서부원 없이도 대출 가능'이라는 내용은, 과거에는 도서부원이 있을 때에만 도서 대출이 가능했다는 내용과 함께 (나)에 반영되었다.
- ④ ㉣을 바탕으로 작성된 메모의 '도서 분실 방지 장치 작동 중'이라는 내용은, 정보음이 울렸을 때의 대처 방안과 함께 (나)에 반영되었다.
- ⑤ ㉤을 바탕으로 작성된 메모의 '도서관 행사'는, 행사의 종류를 소개하는 내용으로 (나)에 반영되었다.

[English translation]

(A)

Member 1: It's been a week since we installed the RFID system and renovated the library! Students seem to like it. Don't you think?

Member 2: Yes. Many classmates say using the library has become more convenient. Hearing that, I felt it was a good idea to have proposed adopting the RFID system.

Member 3: Before, you had to wait in long lines to borrow or return books, but now you don't, and they like that. (omitted)

(B)

Our school library has finished its renovation. We introduce to you how it has changed with the RFID system.

[Easier borrowing and returning] You can borrow and return books without waiting in long lines. Thanks to the RFID system, books are processed several at once without scanning each barcode. (omitted) A more convenient library—we hope you keep using it. — ○○ High School Library

40. The following is part of a memo written by “Member 1” based on ㉠–㉢ in (A). Which description of how the memo’s content is reflected in (B) is not appropriate?

- ① The memo’s “multiple books processed at once” (from ㉠) is reflected in (B) as reduced borrowing time.
- ② The memo’s “benefit of using the inventory checker” (from ㉡) is reflected in (B) as arranging books for user convenience.
- ③ The memo’s “borrowing possible without a library member” (from ㉢) is reflected in (B) together with the claim that borrowing was previously possible only when a member was present.
- ④ The memo’s “anti-loss device in operation” (from ㉢) is reflected in (B) together with how to respond when the alarm sounds.
- ⑤ The memo’s “library events” (from ㉢) is reflected in (B) as an introduction to the types of events.

Rationale. The student must take the source material (the planning dialogue and memo) and apply it to a new written artifact, judging whether each planning point was correctly carried over into the promotional text. This requires applying given content to a concrete writing task rather than recalling or merely comprehending it → *Applying*.

Reading × Analyzing

[Korean original]

(가)

미학은 예술과 미적 경험에 관한 개념과 이론에 대해 논의하는 철학의 한 분야로서, 미학의 문제들 가운데 하나가 바로 예술의 정의에 대한 문제이다. 예술이 자연에 대한 모방이라는 아리스토텔레스의 말에서 비롯된 모방론은, 대상과 그 대상의 재현이 닮은꼴이어야 한다는 재현의 투명성 이론을 ㉠전제한다. (중략) 예술의 정의와 관련된 이 논의들은 예술로 분류할 수 있는 작품들의 공통된 본질을 찾는 시도이자 예술의 필요충분 조건을 찾는 시도이다.

(나)

예술 작품을 어떻게 감상하고 비평해야 하는지에 대해 다양한 논의들이 있다. 예술 작품의 의미와 가치에 대한 해석과 판단은 작품을 비평하는 목적과 태도에 따라 달라진다. 예술 작품에 대한 주요 비평 방법으로는 맥락주의 비평, 형식주의 비평, 인상주의 비평이 있다. (중략) 즉, 인상주의 비평가는 작가의 의도나 그 밖의 외적인 요인들을 고려할 필요 없이 비평가의 자유 의지로 무한대의 상상력을 가지고 작품을 해석하고 판단한다.

20. (가)와 (나)의 공통적인 내용 전개 방식으로 가장 적절한 것은?

- ① 대립되는 관점들이 수렴되어 가는 역사적 과정을 밝히고 있다.
- ② 화제에 대한 이론들을 평가하여 종합적 결론을 도출하고 있다.
- ③ 화제가 사회에 미치는 영향들을 분석하여 서로 간의 차이를 밝히고 있다.
- ④ 화제와 관련된 관점의 문제점을 제시하고 대안적 관점을 소개하고 있다.
- ⑤ 화제와 관련된 하나의 사례를 중심으로 다양한 이론을 시대순으로 나열하고 있다.

[English translation]

(A)

Aesthetics is a branch of philosophy that discusses concepts and theories about art and aesthetic experience; one of its problems is the definition of art. Mimeticism, originating from Aristotle's claim that art is an imitation of nature, ③ presupposes the theory of representational transparency—that an object and its representation must resemble each other. (omitted) These discussions on the definition of art are attempts to find the common essence of works classifiable as art, i.e., the necessary and sufficient conditions for art.

(B)

There are various discussions on how to appreciate and criticize artworks. The interpretation and evaluation of an artwork's meaning and value differ according to the purpose and attitude of criticism. Major critical methods include contextualist, formalist, and impressionist criticism. (omitted) That is, the impressionist critic interprets and judges a work with unlimited imagination by their own free will, without needing to consider the artist's intention or other external factors.

20. Which is the most appropriate description of the method of development *common to both* (A) and (B)?

- ① Revealing a historical process in which opposing viewpoints converge.
- ② Evaluating theories on the topic and drawing a comprehensive conclusion.
- ③ Analyzing the topic's effects on society and revealing differences among them.
- ④ **Presenting problems with a viewpoint on the topic and introducing an alternative.**
- ⑤ Listing various theories chronologically around a single example related to the topic.

Rationale. The item requires reading two separate passages, breaking each down into its rhetorical structure (how each develops its topic), and comparing those structures to identify a shared pattern. This cross-passage decomposition and comparison goes beyond comprehending a single text → *Analyzing*.

B CSAT–Bloom Mapping: Cross-tabulation

Bloom's Level	Literature	Reading	Speaking	Writing	Grammar	Cl. Grammar	Total
Remembering	0	29	0	0	38	9	76
Understanding	200	135	46	15	0	0	396
Applying	1	24	5	12	0	0	42
Analyzing	46	15	0	0	0	0	61
Evaluating	0	16	1	6	0	0	23
Creating	1	0	1	9	0	0	11
Total	248	219	53	42	38	9	609

Table 10: Cross-tabulation of CSAT Korean items by subdomain and Bloom's taxonomy level. Row totals match the per-level counts in Table 1 and Table 3. "Cl. Grammar" denotes Classical Grammar.

C Criteria for Mapping CSAT Understanding Layers to Bloom's Taxonomy

CSAT Understanding Layer	Bloom's Level	Mapping Criterion
Vocabulary & Concepts	Remembering	Grammar and vocabulary questions.
Factual Understanding	Understanding	Questions whose answers are directly found within the text.
Inferential Understanding	Understanding	Questions involving multiple cognitive levels.
	Applying	Inferential reasoning questions.
	Analyzing	Questions comparing two passages or elements.
Critical Understanding	Evaluating	Questions requiring critique or revision through critical analysis.
Application/Creativity	Creating	Questions involving planning, writing, or completing texts and creating media.

Table 11: Criteria for mapping CSAT Korean understanding layers to Bloom's taxonomy. The Inferential Understanding layer maps to multiple Bloom levels depending on the cognitive operation each item requires. Annotators were instructed to assign each item to the Bloom level that best matched the primary cognitive operation required to solve the item, using the criteria in Table 11. They were compensated at an hourly rate for their work(\$10/hour).

D Model Details / Addresses

Category	Model	Version
Proprietary	Claude 3.5	claude-3-5-sonnet-20241022
	GPT-4o	gpt-4o-2024-11-20
	Gemini 2.0	gemini-2.0-flash
	CLOVA	HCX-DASH-001
Proprietary reasoning	GPT-o1	o1-2024-12-17
	Gemini 2.0-thinking	gemini-2.0-flash-thinking
Light open-source	Llama 3.2 3B	meta-llama/Llama-3.2-3B-Instruct
	Llama 3.2 3B KO*	Bllossom/llama-3.2-Korean-Bllossom-3B
Small open-source	Qwen 2.5 7B	Qwen/Qwen2.5-7B-Instruct
	Llama 3.1 8B	meta-llama/Llama-3.1-8B-Instruct
	Llama 3.1 8B KO*	Bllossom/llama-3.1-Korean-Bllossom-Vision-8B
Middle open-source	Qwen 2.5 14B	Qwen/Qwen2.5-14B-Instruct
	Qwen 2.5 32B	Qwen/Qwen2.5-32B-Instruct
Big open-source	Qwen 2.5 72B	Qwen/Qwen2.5-72B-Instruct
	Llama 3.1 70B	meta-llama/Llama-3.1-70B-Instruct
	Llama 3.1 70B KO*	Saxo/Linkbricks-Horizon-AI-Korean-llama3.1-sft-dpo-70B

Table 12: Evaluated models and their exact versions or HuggingFace identifiers. Asterisks (*) indicate Korean fine-tuned models. Proprietary models were accessed via provider APIs, while open-source models were evaluated using their released checkpoints; exact computational cost is therefore not directly comparable across models.

E Experiment Prompts

Prompt Template

[Korean original — system prompt]

당신은 한국어를 모국어로 하는 화자입니다. 당신은 지금부터 아래 문제에 대한 답변을 제출할 것입니다. 문제는 전체 문제 집합에 대한 문항 범위와 대분류에서의 문제 내용을 알려주는 `meta_question`, 문제에 대한 지문인 `context`, <보기>가 있을 경우 참고하는 `sub_context`, 그리고 문제인 `question`과 선택지인 `choice`로 구성되어 있으며, 이 정보들을 참고해서 문제를 풀어야 합니다. 응답은 선택지 ①~⑤ 중 하나를 고르면 됩니다. 이제 문제를 푸세요.

[User message format]

메타 질문: {`meta_question`}
문맥: {`context`}
보조 문맥: {`sub_context`}
질문: {`question`}
선택지: {`choice`}

[English translation — system prompt]

“You are a native speaker of Korean. You will now submit an answer to the problem below. Each problem consists of: `meta_question`, which gives the item’s position in the full item set and its top-level category; `context`, the passage for the problem; `sub_context`, referenced when a <box> (*bogi*) is present; and `question`, the question itself, with `choice`, the answer options. You must solve the problem using this information. Respond by choosing one of options ①–⑤. Now solve the problem.”

[User message format, translated]

`meta_question`: {`meta_question`}
`context`: {`context`}
`sub_context`: {`sub_context`}
`question`: {`question`}
`choice`: {`choice`}

Note. All models received the same zero-shot prompt. The system prompt and the formatted item were concatenated into a single input. The model’s selected option (①–⑤) was parsed from its response.

F Full Per-Group Results: OA/MAE/PA/DA

Table 13: Per-group CAPE metrics by Bloom’s taxonomy level. OA significance: $*p < .05$, $**p < .01$, $***p < .001$ (permutation). Dashes (–) denote values undefined at the Creating level for the Reasoning group, where all items were answered correctly (no accuracy variance for OA; no distractor mass for DA). These are Bloom-level breakdowns; the aggregate group scores in Table 6 differ because pooling all items yields smoother distributions.

Group	Bloom Level	N	OA	OA p	MAE	PA	DA
Full AI	Remembering	76	0.38**	0.001	0.34	0.59	0.64
	Understanding	396	0.37***	0.000	0.20	0.68	0.58
	Applying	42	0.53***	0.000	0.16	0.71	0.61
	Analyzing	61	0.47***	0.000	0.21	0.68	0.63
	Evaluating	23	0.49*	0.017	0.19	0.69	0.56
	Creating	11	0.39	0.233	0.21	0.67	0.58
Proprietary	Remembering	76	0.30**	0.008	0.26	0.57	0.42
	Understanding	396	0.31***	0.000	0.17	0.65	0.36
	Applying	42	0.48**	0.001	0.20	0.63	0.36
	Analyzing	61	0.52***	0.000	0.21	0.61	0.40
	Evaluating	23	0.30	0.170	0.23	0.62	0.40
	Creating	11	0.65*	0.031	0.12	0.72	0.41
Reasoning	Remembering	76	0.36**	0.001	0.27	0.56	0.38
	Understanding	396	0.24***	0.000	0.25	0.61	0.33
	Applying	42	0.37*	0.016	0.29	0.57	0.35
	Analyzing	61	0.41**	0.001	0.29	0.56	0.39
	Evaluating	23	0.21	0.342	0.31	0.57	0.33
	Creating	11	–	–	0.21	0.69	–
Light	Remembering	76	0.02	0.860	0.55	0.27	0.33
	Understanding	396	0.01	0.828	0.57	0.25	0.29
	Applying	42	-0.14	0.394	0.58	0.21	0.30
	Analyzing	61	0.03	0.817	0.51	0.29	0.36
	Evaluating	23	0.36	0.091	0.40	0.34	0.31
	Creating	11	0.50	0.116	0.75	0.14	0.39
Small	Remembering	76	0.19	0.100	0.50	0.33	0.39
	Understanding	396	0.14**	0.007	0.46	0.39	0.39
	Applying	42	0.26	0.100	0.38	0.43	0.41
	Analyzing	61	0.23	0.071	0.41	0.38	0.38
	Evaluating	23	0.55**	0.007	0.39	0.40	0.40
	Creating	11	0.08	0.813	0.46	0.37	0.32
Middle	Remembering	76	0.25*	0.029	0.45	0.37	0.34
	Understanding	396	0.30***	0.000	0.28	0.57	0.35
	Applying	42	0.36*	0.020	0.29	0.53	0.32
	Analyzing	61	0.35**	0.006	0.34	0.47	0.35
	Evaluating	23	0.21	0.336	0.31	0.51	0.28
	Creating	11	0.12	0.726	0.20	0.64	0.28

Group	Bloom Level	N	OA	OA p	MAE	PA	DA
Big	Remembering	76	0.25*	0.029	0.43	0.41	0.42
	Understanding	396	0.29***	0.000	0.29	0.54	0.34
	Applying	42	0.34*	0.027	0.26	0.53	0.31
	Analyzing	61	0.37**	0.004	0.37	0.44	0.41
	Evaluating	23	0.30	0.161	0.35	0.48	0.34
	Creating	11	0.13	0.697	0.24	0.57	0.35
Korean	Remembering	76	0.10	0.402	0.48	0.35	0.41
	Understanding	396	0.24***	0.000	0.40	0.45	0.42
	Applying	42	0.33*	0.031	0.39	0.45	0.44
	Analyzing	61	0.29*	0.023	0.39	0.46	0.50
	Evaluating	23	0.50*	0.016	0.31	0.49	0.42
	Creating	11	0.05	0.893	0.52	0.36	0.41

G Representative Response Distributions

Table 14 reports the full choice distributions for the representative Remembering and Evaluating items discussed in Section 5.3. These examples illustrate the aggregate patterns described in the main text and are not intended as additional statistical evidence.

Bloom	Domain	Source	①	②	③	④	⑤
Remembering	Middle Korean	Human 16 LLMs	84.0 18.75	4.0 12.50	8.0 43.75	3.0 18.75	1.0 6.25
Evaluating	Writing	Human 16 LLMs	2.0 6.25	2.0 18.75	2.0 0.0	90.1 75.0	4.0 0.0

Table 14: Human and LLM choice distributions (%) for the representative Remembering and Evaluating items discussed in Section 5.3. Bold values indicate the correct option. Human choice shares are reported as rounded percentages from the source response statistics and therefore may not sum exactly to 100. LLM choice shares are computed from the responses of the 16 evaluated models.