ProcrastiCoT: A Dual-Encoder Cross-Attention Framework for Multi-Dimensional Procrastination Motives Assessment

Chia-Ni Chiang Chia-Yun Lee Hsin-Hsuan Chen Jheng-Long Wu
Department of Data Science, Soochow University, Taiwan
nianna9316@gmail.com jessicaleej0033@gmail.com
jaylachen1012@gmail.com jlwu@gm.scu.edu.tw

Abstract

Procrastination is driven by emotional self-regulation failures and cognitive distortions that narratives often mask. We introduce **ProcrastiType-ZH**, to our knowledge the first Traditional Chinese multi-dimensional procrastination dataset, with 1,000 LLM-generated narratives and human consensus labels for four motives on a 1–5 severity scale (effective range 1–4). We propose **ProcrastiCoT**, a theory-informed Chain-of-Thought framework, and inject its LLM-generated reasoning into BERT-size encoders either by concatenation or through a **Dual-Encoder Cross-Attention** architecture that gives the reasoning its own context window. Across backbones, reasoning injection gives the best single-run QWK on Avoidant, Decisional, and Irrationality. On our primary CKIP-BERT backbone, the Dual-Encoder scores above its pure-encoder counterpart on every dimension in each of four paired training runs, significantly (uncorrected) on Arousal, Decisional, and Irrationality but not Avoidant; on Decisional its mean (0.316) only matches a lexical baseline. A Qwen2.5-7B model fine-tuned on the same labels, with about 70 times more parameters, is strongest on Arousal (QWK 0.709) but below the Dual-Encoder on Irrationality and Avoidant in all four runs. A variant trained to also generate the teacher's reasoning scores lower on every dimension, with predictions closer to the teacher's ratings; one possible explanation is a conflict between that reasoning and the human labels.

1 Introduction

Procrastination is increasingly seen as an emotion-regulation problem rather than a time-management deficit (Sirois and Pychyl, 2013), driven by latent motives such as thrill-seeking (*Arousal*), fear of failure (*Avoidant*), and difficulty choosing (*Decisional*) (Ferrari, 1992; Mann et al., 1997). This typology is contested: Steel (2010) finds that such items load largely on a single general factor. We therefore add *Irrationality* for non-specific self-regulation failure and test in §3.3 whether the four dimensions behave as distinct constructs in our data. Detecting these motives from self-narratives could inform personalized interventions.

Assessing them computationally faces three challenges: to our knowledge, no specialized Chinese procrastination narrative dataset exists; motives are often masked by rationalization (e.g., claiming "I work better under pressure" to hide anxiety); and severity is ordinal and multi-dimensional rather than binary. We make three contributions:

1. **ProcrastiType-ZH**, to our knowledge the first Traditional Chinese multi-dimensional procrastination narrative dataset (1,000 LLM-generated narratives) with human consensus severity labels; the corpus, rubric, and generated reasoning are available from the authors on request for non-commercial research use.

2. **ProcrastiCoT**, a theory-informed Chain-of-Thought framework whose four nodes (emotional regulation, decisional conflict, temporal discounting, and sensation-seeking), drawing on cognitive therapy (Beck, 1976), decisional-conflict theory, temporal motivation theory, and procrastination typologies, turn narratives into interpretable representations of latent motives.

3. **Reasoning injection** into BERT-size encoders, either by concatenation within a single stream or through a **Dual-Encoder Cross-Attention** architecture in which narrative tokens *query* separately encoded reasoning. Across backbones it gives the best single-run QWK on three of the four dimensions (two on our primary CKIP-BERT), and on CKIP-BERT the Dual-Encoder scores above its pure encoder on every dimension in all four paired

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1149–1162
November 6-10, 2026 ©2026 Association for Computational Linguistics

training runs, significantly (uncorrected) on three; the teacher call is a one-off preprocessing step for a fixed corpus (§5.2).

LLM-generated reasoning, dual encoders, and cross-attention are each established; we contribute their instantiation for multi-dimensional severity assessment and an empirical characterization: the benefit of reasoning injection does not appear to hinge on context length (§5.3.2); a 7B model fine-tuned on the same labels is a strong but not dominant baseline, which scores lower when also trained to generate the teacher's reasoning (§5.3.4); single-run Decisional differences are within run-to-run variance (§5.3.5); and the reasoning survives an audit whose limits we state (§4.2).

2 Related Work

2.1 NLP for Mental Health and Synthetic Data

Prior NLP work, including the CLPsych workshops, predicts depression, PTSD, and anxiety from social media (Coppersmith et al., 2014; Chancellor and De Choudhury, 2020), mostly as binary classification or risk-level screening; graded, multi-dimensional severity assessment remains under-explored. LLM-generated text is a viable way to build specialized datasets (Wang et al., 2023; Long et al., 2024), and persona-conditioned generation avoids the cost and privacy constraints of real data but introduces systematic biases (Li et al., 2025), which we treat as a limitation while grounding labels in human consensus (§3.2).

2.2 Reasoning Injection and Dual Encoders

Chain-of-Thought prompting advances LLM reasoning (Wei et al., 2022; Kojima et al., 2022) but is slow to generate at deployment; passing such reasoning to a smaller encoder such as BERT (Devlin et al., 2019) is what we term *reasoning injection*. Dual (Siamese) encoders encode sequences independently (Reimers and Gurevych, 2019), which also gives each its own context window. Bi-encoders interact only through a final similarity, and late-interaction models such as ColBERT (Khattab and Zaharia, 2020) match token-level representations; we let narrative tokens attend to the reasoning through cross-attention.

3 The ProcrastiType-ZH Dataset

3.1 Narrative Generation Pipeline

To generate diverse Traditional Chinese procrastination texts, we designed a structured persona-scenario matrix. We used GPT-4o-mini (OpenAI, 2024) (accessed April 2026) to generate 1,000 narratives. We crossed 8 distinct personas (e.g., Anxious, Rational, Complaining) with 7 realistic task contexts spanning academic delay, high-stakes deadlines, tedious routine chores, and ambiguous technical choices; the full matrix is listed in Appendix B. Each narrative was generated from four target intensities (1–5, one per dimension), whose sum set its intended severity level (Low 4–9, Medium 10–15, High 16–20; 400, 200, and 400 narratives); annotators did not see these targets. Persona–context cells hold 9–26 narratives each. Each context is posed as a job-interview question (e.g., "Please share an experience of handling a difficult project"), which the generator answers in the first person as a candidate, and the prompt asks it to present the delay as a professional challenge and to avoid explicit terms such as "procrastination" or "distraction" (325 of the 1,000 narratives nonetheless use one of these two terms).

3.2 Annotation Protocol

We employed a structured annotation protocol involving three human annotators, all undergraduate students with a minor in psychology, trained on the behavioral severity rubric. The annotators are three of the authors of this paper; they received no separate payment for annotation, which they carried out as part of their own research, and knew the narratives were synthetic. The process was designed to measure observable *Behavioral Severity* rather than speculative subjective intent.

The protocol ran in three phases. In the **calibration phase**, a pilot subset of 50 narratives was annotated *independently* by each annotator; the team then reviewed all cases to unify scoring boundaries, particularly for distinguishing thrill-seeking (Arousal) from fear-of-failure (Avoidant). In the **consensus annotation phase**, the remaining 950 narratives were reviewed *jointly* by all three annotators, who produced a single agreed rating per dimension through real-time discussion against textual evidence. Ratings use a 5-point behavioral scale from 1 (no procrastination indicators) to 5 (delay causing task failure or harm to others); anchors are in Appendix C. Finally, a **reliability validation**

Dimension	Score 1	Score 2	Score 3	Score 4	Score 5
Arousal	81.4%	12.3%	1.0%	5.3%	–
Avoidant	26.1%	1.1%	5.6%	67.1%	0.1%
Decisional	0.8%	–	39.1%	60.1%	–
Irrationality	71.5%	–	2.4%	26.0%	0.1%

Table 1: Label distribution: percentage of narratives assigned each severity level in the final human consensus labels ($N = 1{,}000$). This table reports the *distribution of labels*; it is not an inter-annotator agreement statistic. Reliability evidence is reported in §3.2.

phase applied the three post-hoc checks below.

On the absence of classical inter-annotator agreement. Following the *prescriptive* paradigm of Röttger et al. (2022), we adopt consensus because our task needs a single defensible severity label rather than a distribution of opinions. A direct consequence is that no independent per-annotator scores exist for the 950 consensus-annotated narratives, so classical agreement statistics (Cohen's or Fleiss' κ) *cannot* be computed for the main corpus. We compensate with three checks: a two-week **test–retest** on 50 re-annotated narratives ($\kappa_w = 0.819$, 0.790, 0.634, 0.986 for Arousal, Avoidant, Decisional, Irrationality), the closest available analogue to an agreement coefficient, though on the 50 calibration narratives whose boundaries had already been reconciled, so an upper bound for the corpus; **gold-standard validation** against 20 researcher-authored prototypes with known intended severities (exact match 80–100%); and **scale-prototype validation**, in which the team correctly classified all four canonical prototype items drawn from the source constructs. The calibration pilot *was* scored independently, but its per-annotator ratings were not preserved, so no κ can be reported for it. A worked example is in Appendix E.

3.3 Dataset Analysis and Statistical Validity

3.3.1 Label Distribution and Imbalance

As Table 1 shows, the *Avoidant* dimension dominates (67.2% rated ≥ 4), a property of this generated corpus rather than an estimate of real-world prevalence. Arousal is scarce (5.3% rated ≥ 4).

The effective operating range is 1–4. Score 5 accounts for $\leq 0.1\%$ of labels in every dimension, so QWK, although computed over the full 1–5 scale, describes discrimination *within* 1–4. We did not rebalance the distribution, which reflects what annotators observed, but models trained here should not be expected to identify the most severe cases.

3.3.2 Construct Validity

As shown in Table 5 (Appendix D), the low average absolute pairwise Pearson correlation ($|r| = 0.102$) is consistent with the four dimensions representing distinct, non-redundant constructs in our data. The rubric could affect these correlations in either direction: shared severity anchors may inflate them, whereas calibration that separated the dimensions may deflate them. The weak positive correlation between Avoidant and Decisional ($r = +0.221, p < 0.001$) is consistent with decisional-conflict accounts in which avoidant coping extends to postponing choices (Mann et al., 1997).

3.3.3 Known-Group Validity

We partitioned narratives into Low ($n = 401$), Medium ($n = 434$), and High ($n = 165$) severity groups by the sum of the human ratings (cut points ≤ 9, 10–12, > 12). Because the grouping derives from the same ratings, this checks internal ordering rather than an external criterion: mean scores for all dimensions increase monotonically across the three groups, and the Low-versus-High contrast is significant for every dimension (two-sided Mann-Whitney U, $p < 0.001$; Table 6, Appendix D).

3.3.4 Dual-Framework Discrepancy

Finally, we analyze the gap between the LLM's intended generation severity and actual human consensus (Table 7, Appendix D). The exact match rate is only 49.2%, with humans systematically rating narratives lower by a -2.55 score-point margin on the summed severity scale. One plausible explanation is that the generator signals severity through explicit emotional vocabulary while the annotators graded functional impairment; either way, the labels are not an echo of the generation intent.

4 Methodology

4.1 ProcrastiCoT Prompt Design

ProcrastiCoT translates theory-grounded psychometric assessment into structured, machine-interpretable nodes. Rather than prompting for direct classification, we design a multi-node theory-guided reasoning chain rooted in cognitive therapy (Beck, 1976) and Temporal Motivation Theory (TMT) (Steel, 2007).

The framework decomposes assessment into four theory-guided nodes, each targeting one dimension: **Emotion & Cognitive Defense**

($\mathcal{N}_{\text{emotion}}$; Avoidant), grounded in Beck's cognitive therapy; **Decisional Conflict** ($\mathcal{N}_{\text{decision}}$; Decisional), grounded in the Melbourne Decision Making Questionnaire (Mann et al., 1997); **Temporal Discounting & Self-Regulation** ($\mathcal{N}_{\text{time}}$; Irrationality), grounded in TMT; and **Sensation Seeking & Self-Handicapping** ($\mathcal{N}_{\text{arousal}}$; Arousal), grounded in Ferrari (1992). Appendix B details each node.

The system prompt enforces a standardized JSON schema so that the model outputs a single well-formed object containing the reasoning texts for each node and the final severity ratings:

$$\text{Output} = \{\text{cot} : \{\mathcal{N}_{\text{emo}}, \mathcal{N}_{\text{dec}}, \mathcal{N}_{\text{time}}, \mathcal{N}_{\text{aro}}\},\ \text{scores} : \{y_{\text{aro}}, y_{\text{avo}}, y_{\text{dec}}, y_{\text{irr}}\}\} \tag{1}$$

Using GPT-5.4-mini (OpenAI, 2026) as the teacher LLM with this theory-aligned system prompt, we generate structured theory-guided analytical reasoning ($\mathcal{R}_{\text{ProcrastiCoT}}$) for all 1,000 narratives.

4.2 Auditing the Generated Reasoning

Every reasoning-consuming model depends on this reasoning, so we audited it. Three psychology-trained raters, who are the annotators (§3.2) and authors, independently scored every node of a 50-narrative sample (600 judgments; stratified by severity and enriched for rare high-severity cases) for analytical soundness (1–5) and applied three narrative-level flags: hallucination, situational mismatch, and over-diagnostic language (Appendix C). Mean soundness was $4.38/5$ ($\kappa_w = 0.93$); 91% of judgments were sound (≥ 4) and 0.5% unsound. By majority vote, 1 of the 50 narratives (2%) was flagged for hallucination and none for mismatch, but 13 (26%) for diagnostic or pathologizing language beyond what the narrative evidences. In a mental-health setting this is a content error, so we call ProcrastiCoT's output *theory-guided analytical* rather than diagnostic reasoning. Its likely cause is the generation prompt, which instructs the teacher to use clinical terminology; Appendix G gives a revised, untested prompt that ties each clinical term to a quoted narrative span. The audit is internal to the author team (see Limitations).

4.3 Generative Fine-Tuning (Decoder-only SFT)

We fine-tune a *Qwen2.5-7B-Instruct* model (Qwen Team, 2024) using QLoRA (Dettmers et al., 2023) in two settings (below), each trained under two supervision sources. With **teacher scores** (teacher-score distillation, our original configuration), the targets $\mathcal{Y}_{\text{scores}}$ are the teacher's own ratings; these differ from the human labels on 61.8% of narrative–dimension pairs and agree with them at only QWK 0.358, 0.427, 0.250, and 0.285 (five-fold means; Arousal, Avoidant, Decisional, Irrationality). The teacher's Irrationality agreement is far below GPT-5.4-mini's zero-shot 0.862 (Table 2): under the four-node prompt it rates only 16.2% of narratives as 1, against 71.5% of human labels. With **human scores**, the targets are the consensus labels every encoder regresses on, which makes this the same-supervision comparison; the CoT variant keeps the teacher's reasoning verbatim, except for removing a nested copy of the teacher's scores in 38 outputs (Appendix A), and replaces the target scores with the human labels. All models are evaluated against the human labels. The two settings are:

1. **Direct Prediction (No-CoT)**: The model is trained to directly output the JSON scores:

$$\mathcal{X}_{\text{narrative}} \xrightarrow{\text{SFT}} \mathcal{Y}_{\text{scores}} \tag{2}$$

2. **Reasoning Fine-Tuning (CoT)**: The model is trained to generate both the reasoning and the scores:

$$\mathcal{X}_{\text{narrative}} \xrightarrow{\text{SFT}} \mathcal{R}_{\text{ProcrastiCoT}} \oplus \mathcal{Y}_{\text{scores}} \tag{3}$$

4.4 Reasoning Injection via Concatenation

As an initial approach to inject CBT reasoning, we construct a concatenation variant (`+ ProcrastiCoT (Concat)`). We join the narrative $\mathcal{X}_{\text{narrative}}$ and the generated CBT reasoning $\mathcal{R}_{\text{ProcrastiCoT}}$ into one string, each preceded by a short field label and separated by a literal `[SEP]` (template in Appendix A), and tokenize it as a single input. All tokens therefore share one segment ID; the interaction between narrative and reasoning is modeled only by self-attention over the concatenated sequence. This approach is constrained by the encoder's 512-token limit, which truncates the tail of the reasoning for most narratives (§6.1).

4.5 Dual-Encoder Cross-Attention Architecture

To allow smaller models to use full-length reasoning while bypassing the 512-token context limitation of standard BERT encoders, we propose a **Dual-Encoder Cross-Attention** paradigm. Rather than concatenating the texts, we independently encode the narrative $\mathcal{X}_{\text{narrative}}$ and the generated CBT

reasoning $\mathcal{R}_{\text{ProcrastiCoT}}$ using a Siamese (shared-weight) encoder:

$$\begin{aligned} H_{\text{narr}} &= \text{BERT}(\mathcal{X}_{\text{narrative}}) \\ H_{\text{CBT}} &= \text{BERT}(\mathcal{R}_{\text{ProcrastiCoT}}) \end{aligned} \tag{4}$$

By decoupling the inputs, each branch has its own context window (up to 512 tokens), so the reasoning no longer competes with the narrative for a single budget; in our configuration (§5.1) only 2 of the 1,000 reasoning chains exceed their window.

To fuse the representations, we apply a Multi-Head Cross-Attention layer where the narrative acts as the Query (Q), dynamically attending to the CBT reasoning features represented as the Key and Value ($K = V$):

$$H_{\text{fuse}} = \text{LayerNorm}\Big(H_{\text{narr}} + \text{CrossAttn}(Q = H_{\text{narr}}, K = V = H_{\text{CBT}})\Big) \tag{5}$$

Padding tokens in the CBT reasoning branch are masked during cross-attention to prevent noise. The final prediction is read from the `[CLS]` position of the fused sequence H_{fuse} and passed through a dropout and linear projection head; module-level initialization and dimensions are given in Appendix A. The model is optimized via Mean Squared Error (MSE) loss against the human severity consensus:

$$\mathcal{L} = \frac{1}{4}\sum_{d=1}^{4} \text{MSE}(\hat{y}_d, y_d) \tag{6}$$

5 Experiments

5.1 Experimental Setup

We use stratified 5-fold cross-validation on the summed human severity (200 test narratives per fold; Appendix A). Encoders hold out 10% of each training fold (80 narratives, seed 42) for learning-rate and checkpoint selection; the generative models train on all 800 narratives for a fixed number of epochs. We report means over the five test folds.

Complete hyperparameters are in Appendix A. **Majority-Class (Dummy)** predicts the most frequent training severity. **TF-IDF + Ridge** uses *jieba*-segmented word n-grams with hyperparameters grid-searched on the training folds. **Zero-shot** and **few-shot LLMs** prompt *GPT-4o-mini* and *GPT-5.4-mini* with dimension definitions and 1–5 anchors to emit JSON ratings; the few-shot prompt adds three exemplars (none illustrating Decisional). GPT-4o-mini is evaluated on all 1,000 narratives and GPT-5.4-mini on one random 100-narrative subset (seed 42; 14–26 per fold), so its figures are not directly comparable to the other rows. **Generative fine-tuning** adapts *Qwen2.5-7B-Instruct* (Qwen Team, 2024) with 4-bit QLoRA (Dettmers et al., 2023) in direct-prediction (*SFT No-CoT*) and reasoning (*SFT CoT*) settings, each trained on the teacher's or the human scores (§4.3).

Lightweight encoders fine-tune *bert-base-chinese* and *CKIP-BERT* (Yang, 2020), each as a pure encoder and as a concatenation variant (`+ ProcrastiCoT (Concat)`). CKIP-BERT, pre-trained for Traditional Chinese, is our primary backbone, since *bert-base-chinese* is trained on Simplified-dominant text. Our **Dual-Encoder Cross-Attention** uses 8 heads at hidden size 768, with narratives truncated to 256 tokens (35 shortened) and reasoning to 512 (2 shortened); for 38 narratives the stored reasoning also carries a copy of the teacher's scores (Appendix A).

A **long-context baseline** fine-tunes *mmBERT-base* (Marone et al., 2025), a multilingual encoder built on ModernBERT (Warner et al., 2025) with an 8,192-token window, narrative-only and with the full reasoning concatenated at *zero truncation*; to our knowledge no ModernBERT-class encoder is dedicated to Traditional Chinese.

We evaluate against the human consensus labels with two ordinal metrics: **MAE** between rounded predictions and human scores, and **Quadratic Weighted Kappa** (QWK) (Cohen, 1968), computed over the full 1–5 label set, which penalizes ordinal errors quadratically (1 = perfect agreement, 0 = chance; guidelines in Cicchetti, 1994).

All training and evaluation ran on a single NVIDIA RTX 4090 (24GB): roughly 15 minutes per fold for the BERT-size encoders and 30 minutes per fold for the 7B QLoRA models.

5.2 Computational Overhead of Reasoning Injection

Reasoning injection requires a teacher LLM call before the encoder can predict. Assuming one narrative per uncached request, a call averages 2,915 input tokens and 334 reasoning tokens,[1] an estimated $\sim$4 seconds of decoding at an assumed 90 tokens/s (not a measured latency; time-to-first-token and network overhead are excluded); at GPT-5.4-mini rates ($0.75/$4.50 per million input/output tokens) this gives about $3.69 per 1,000 narratives,

excluding JSON overhead, with reasoning effort at the API default of none. These are estimates, not a measured API bill. The input is dominated by the fixed, cacheable system prompt (2,732 tokens, 94% of the input). Reasoning is generated *once per narrative* and stored, so for a fixed corpus this is a one-off preprocessing cost, and the predictor remains a BERT-size encoder. We did not benchmark end-to-end deployment on new text against generative baselines, so we make no end-to-end cost-advantage claim.

5.3 Main Quantitative Results and Thematic Discussion

Table 2 reports results on the held-out test folds.

5.3.1 Effect of Encoder Pre-training Corpus

Reasoning injection benefits differ by encoder. CKIP-BERT + ProcrastiCoT (Concat) achieves the top Avoidant QWK (0.742, +0.071 over its pure-encoder baseline), while the Dual-Encoder variant reaches the highest Arousal score among the Table 2 encoders (0.591 ± 0.082), though its margin over *bert-base-chinese* (0.564 ± 0.072) is not statistically significant ($p = 0.150$); §5.3.2 reports a long-context baseline that scores higher still. Conversely, reasoning injection degrades *bert-base-chinese* on Arousal ($0.564 \rightarrow 0.500$), possibly because the benefit depends on alignment between the teacher LLM's output register and the encoder's pre-training corpus.

5.3.2 Does a Long Context Window Suffice?

One might argue that the 512-token limit our architecture works around is an artifact of older encoders. Table 3A does not support this: with *zero* truncation, concatenation *degrades* mmBERT against its narrative-only baseline on every dimension, most sharply on Arousal ($0.605 \rightarrow 0.507$) and negligibly on Irrationality ($0.912 \rightarrow 0.909$), and our Dual-Encoder outperforms the long-context concatenation on Arousal in all four of its runs ($p \leq 0.001$) and on Decisional in one (§5.3.5). Fitting the reasoning into the window is not the bottleneck; exploiting it is. One untested reading is that a single `[CLS]` vector summarizing two texts of different register lets the reasoning dilute rather than inform it.

[1]Teacher-tokenizer counts: 2,732 prompt plus 183 narrative tokens in, 334 reasoning tokens out; §6.1 uses BERT-tokenizer counts (200 and 345), which the encoder's limit applies to.

mmBERT narrative-only, however, is a strong competitor: its Arousal score (0.605) lies within the Dual-Encoder's range across runs (0.591–0.632), its Avoidant score (0.741) is essentially the best (0.742), and the Dual-Encoder never beats it significantly, while the reverse exploratory test favors mmBERT on Avoidant in two runs ($p = 0.008$, 0.020) and on Decisional in one ($p = 0.042$; uncorrected). Against such an encoder, what the Dual-Encoder adds is the interpretable reasoning trace.

5.3.3 Decisional Dimension

On Decisional, the lexical baseline (TF-IDF + Ridge, QWK 0.319) matches or exceeds both concatenation variants (0.235 and 0.318), since decisional conflict often surfaces in explicit markers (e.g., "hesitate"). Our Dual-Encoder clears this ceiling on both backbones (0.353 on CKIP-BERT, 0.356 on *bert-base-chinese*), as do the human-supervised SFT No-CoT (0.346) and mmBERT narrative-only (0.352, Table 3). The margins are at most +0.037 QWK, and averaged over four runs the CKIP Dual-Encoder scores 0.316, level with the lexical baseline, which none of its runs beats significantly (§5.3.5). Decisional remains the weakest dimension, an open challenge (§6.2).

5.3.4 Generative Baselines: Calibration and Supervision

Prompted generative LLMs are poorly calibrated to the human scale (GPT-4o-mini's zero-shot MAE exceeds 1.0 on Arousal and Irrationality), and GPT-5.4-mini's zero-/few-shot differences are within fold-level noise on its 100-narrative subset (Appendix H). For the fine-tuned Qwen2.5-7B, the gap to the encoders was largely a supervision effect. Trained on teacher scores, SFT tracks the teacher's modest agreement with the humans (§4.3). Trained on human scores, SFT No-CoT reaches the highest Arousal QWK of any system (0.709; above mmBERT and every Dual-Encoder run, $p \leq 0.025$), is on par with the Dual-Encoder on Decisional (0.346 vs. 0.280–0.353 across its runs), and scores below every Dual-Encoder run on Avoidant and Irrationality (significantly in three of four on each). With about 7.6B parameters against the Dual-Encoder's 105M, it is strong but not dominant. With human-score targets, SFT CoT scores below SFT No-CoT on all four dimensions ($p \leq 0.008$), and its predictions lie closer to the teacher's ratings (QWK against the teacher vs. the humans: 0.600 vs. 0.324 on Avoidant, 0.414 vs. 0.274 on Decisional). The

Model	Arousal		Avoidant		Decisional		Irrationality	
	QWK	MAE	QWK	MAE	QWK	MAE	QWK	MAE
Baselines								
Majority-Class (Dummy)	0.000	0.302	0.000	0.862	0.000	0.415	0.000	0.832
TF-IDF + Ridge	0.396	0.291	0.474	0.847	0.319	0.320	0.743	0.535
GPT-4o-mini (Zero-Shot)	0.060	2.040	0.381	0.787	0.304	0.439	0.405	1.122
GPT-4o-mini (Few-Shot)	0.183	1.142	0.441	0.787	0.335	0.458	0.533	0.888
GPT-5.4-mini (Zero-Shot)	0.436	0.521	0.535	0.823	0.246	0.827	0.862	0.319
GPT-5.4-mini (Few-Shot)	0.484	0.391	0.545	0.804	0.215	0.891	0.799	0.364
Generative Fine-Tuning (Qwen2.5-7B)								
SFT No-CoT, teacher scores†	0.331	1.003	0.435	0.885	0.289	0.786	0.179	1.439
SFT CoT, teacher scores†	0.449	0.632	0.332	1.020	0.243	0.944	0.240	1.305
SFT No-CoT, human scores	**0.709**	0.231	0.664	**0.414**	0.346	0.328	0.881	**0.153**
SFT CoT, human scores	0.508	0.289	0.324	0.834	0.274	0.452	0.183	0.813
Google bert-base-chinese								
Pure Encoder (Tuned)	0.564	0.236	0.690	0.515	0.243	0.348	0.912	0.201
+ ProcrastiCoT (Concat)	0.500	0.250	0.686	0.518	0.235	0.384	0.903	0.201
Academia Sinica CKIP-BERT								
Pure Encoder (Tuned)	0.448	0.290	0.671	0.500	0.231	0.345	0.903	0.175
+ ProcrastiCoT (Concat)	0.587	0.229	**0.742**	0.437	0.318	0.316	0.891	0.185
Dual-Encoder Cross-Attention (Ours)								
Google bert-base-chinese	0.505	0.251	0.716	0.494	**0.356**	0.320	0.918	0.174
Academia Sinica CKIP-BERT	0.591	**0.218**	0.702	0.541	0.353	**0.308**	**0.920**	0.180

Table 2: QWK and MAE averaged over five-fold cross-validation (bold: best per column; per-fold SDs in Table 10). GPT-5.4-mini rows use a 100-narrative subset (§5.1). †Trained on the teacher's scores; all other trained rows use the human labels, against which every row is evaluated. One-sided paired bootstrap (1,000 iterations, pooled folds, uncorrected): the CKIP-BERT Dual-Encoder beats its pure encoder on all four dimensions ($p < 0.001$, 0.038, < 0.001, 0.018) and the concatenation variant only on Irrationality ($p = 0.006$; significant in three of four Dual-Encoder runs). Training variance is analyzed in §5.3.5. Non-significance does not imply equivalence.

	Arou.	Avoi.	Deci.	Irra.
A. Long-context control				
mmBERT, narrative only	**0.605**	**0.741**	0.352	0.912
mmBERT + reasoning	0.507	0.702	0.282	0.909
B. Fusion / context ablation				
no cross-attention	0.604	0.701	0.317	0.915
truncated reasoning	0.595	0.729	0.306	0.917
Full Dual-Encoder (CKIP)	0.591	0.702	**0.353**	**0.920**

Table 3: Controls (QWK, same five-fold protocol). **A**: mmBERT-base (8,192-token window), narrative only or with the complete reasoning concatenated without truncation. **B**: our CKIP-BERT Dual-Encoder without cross-attention, or with the reasoning cut to 256 tokens (below the ∼289 that concatenation retains on average). Discussed in §5.3.2 and §6.1.

two settings also differ in system prompt and output target (Appendix A), so the gap cannot be attributed to the reasoning alone; one possible explanation is a conflict between reasoning written to justify the teacher's scores and the human labels, and conditioning the reasoning on the labels is untested.

5.3.5 Run-to-Run Variance

Every system in Table 2 is a single training run, and our paired bootstrap resamples test narratives, so it captures test-set sampling but not training randomness (initialization, data order, dropout, and the per-fold learning-rate choice, which itself changes between runs). For our central comparison we trained the CKIP-BERT Dual-Encoder and its pure encoder three more times each, with identical inputs and fixed seeds (Table 4). The spread is largest on Decisional, where the Dual-Encoder ranges from 0.280 to 0.353; two of its runs even differ "significantly" under our bootstrap ($p = 0.018$). The Dual-Encoder nevertheless scores above the pure encoder on every dimension in every run, and across runs the gain is significant on Arousal ($+0.201$), Decisional ($+0.097$), and Irrationality ($+0.013$), but not on Avoidant ($+0.023$); with a Holm correction over the four dimensions, Arousal and Decisional remain significant ($p < 0.001$, 0.023) but Irrationality does not ($p = 0.086$). Differences between the Dual-Encoder and single-run systems smaller than its own range should be read as unresolved; §§5.3.2–5.3.4 note which comparisons

	Arou.	Avoi.	Deci.	Irra.
Dual-Encoder (CKIP)				
reported run	0.591	0.702	0.353	0.920
mean ± SD	.613±.017	.716±.015	.316±.038	.917±.008
range	.591–.632	.702–.737	.280–.353	.905–.922
Pure encoder (CKIP)				
reported run	0.448	0.671	0.231	0.903
mean ± SD	.412±.026	.693±.023	.220±.027	.904±.006
range	.388–.448	.671–.717	.185–.248	.896–.910
Welch p	$<.001$	.151	.008	.043

Table 4: QWK over four training runs per model with identical inputs and settings: the unseeded run reported in Table 2 and three re-runs with seeds 42–44. Welch's t-test compares the two models across runs (uncorrected).

hold in all four of its runs.

6 Discussion

6.1 Overcoming Context Truncation

Concatenation is bounded by BERT's 512 tokens. Under the template it actually uses (Appendix A), the input averages 565 tokens (max 812) and exceeds 512 for 86.7% of narratives on CKIP-BERT (85.8% on bert-base-chinese), truncating the tail of the reasoning, the terminal Arousal node first; about 289 reasoning tokens survive on average.[2] The Dual-Encoder gives narrative and reasoning separate windows (256 and 512 tokens) and encodes all but 2 of the 1,000 chains in full.

In the context-budget sensitivity analysis of Table 3B, cutting the reasoning branch to a fixed 256 tokens, somewhat below the average concatenation budget, leaves Arousal at 0.595 against 0.591 for the full model. We therefore see no sign that the Arousal gain depends on recovering truncated reasoning, consistent with §5.3.2, although a fixed cut does not match each narrative's budget.

Decisional shows the largest drop under both ablations (0.317 without cross-attention and 0.306 with truncated reasoning, against 0.353), but neither drop is significant ($p = 0.111$ and 0.088) and both lie within the full model's range across runs (0.280–0.353, §5.3.5), so we do not interpret them.

6.2 Error Analysis: The Decisional Bottleneck

Decisional is the hardest dimension (best single run 0.356; four-run CKIP-BERT mean 0.316). In an informal manual inspection of errors, choice paralysis was often masked by rationalized perfectionism (e.g., "I am gathering more data to ensure the best outcome"), which mimics diligent deliberation, so surface cues may not suffice; behavioral signals such as time-on-task may be needed.

6.3 Reasoning Node Ablation

Restricting the Dual-Encoder's reasoning input to a single CBT node (Table 9, Appendix F) lowers Arousal and Decisional under every node, while single nodes match the full chain on Irrationality (Node 4: 0.920) and Avoidant (Node 4: 0.707 vs. 0.702). These are single runs, with differences of the size run-to-run variation can produce (§5.3.5), so we read them only as suggesting that reasoning injection is not uniformly beneficial.

7 Conclusion and Future Work

We introduced **ProcrastiType-ZH**, the theory-informed **ProcrastiCoT** reasoning framework, and reasoning injection via concatenation and a **Dual-Encoder Cross-Attention** architecture. Injecting a teacher LLM's reasoning into BERT-base encoders gives the best single-run scores on three of the four dimensions across backbones (two on CKIP-BERT), at the cost of a one-off teacher call for a fixed corpus. Across four training runs the Dual-Encoder's gains over the pure encoder are significant (uncorrected) on Arousal, Decisional, and Irrationality but not Avoidant, its Decisional mean only matches a lexical baseline (§5.3.5), and a 7B model fine-tuned on the same labels, about 70 times larger, is stronger on Arousal but weaker on Irrationality and Avoidant; its variant that also generates the teacher's reasoning scores lower (§5.3.4). The gain does not appear to depend on recovering truncated context (§5.3.2, §6.1), a narrative-only long-context encoder remains competitive (§5.3.2), the gain varies across backbones and dimensions (§5.3.1, §5.3.5, §6.3), and *Decisional* remains weak in absolute terms (§6.2).

Future work will test reasoning conditioned on the human labels in generative fine-tuning, regenerate the reasoning under the revised prompt of Appendix G to test whether the 26% over-diagnostic flag rate falls without costing QWK, evaluate every system over multiple seeds, distill the teacher away at deployment, build a human-authored evaluation set, and engage licensed psychologists as external auditors.

[2] Narratives average 200 tokens (max 296) and reasoning 345 (max 537), excluding `[CLS]`/`[SEP]` (bert-base-chinese tokenizer); concatenated-input counts include them and the field labels.

Limitations

Our study has nine main limitations. First, the narratives are synthetic: persona-prompted LLM text, written as job-interview answers in a self-presentational register, may differ from real counseling transcripts or spontaneous self-reports and carries known biases relative to real populations (Li et al., 2025). Second, generative baselines are evaluated on text from the same model family (source-model bias); human consensus labels, which often diverge from the generation intent (§3.3), reduce but cannot rule out an advantage from distributional familiarity. Third, the labels come from joint consensus by three author-annotators, so no classical inter-annotator agreement is available for the main corpus (§3.2), and the test–retest check is an upper bound. Fourth, $N = 1{,}000$ is modest: with 200 test narratives per fold (14–26 for GPT-5.4-mini), fold estimates are noisy (Appendix H). Fifth, score 5 accounts for $\leq 0.1\%$ of labels, so the effective range is 1–4. Sixth, any new narrative needs a teacher-LLM call before prediction, adding an estimated few seconds and an API dependency at deployment (§5.2). Seventh, GPT-5.4-mini is both the reasoning teacher and a prompted baseline, so prompt-level familiarity effects cannot be excluded. Eighth, the reasoning audit (§4.2) was carried out by psychology-trained undergraduate students who are also the annotators and authors; it supports internal coherence and faithfulness to the narratives, not clinical appropriateness, and an external replication by licensed psychologists could find lower soundness or more flags. The 26% over-diagnostic rate shows that the reasoning can attribute unsupported pathology that downstream models may inherit; the revised prompt (Appendix G) is untested. Ninth, only the CKIP-BERT Dual-Encoder and pure encoder were trained more than once (§5.3.5); every other system is a single run, and since the Dual-Encoder's Decisional QWK spans 0.280–0.353 across runs, single-run differences of that size should not be read as established.

Ethical Considerations

Evaluating psychological traits from text raises ethical concerns. Our dataset, ProcrastiType-ZH, consists entirely of synthetic narratives generated via persona-based prompting and annotated by three of the authors. No real patients' private data or actual medical records were used, substantially reducing privacy risks. We did not seek institutional ethics review. The study involves no human participants and no real personal data: every narrative is synthetic, and the only people who read, annotated, or audited the data are the authors themselves, as part of their own research. We acknowledge that our pipeline involves multiple LLM-dependent stages: narrative generation (GPT-4o-mini), CBT reasoning generation (GPT-5.4-mini), and LLM-based baselines. While human consensus labels break the self-evaluation loop, the text and reasoning features remain LLM-generated, and we urge caution in generalizing our findings to authentic narratives authored by real patients. ProcrastiCoT and ProcrastiType-ZH are intended for research only. They have not been validated on real individuals and **should not be used to screen or diagnose people**, or to inform decisions about them such as hiring, admissions, or performance evaluation; because the narratives are framed as interview answers, use in hiring is a foreseeable misuse that we explicitly rule out. The corpus, rubric, and generated reasoning are shared on request for non-commercial research only, under terms that prohibit these uses.

Acknowledgements

This paper was partially supported by the National Science and Technology Council, Taiwan (Grant number: NSTC 115-2221-E-031-003).

We used OpenAI GPT-4o-mini for narrative generation (§3) and GPT-5.4-mini for CBT reasoning generation (§4). Generative AI tools assisted with language polishing, checking the correctness of code, and organizing data; all outputs were reviewed and validated by the authors, who take full responsibility for the content.

References

Aaron T Beck. 1976. *Cognitive therapy and the emotional disorders*. International Universities Press.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43.

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseph R Ferrari. 1992. Psychometric validation of two procrastination inventories for adults: Arousal and avoidance measures. *Journal of Psychopathology and Behavioral Assessment*, 14(2):97–110.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 39–48. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Leon Mann, Paul Burnett, Mark Radford, and Steve Ford. 1997. The Melbourne decision making questionnaire: An instrument for measuring patterns for coping with decisional conflict. *Journal of Behavioral Decision Making*, 10(1):1–19.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2026. GPT-5.4 mini. `https://developers.openai.com/docs/models/gpt-5.4-mini`. Accessed May 2026.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Fuschia Sirois and Timothy Pychyl. 2013. Procrastination and the priority of short-term mood regulation: Consequences for future self. *Social and Personality Psychology Compass*, 7(2):115–127.

Piers Steel. 2007. The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1):65–94.

Piers Steel. 2010. Arousal, avoidant and decisional procrastinators: Do they exist? *Personality and Individual Differences*, 48(8):926–934.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances*

in Neural Information Processing Systems, 35:24824–24837.

Mu Yang. 2020. CKIP transformers: Traditional Chinese NLP tools. https://github.com/ckiplab/ckip-transformers.

A Implementation Details

Encoders. All encoders (pure, concatenation, Dual-Encoder, and mmBERT) use AdamW for 10 epochs, keeping the checkpoint with the lowest validation loss (the patience of 10 never triggers early stopping), and a per-fold learning rate from $\{10^{-4}, 10^{-5}, 10^{-6}\}$ chosen by validation loss on the split of §5.1. Inputs and batch sizes: the pure encoder truncates the narrative to 256 tokens (35 narratives are longer), batch 16; the concatenation variant uses 512 tokens, batch 8; the Dual-Encoder uses 256 narrative plus 512 reasoning tokens (2 chains are longer), batch 8; mmBERT caps inputs at 1,024 tokens, batch 8, and its longest input is 600 tokens, so none is truncated. The concatenation input is "[Narrative text]: {narrative} `[SEP]` [Clinical-psychology analysis]: {reasoning}" (field labels in Chinese), tokenized as one sequence with all token-type IDs set to 0. The Dual-Encoder's shared encoders start from pre-trained weights, while the cross-attention layer, LayerNorm, and head are randomly initialized; the fused `[CLS]` vector passes through dropout ($p = 0.3$) and a linear layer to 4 outputs. The fusion ablations (Table 3B) use the same settings, with the reasoning cut to 256 tokens in the truncated variant.

Reasoning input and two data defects. Encoders receive the reasoning as the string form of the teacher's four-node object, node keys included. In 38 narratives (3.8%) that object also holds a copy of the teacher's scores. It sits at the end of the string, so the Dual-Encoder, its no-cross-attention ablation, and the mmBERT concatenation keep it, whereas 512-token concatenation truncates it for all but a few narratives and the 256-token ablation for all. Two further reasoning strings are malformed. We found both defects after the experiments and report the results unchanged. In three seed-matched Dual-Encoder pairs (seeds 42–44) with the copy kept or removed, predictions on the 38 narratives match the teacher's scores 40.4% vs. 40.1% of the time (37% elsewhere), and removal changes QWK by $+0.001$, $+0.010$, -0.020, and -0.002 on average (paired t-test across seeds, $p \geq 0.235$); the largest single change (-0.057 on Decisional, seed 42) is not repeated for seeds 43–44. We find no evidence that the copy affected our results.

Generative fine-tuning. *Qwen2.5-7B-Instruct* is adapted with 4-bit QLoRA (Dettmers et al., 2023) ($r = 16$, $\alpha = 32$, dropout 0.05, all linear projections `q/k/v/o/gate/up/down_proj`) for 3 epochs at learning rate 2×10^{-5}, effective batch size 8 (1×8 gradient accumulation), and a cosine schedule with 10% warmup, without early stopping. Teacher-score targets are the stored ProcrastiCoT output (the full JSON for SFT CoT, its `scores` field for SFT No-CoT); human-score targets replace the scores with the consensus labels and keep the reasoning verbatim, minus the teacher-score copy. Sequences are capped at 1,024 tokens, which truncates no training example, and the loss covers the assistant turn only. SFT CoT uses a shorter ProcrastiCoT system prompt (365 Qwen tokens) than the teacher (§5.2), and SFT No-CoT a direct-rating prompt (125 tokens), each identical at training and inference. Decoding is greedy with up to 1,024 new tokens; all 4,000 outputs across the four settings parsed into valid 1–5 scores.

Splits and lexical baseline. Folds come from `StratifiedKFold` (5 splits, shuffled, seed 42) on the summed human severity, clipped to 7–14 to merge sparse extremes. TF-IDF + Ridge uses *jieba*-segmented word n-grams and a separate Ridge regressor per dimension; the n-gram range (unigrams, or unigrams plus bigrams), vocabulary size ($\{3{,}000, 5{,}000, 10{,}000\}$), and α ($\{0.1, 1, 10\}$) are chosen per dimension and fold by 5-fold grid search on the training data (MSE), and predictions are rounded and clipped to 1–5.

B Generation Details

Persona-scenario matrix. The matrix crosses 8 personas (Anxious, Nagging, Observant, Rational, Efficient, Complaining, Casual, and Chatty) with 7 task contexts, each posed as a job-interview question: (1) recovering from severe academic delays, (2) executing high-stakes missions under expectations, (3) handling difficult projects, (4) completing tasks under extreme time pressure, (5) facing tedious and unchallenging daily tasks, (6) avoiding non-urgent daily chores, and (7) navigating ambiguous requirements or technical choices.

Node specifications. **Node 1, Emotion & Cognitive Defense** ($\mathcal{N}_{\text{emotion}}$; Beck's cognitive therapy) identifies cognitive distortions (e.g., catas-

trophizing, emotional reasoning) and short-term mood repair, i.e., delay as an emotional defense against anxiety, boredom, or fear of failure (Avoidant). **Node 2, Decisional Conflict** ($\mathcal{N}_{\text{decision}}$; MDMQ (Mann et al., 1997)) detects defensive avoidance or hypervigilance used to postpone responsibility for a possibly sub-optimal choice (Decisional). **Node 3, Temporal Discounting & Self-Regulation** ($\mathcal{N}_{\text{time}}$; TMT (Steel, 2007)) evaluates Expectancy, Value, and Sensitivity to Delay, i.e., hyperbolic discounting in which immediate low-value activities (e.g., social media) override long-term goals (Irrationality). **Node 4, Sensation Seeking & Self-Handicapping** ($\mathcal{N}_{\text{arousal}}$; Ferrari, 1992) tests whether delay is an intentional bid for arousal under deadline pressure and checks for self-handicapping, such as creating obstacles that excuse failure (Arousal).

C Annotation and Audit Rubrics

Severity anchors. **Score 1 (No Mention)**: absence of any procrastination indicators. **Score 2 (Passive Delay)**: a passive state without explicit dilatory intention or emotional distress. **Score 3 (Awareness & Distress)**: active awareness of delay accompanied by immediate remorse, or distraction under low pressure. **Score 4 (Compensatory Completion)**: active delay, but the task was successfully completed under high pressure. **Score 5 (Dysfunctional Failure)**: active delay resulting in final task failure, severe quality degradation, or negative external impact on others.

Reasoning audit. Raters scored each node for analytical soundness (5: fully supported by the narrative and consistent with the node's construct; 1: unsupported or misattributed) and applied three flags once per narrative: *hallucination* (facts absent from the narrative), *situational mismatch* (inconsistent with its core situation), and *over-diagnostic language* (pathologizing beyond the evidence). The 50 narratives come from 100 stratified draws (25 per dimension, across its severity levels) among the 998 with complete reasoning, reduced to 50 that keep every rare case (Arousal ≥ 3 or any score of 5); rare cases are thus enriched (Arousal ≥ 3: 20% vs. 6.3% of the corpus) and some levels are absent (e.g., Avoidant 2), so audit rates describe this sample. Flags use majority vote (any-rater rates: 2%, 0%, 32%); κ is averaged over nodes and rater pairs.

D Extended Dataset Statistics

Tables 5–7 give the statistics summarized in §3.3. In addition, Arousal correlates negatively with Avoidant ($r = -0.146$). The known groups use cut points Low ≤ 9, Medium 10–12, and High > 12 on the same summed ratings; separation is largest for Irrationality ($|r| = 0.913$) and weakest for Arousal ($|r| = 0.167$). Of the 400 narratives generated as High, 16.2% fell in the human Low band.

Dimension	Arou.	Avoi.	Deci.	Irra.
Arousal	1.000	-0.146***	-0.032	+0.094**
Avoidant	-0.159***	1.000	+0.221***	+0.030
Decisional	+0.032	+0.250***	1.000	+0.088**
Irrationality	+0.073*	+0.046	+0.121***	1.000

Table 5: Inter-dimensional correlation matrix. Upper triangle displays Pearson correlation r; lower triangle displays Spearman rank correlation ρ ($N = 1{,}000$). ***, **, and * indicate $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively.

Dimension	Low μ	Med μ	High μ	Δ	$\lvert r \rvert$
Arousal	1.170	1.306	1.612	+0.443	0.167***
Avoidant	2.282	3.647	3.891	+1.609	0.615***
Decisional	3.190	3.809	3.958	+0.768	0.744***
Irrationality	1.182	1.659	3.867	+2.685	0.913***

Table 6: Known-group discriminant validity across human severity groups: Low ($n = 401$), Medium ($n = 434$), and High ($n = 165$). Δ represents the mean difference from Low to High. $|r|$ is the rank-biserial correlation for the Low-vs-High contrast. *** indicates two-sided Mann-Whitney U significance at $p < 0.001$ for the Low-versus-High contrast.

	Human Level		
LLM Setup	**Low**	**Med**	**High**
Low ($n = 400$)	**64.5%**	31.8%	3.8%
Medium ($n = 200$)	39.0%	**51.5%**	9.5%
High ($n = 400$)	16.2%	51.0%	**32.8%**

Table 7: Human severity level (by summed consensus ratings) against the intended severity level used for generation, as row percentages.

E Qualitative Example

Table 8 shows one narrative (a translated excerpt), its consensus scores, and two of its four reasoning nodes. Its surface reads as diligence, yet the annotators rated Avoidant, Decisional, and Irrationality

at 4 and Arousal at 1, and Nodes 1–2 reformulate the delay as mood repair and decisional paralysis, which is the signal the encoder consumes. Node 1's "closed loop of emotional avoidance and cognitive rationalization" also goes beyond what the narrative evidences, an instance of the over-diagnostic language flagged in §4.2; the example illustrates the reasoning, not its faithfulness.

F Reasoning Node Ablation

Table 9 replaces the four-node reasoning input of the CKIP-BERT Dual-Encoder with a single node (the narrative stays the query, the node text is the key/value sequence), holding the splits fixed; entries are five-fold mean QWK $\pm$ SD. Decisional falls under every node ($0.353 \rightarrow 0.255$–0.327) and Arousal more weakly, while Node 4 matches the full chain on Irrationality and Avoidant. These are single runs, so the differences are descriptive (§5.3.5).

G Revised Generation Prompt (Diagnostic Language)

The original prompt instructs the teacher to "be sure to use professional clinical-psychology terminology", which §4.2 identifies as the likely cause of the over-diagnostic flags. The revision keeps the node specifications (Appendix B) but frames the task as text analysis rather than diagnosis, admits a clinical term only alongside a quoted supporting span, and requires an explicit statement of insufficient evidence where the narrative is silent. It is untested, and all results use the original prompt; it is available with the corpus on request.

H Per-Fold Standard Deviations

Table 10 gives the population SD across the five test folds for every cell of Table 2; it reflects variability across test folds, not training runs (§5.3.5). The GPT-5.4-mini rows use a 100-narrative subset (14–26 per fold) and are the least stable (Arousal QWK ± 0.224 and ± 0.249), so we do not interpret their zero-/few-shot differences.

Statistical tests. Single-run comparisons use a one-sided paired bootstrap over the pooled 1,000 test predictions; across-run comparisons use Welch's t-test (Table 4); the score-copy control uses a paired t-test across seeds (Appendix A); none is corrected for multiple comparisons. Because the bootstrap tests pooled QWK while the tables report five-fold means, counts of significant Dual-Encoder runs are approximate and can shift by one under fold-stratified resampling.

Narrative Example (Translated from Traditional Chinese)

"Honestly, when I was organizing work files recently, my expectations for myself were too high. I didn't want to just turn in anything, so I ended up overthinking and procrastinating for days. Every time I tried to decide how to handle it, I felt confused, constantly evaluating different options and their pros and cons, which ended up delaying the progress. Of course, my attention was also scattered by trivial things; for instance, scrolling on my phone distracted me, making tasks that should be finished quickly more and more delayed. Finally, after communicating with colleagues, I realized the importance of timely action, which made me determined to start, and I ultimately finished it before the deadline..."

Human Consensus Severity Scores

Arousal: 1 (No Mention) **Avoidant**: 4 (Compensatory Completion) **Decisional**: 4 (Compensatory Completion)
Irrationality: 4 (Compensatory Completion)

ProcrastiCoT CBT Reasoning (Nodes 1–2 of 4)

Node 1 (Emotion): Excessively high self-expectations and an unwillingness to hand in anything less than good produce clear perfectionistic anxiety, and delay then serves as *short-term mood repair*. Phone scrolling is a typical low-pressure substitute behavior: it offers immediate relief but deepens later guilt, forming a closed loop of emotional avoidance and cognitive rationalization.

Node 2 (Decision): Decisional delay and over-comparison co-occur: repeatedly evaluating how to proceed and weighing pros and cons indicates choice conflict and *decisional paralysis*. The defensive coping consists mainly of postponing the decision and turning to trivial tasks (*defensive avoidance*); only after talking with colleagues is the impasse broken.

Table 8: Qualitative example from ProcrastiType-ZH: a narrative excerpt, its human consensus scores, and two of the four CBT reasoning nodes generated by ProcrastiCoT (GPT-5.4-mini teacher), translated from Traditional Chinese.

Dimension	Full (4 Nodes)	Node 1 (Emo)	Node 2 (Dec)	Node 3 (Time)	Node 4 (Arousal)
Arousal	**0.591**$_{\pm 0.082}$	0.568$_{\pm 0.108}$	0.580$_{\pm 0.142}$	0.558$_{\pm 0.145}$	0.570$_{\pm 0.192}$
Avoidant	0.702$_{\pm 0.039}$	0.667$_{\pm 0.014}$	0.705$_{\pm 0.026}$	0.700$_{\pm 0.026}$	**0.707**$_{\pm 0.035}$
Decisional	**0.353**$_{\pm 0.052}$	0.263$_{\pm 0.049}$	0.294$_{\pm 0.042}$	0.327$_{\pm 0.105}$	0.255$_{\pm 0.115}$
Irrationality	**0.920**$_{\pm 0.015}$	0.910$_{\pm 0.034}$	0.905$_{\pm 0.013}$	0.902$_{\pm 0.034}$	**0.920**$_{\pm 0.012}$

Table 9: Reasoning Node Ablation (QWK) on the Dual-Encoder CKIP-BERT architecture. Each variant provides only a single CBT node as the reasoning input (the key/value sequence).

	Arousal		Avoidant		Decisional		Irrationality	
Model	**QWK**	**MAE**	**QWK**	**MAE**	**QWK**	**MAE**	**QWK**	**MAE**
Baselines								
Majority-Class (Dummy)	0.000	0.034	0.000	0.048	0.000	0.019	0.000	0.045
TF-IDF + Ridge	0.054	0.013	0.026	0.019	0.080	0.039	0.026	0.019
GPT-4o-mini (Zero-Shot)	0.008	0.026	0.014	0.024	0.027	0.024	0.021	0.024
GPT-4o-mini (Few-Shot)	0.021	0.038	0.016	0.026	0.061	0.030	0.026	0.029
GPT-5.4-mini (Zero-Shot)	0.224	0.135	0.103	0.136	0.116	0.140	0.058	0.090
GPT-5.4-mini (Few-Shot)	0.249	0.069	0.102	0.157	0.113	0.154	0.099	0.142
Generative Fine-Tuning (Qwen2.5-7B)								
SFT No-CoT, teacher scores	0.045	0.065	0.019	0.017	0.037	0.037	0.019	0.045
SFT CoT, teacher scores	0.057	0.054	0.049	0.060	0.026	0.043	0.024	0.045
SFT No-CoT, human scores	0.043	0.023	0.056	0.051	0.050	0.026	0.019	0.020
SFT CoT, human scores	0.075	0.038	0.022	0.025	0.065	0.024	0.041	0.056
Google bert-base-chinese								
Pure Encoder (Tuned)	0.072	0.026	0.046	0.027	0.020	0.025	0.020	0.048
+ ProcrastiCoT (Concat)	0.058	0.019	0.029	0.049	0.031	0.031	0.033	0.051
Academia Sinica CKIP-BERT								
Pure Encoder (Tuned)	0.123	0.097	0.060	0.056	0.068	0.028	0.026	0.017
+ ProcrastiCoT (Concat)	0.116	0.015	0.035	0.040	0.078	0.029	0.050	0.057
Dual-Encoder Cross-Attention (Ours)								
Google bert-base-chinese	0.202	0.053	0.040	0.041	0.099	0.065	0.023	0.022
Academia Sinica CKIP-BERT	0.082	0.019	0.039	0.055	0.052	0.028	0.015	0.029

Table 10: Standard deviation across the five cross-validation test folds for each cell of Table 2, in identical row and column order. GPT-5.4-mini rows come from a 100-narrative subset (14–26 narratives per fold) and are not comparable to the full-corpus rows.