Growth Above All: Goals and Beneficiaries of LLM Silver Linings in Thai and US Settings

Pachaya Sailamul
National Electronics and Computer Technology Center (NECTEC), NSTDA
Pathum Thani, Thailand
pachaya.sai@nectec.or.th

Sirirat Ularntinon
Queen Sirikit National Institute of Child Health
Bangkok, Thailand
siriratul@gmail.com

Nutchanon Yongsatianchot
Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
nutchanon.yon@mahidol.ac.th

Abstract

Large language models (LLMs) could widen access to mental health support, but support should also suit the person's culture. This work asks what LLMs offer in the silver-linings exercise of Positive Affect Treatment (PAT), where a person lists positive aspects of an event. Following appraisal theory, we treat each silver lining as a claim that the event serves a goal. We wrote 18 high-school events in four versions, set in the United States or Thailand and in English or Thai. Two widely used LLMs gave six silver linings for each version, and two LLM coders coded which goal each silver lining serves and who gains from it, across all 864 lists of silver linings. Growth, what the student learned or can now do, was the most frequent goal in every condition and for both models. Without the country named, the share of linings that benefit others or the group changed by under three percentage points with language, setting, or their interaction. Naming the country raised the setting contrast by under five points. These results suggest little sensitivity to cultural cues, so a system that uses these models to guide the exercise cannot assume its suggestions fit the person asking.

1 Introduction

Mental health problems affect people in every country and culture. In 2019, about one in eight people worldwide lived with a mental disorder, yet even in high-income countries only one third of people with depression received formal care (World Health Organization, 2022). In Thailand, about one in seven secondary-school students aged 13 to 17 had a current psychiatric disorder (Ekasawin et al., 2016), and only 11.5% of Thai adults with a mental health or substance use problem received any care within a year (Kittirattanapaiboon et al., 2017). Large language models (LLMs) could help to close this gap, because they can hold a supportive conversation with anyone, at any time, and in the person's own language. In randomized trials and meta-analyses, conversational agents reduced symptoms of depression, anxiety, and distress, although much of this evidence concerns earlier systems with scripted responses (Fitzpatrick et al., 2017; He et al., 2023; Li et al., 2023; Zhong et al., 2024; Heinz et al., 2025). In the United States, about one in eight adolescents and young adults already use generative AI for mental health advice (McBain et al., 2025). However, good support should also suit the person who receives it, including their goals, values, and cultural background. We therefore need to ask not only whether a model can deliver a therapeutic exercise, but also whether what it delivers fits the person who asked.

Many therapeutic exercises teach people new ways to cope with stress. Coping theory distinguishes two broad kinds (Lazarus and Folkman, 1984). In problem-focused coping, a person acts to change the situation. In emotion-focused coping, a person changes how they perceive it by changing their beliefs or goals about it. For example, a student with a low grade may cope by changing how they study, or by deciding that what they learned matters more than the grade. The second option is also called reappraisal, and positive reappraisal is the form that looks for what is good in a stressful situation (Folkman, 1997). Positive Affect Treatment (PAT) teaches positive reappraisal among other practices that build positive emotion (Craske et al., 2022). In randomized trials against a treatment that targeted negative affect, PAT produced larger gains in positive affect (Craske et al., 2019) and in overall clinical status (Craske et al., 2023; Meuret et al., 2026).

In this work, we focus on the silver-linings exercise in the PAT workbook (Meuret et al., 2022, pp. 68–73). The exercise asks a person to choose a situation, which may be positive, neutral, or negative, and to write down at least six silver linings of

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 80–104
November 9, 2026 ©2026 Association for Computational Linguistics

it. The workbook defines a silver lining as any positive aspect of a situation, however small or silly. However, it does not say what makes one silver lining better than another, or how a silver lining differs from reassurance or advice.

LLMs could act as a coach who guides people through exercises like this one. Language models already help people reframe negative thoughts in deployed tools (Sharma et al., 2023, 2024). When a person asks a model for silver linings, the model has to decide which positive aspects to offer, because the workbook leaves this choice open. To study these choices, we draw on appraisal theory (Lazarus, 1991). In this theory, a person's emotion depends on whether they appraise an event as helping or hindering their goals. We therefore define a silver lining as a claim that some aspect of an event serves a goal. This definition gives each silver lining two properties that we can code, namely which goal it serves and who benefits.

The goals that a silver lining serves may differ across cultures. Markus and Kitayama (1991) distinguished an independent view of the self, which emphasizes personal attributes and goals, from an interdependent view, which emphasizes relationships and the group. For instance, "you proved you can do it on your own" may be a welcome silver lining in one setting, while in another setting the good in an event may be that the group stayed together. Thailand is a useful case for this question, because Thai accounts of coping and social life stress acceptance (thum-jai) and consideration for others (kreng-jai), as Section 2 describes. However, the emotional output of multilingual LLMs reflects Western norms even in other languages (Havaldar et al., 2023), so a model may not adjust its silver linings to the person's culture on its own. To our knowledge, no LLM study has examined what positive reappraisals are about in a language other than English.

This work addresses this gap by testing whether the silver linings that LLMs offer a high-school student change with cues to the student's culture. We wrote 18 high-school events across the social, academic, and sport domains, which are common events that students may need to cope with, such as trying out for the basketball team, getting a midterm grade, or being invited out by friends. Each event is set in both the United States and Thailand and written in both English and Thai. These four versions cross the setting of the event with the language of the request. The student never states their nationality, so the language and the setting are implicit cues to the student's culture. In a second condition, the student also states the country, as in "I'm in high school in Thailand", which gives an explicit cue. We asked two widely used models for six positive aspects of each version under three wordings, which produced 864 lists of suggested silver linings. GPT-5.6 Luna was the default model for free ChatGPT users when we generated the lists (OpenAI, 2026), and DeepSeek V4.1 Flash is the newest model from DeepSeek (DeepSeek, 2026). Two LLM coders from different model families coded every item, and we read the cues against the share of silver linings that benefit another person or the group, which we call the other-benefit share. We ask two research questions.

- **RQ1.** Do the goals and beneficiaries of the models' silver linings change with the language of the request, the setting of the event, or their interaction, and do they change in the direction that cross-cultural research on people suggests (Section 2)?
- **RQ2.** How does the change under these implicit cues compare with the change when the student names the country?

Growth was the most frequent goal in all four conditions and for both models, and we found no evidence of a shift toward silver linings that benefit others or the group, or toward acceptance, under these cues. Without the country named, the contrasts in other-benefit share were within three percentage points of zero under both coders, and their intervals ruled out a ten-point shift, which we treat as the smallest meaningful shift. Naming the country raised the setting contrast by about four and a half points, while the valence of the event moved the same measure by about seventeen. In a check on known-answer items, both LLM coders gave the intended codes for at least 142 of 144, including every item with Thai cultural content.

Our work makes three contributions. First, we provide 18 matched high-school events in four combinations of setting and language, with a check that the setting is identifiable. Second, we propose a coding scheme that asks whether a generated item names a gain at all before coding its goal and beneficiary. Finally, we report which goal each silver lining serves and who gains from it, and how little this changes with language, setting, or an explicit cue.

2 Related Work

Positive reappraisal and silver linings. Positive reappraisal means seeing a stressful or negative event in a more positive light, and in a laboratory study it produced larger gains in positive affect (McRae et al., 2012). Exercises that ask people to find the good in an event do help. For instance, in a three-week online program, participants wrote three bright sides of a frustrating event each day alongside a goal-visualization exercise, and pessimistic participants reported fewer depressive symptoms afterward (Sergeant and Mongrain, 2014). Research on benefit finding also groups what people report into domains, such as relating to others or spiritual change (Tedeschi and Calhoun, 1996; Sodergren and Hyland, 2000), but it asks a person to report the benefits of an adversity they lived through (Sears et al., 2003; Helgeson et al., 2006). We instead ask what a model offers a person it knows almost nothing about, and we read that offer through appraisal theory. Our definition of a silver lining follows computational models of coping, in which emotion-focused coping changes a person's beliefs about an event or the goals against which they evaluate it (Marsella and Gratch, 2009; Yongsatianchot and Marsella, 2022, 2023). Uusberg et al. (2019) call a reappraisal that changes the goals repurposing, because the goals used to judge the situation change while its facts stay the same.

Culture and coping. Western and East Asian samples did not differ significantly in how often they reported using reappraisal (Song et al., 2024), but emotion regulation tends to be more adaptive when it fits its cultural context; for instance, suppressing emotions has gone with worse well-being in independent cultures but often not in interdependent ones (Ford and Mauss, 2015). Asian models of the self differ: across 55 cultural groups, the Thai sample stood near the interdependent end on connection to others but near the independent end on self-reliance (Vignoles et al., 2016). Thai college students report a more interdependent self than American students (Christopher et al., 2010), national Thai surveys identified grateful relationship as one of nine value orientations (Komin, 1990a), and Thai adolescents describe a successful life as a way to repay their parents (Balthip et al., 2022). Kreng-jai (เกรงใจ), a reluctance to impose on others, is an interactional value in Thai culture (Komin, 1990b) with a close counterpart, arr-nar, among Burmese and Karen communities (Khirikoekkong et al., 2023). Thum-jai (ทำใจ), accepting and letting go of what cannot be changed, is a Thai coping strategy (Mills et al., 2017, 2019) often understood through Buddhist ideas of impermanence and karma (Cassaniti, 2015). This literature motivates predictions for Thai cues: more silver linings that benefit another person or the group, more relational, community, and transcendent goals, and more consolations that invoke acceptance or fate. These are cultural tendencies, not the preferences of every Thai student.

Culture in LLMs. ChatGPT's responses to a cultural values survey align most closely with those of American respondents (Cao et al., 2023), and prompting in the language of another culture improves alignment with that culture only partly (AlKhamissi et al., 2024). The closest work to ours evaluates the cultural sensitivity of emotional support from LLMs, but not the content of reappraisals (Liu et al., 2026).

LLMs, reappraisal, and appraisal theory. Ziems et al. (2022) annotated positive reframes with six strategies, including growth mindset, impermanence, and optimism. Sharma et al. (2023) found that people preferred empathic and specific reframes to overly positive ones, and raters judged reappraisals from guided LLMs and from GPT-4 above human-written ones (Zhan et al., 2024; Herderich et al., 2026). A related line asks whether LLMs reason about emotion as appraisal theory predicts. LLM responses followed human dynamics of appraisal and coping but did not vary along key appraisal dimensions (Yongsatianchot et al., 2023), and GPT matched human appraisals but had difficulty predicting coping responses when situations were varied (Tak and Gratch, 2023). LLMs that estimated the chance of an outcome a role-played character wanted showed little wishful thinking, which some models showed only when the prompt stated that the character felt hopeful (Yongsatianchot and Marsella, 2025). Their emotion inferences followed appraisal dimensions in most domains, although surface features overrode them in one domain and only for some models (Thejll-Madsen et al., 2026). These studies focus on the emotional reaction, namely how good a reframe is or how well a model infers what a person feels. We focus instead on the content of the reappraisal, the goal it invokes and the person it names as gaining.

LLMs as coders. LLMs can annotate text at or

above the level of crowd workers on several tasks (Gilardi et al., 2023), but across 25 social-science benchmarks they reached only fair agreement with human labels on the classification tasks (Ziems et al., 2024). Model judges also show position, verbosity, and self-enhancement biases (Zheng et al., 2023; Panickssery et al., 2024), and a model can judge the same content inconsistently across languages (Fu and Liu, 2025), so we check our coders on known answers and report their agreement separately for each language.

3 Method

3.1 Design

We used a 2 (setting: US, Thai) × 2 (language: English, Thai) × 3 (domain: social, sport, academic) × 3 (valence: positive, neutral, negative) × 2 (model) × 3 (instruction wording) design in two persona arms. The stimuli are 18 schemas, two for each combination of domain and valence, and they form six story families of a positive, a negative, and a neutral schema from the same domain. A schema describes one everyday high-school event, and we wrote each schema as four texts that tell the same story. The four texts cross the setting of the event with the language of the text, and we call these four combinations cells.

In the two diagonal cells the language matches the setting, and the two off-diagonal cells pair Thai text with a US event and English text with a Thai event, so the language cue and the setting cue separate only there. Because every schema appears in all four cells, the two contrasts are paired within schema and are not confounded with the particulars of any one event.

The asker is a high-school student who speaks in the first person (I / เรา). In the implicit arm, which is the main arm, the persona line is "I'm in high school." ("เราเรียนอยู่ม.ปลาย"), and only situational detail carries the setting. In the explicit arm, the persona line also names the country and nothing else changes, which tests whether a direct cue produces a larger change than the implicit cues.

3.2 Stimuli

The positive and negative schemas of a story family differ in exactly one clause, which states the outcome or how it compared with expectation. Each neutral schema describes a routine logistics event without a positive or negative outcome, such as a gear pick-up or a form handed in. No text says why the event mattered to the student, because the goals are what the model supplies and what we measure. The setting enters each text mainly through one or two nouns and the local details they require. For instance, the US-setting texts mention AMC, Venmo, and Lincoln High where the Thai-setting texts mention Major Cineplex, PromptPay, and Suksawat Wittaya.

We wrote the Thai-setting texts natively in Thai rather than translating them from English, and a second native speaker revised all Thai text. Before generation, we checked that both models could name the intended country of every text, and after one revision every text passed. Appendix A gives the details and all 18 schemas.

3.3 Generation

Each prompt was a single user turn with no system prompt, and it contained the persona line, the situation, and the instruction. The basketball-tryout schema under the first wording gives this prompt in the US-setting English cell: "I'm in high school. I've been going to open gym all summer to get ready for basketball tryouts at Lincoln High. Today the coach posted the roster, and my name wasn't on it. Please help me think of 6 good things about this event". Its Thai-setting Thai twin reads "เราเรียนอยู่ม.ปลาย เราซ้อมบาสมาทั้งปิดเทอมเพื่อไปคัดตัวทีมบาสโรงเรียนสุขสวัสดิ์วิทยา วันนี้โค้ชติดประกาศรายชื่อ ไม่มีชื่อเรา ช่วยคิดหน่อยว่าเรื่องนี้มีสิ่งดีๆ อะไรบ้าง สัก 6 อย่าง". The instruction came from one of three wording pairs, each an English–Thai twin, which ask for "6 good things about this event" (P1), "6 good sides to this event" (P2), or "6 positive ways to look at this event" (P3). Table 3 in Appendix B gives the full strings and the Thai twins. We crossed the wording with the other factors because a single wording per language would confound language with phrasing.

Two widely used models generated the lists. We called GPT-5.6 Luna through the OpenAI API and DeepSeek V4.1 Flash through OpenRouter, both at temperature 1 with reasoning disabled. One text with one model in one arm is a setup, which gives 144 setups per arm. Each setup received three requests, one per wording pair.

We call one model response to one situation a suggestion list, or list for short, and each list should contain six suggested silver linings, which we call items. The two arms therefore give 864 lists and 5,184 requested items, with 432 lists per arm. We kept and flagged responses with the

wrong number of items, the wrong language, a refusal, or meta-commentary, and we did not resample them. Appendix B gives the persona lines, settings, and parsing rules.

3.4 Coding scheme

The first coding question asks what an item gives the asker, and coders answer it with an ordered decision in which the first yes decides. First, an item that is off-task, such as a story angle, a refusal, or a general tip with no link to the event, is not a lining. Second, an item that tells the asker to take an outward step, such as asking the coach for feedback or trying again, is an action. Third, an item that names a gain in the event itself is a lining. The gain may be something the asker or the people around them got, kept, avoided, or learned, a good result of the event, or something they can now do. Fourth, an item that only changes how the asker should see or feel about the event, without naming anything good that the event brought, is a consolation, such as "it could have been worse" or "it is not your fault". A consolation eases the loss, whereas a lining points to a gain. Any other item, such as a retelling of the event, is not a lining. An item that opens with "at least" or says that the chance is not lost is a consolation even when it names a gain (Appendix C.2). For instance, "a chance to make new friends" is a lining even though the asker would have to act on it, while "join a club" is an action.

Only linings receive two further questions. The first asks which goal the gain serves. Coders code the gain that the lead clause names, and when it names two, they take the first that applies in a fixed order of seven goals. Transcendent comes first (meaning beyond the event, such as karma or fate), followed by community (service to or pride for a collective beyond the asker's own relationships, such as the school). Relational comes next (a connection between people, including repaying one's parents and not imposing on friends, family, or the group, a concern Thai speakers call kreng-jai). The four goals that concern the student come last: growth (the student got better or did something new), achievement (a good result or recognition that the item names), wellbeing (a feeling, or a mind spared worry or hassle), and resources (time, money, or things in order). The goal records the kind of gain for whoever gains it, except that what family or friends feel because of the asker is relational. The second question asks who benefits, and it has five options: the asker, a specific other person, the group, a transcendent order, or unspecified. Consolations receive a type instead, such as acceptance, impermanence, or reassurance. Models often wrote each item as a bold header followed by an elaboration. Coders therefore anchored each judgment in the header, or in the first sentence when an item had no header, and read the rest only to settle what the header left open.

The other-benefit share of a list is the share of its linings that benefit a specific other person or the group rather than the asker, and every goal share is likewise a share of the linings. We read the cues against the other-benefit share, because interdependent views of the self emphasize benefits to other people and to the group, and we report the goal and item-type distributions alongside it.

Coders saw the situation and the items but never the model, the cell, or the list identifier. Every judgment is categorical, so cross-cultural differences in the use of rating scales cannot enter the codes. Table 5 in Appendix C lists all coding questions, and Appendix C gives the full definitions and decision rules.

3.5 Coders and checks on them

Two LLM coders coded all 864 lists: GPT-6.1 Sol through the OpenAI API and Claude Opus 5.5 through the Anthropic API, both at medium reasoning effort. Each coded zero-shot from the codebook, one list per call, with the same prompt. Claude Opus 5.5 belongs to a model family that generated none of the lists. Two bilingual Thai coders (Coder A and Coder B), who are authors of this paper, took part in developing the codebook on a separate set of 32 lists from the implicit arm. Because some of the codebook's examples come from these lists, we do not use them to check the coders. Instead, we check the LLM coders in three ways.

First, we wrote a known-answer set of 72 items, six for each of 12 schemas in the Thai setting. Each item has an intended item type and, for a lining, an intended goal and beneficiary, or for a consolation, an intended type. Every item has a Thai and an English version, which the coders saw as separate lists under the Thai-language and the English-language situation of its schema. Twelve items carry Thai cultural content, such as kreng-jai, face for the family, repaying one's parents, karma, and the Buddhist view that nothing is permanent. Both human coders checked every item

and its intended codes. We wrote these items together with the codebook, so they show whether a coder applies its definitions, including to Thai cultural content, but not whether the definitions are right.

Second, the two human coders coded 36 held-out lists that played no part in developing the codebook. The lists cover all four cells of nine schemas from the implicit arm, and they include both generating models and all three wordings. Each human coder worked alone, in a list order shuffled for that coder, without seeing the cell, the model, the LLM codes, or the other human coder's codes. We report their codes as they are, without reconciliation. We compare each LLM coder with each human coder, and the two human coders with each other, per question, per language, and per generating model, as percent agreement and Krippendorff's α. Third, we compare the two LLM coders on all 5,190 items.

3.6 Analysis

The unit of analysis is the list. We averaged other-benefit share within each setup over the three wordings and contrasted the setup means within schema and model. The language contrast is the Thai text minus the English text of the same event, averaged over the two settings, and the setting contrast is the Thai event minus the US event, averaged over the two languages. The interaction is the language contrast in the Thai setting minus the language contrast in the US setting, which is a difference of differences. Each contrast is in percentage points with equal weight per schema, and its 95% interval comes from a percentile bootstrap over lists within setups.

The smallest shift of interest is 10 points of other-benefit share. We read a contrast as a shift when its interval excludes zero and its estimate is at least 10 points, as no shift of practical interest when its interval lies inside ± 10 points, and as inconclusive otherwise. A list with no lining has no other-benefit share, so it drops out of these contrasts, although it still counts in the item-type distribution. We computed the same contrasts in the explicit arm and the change between the arms, and treat every other reported comparison as exploratory.

We computed every contrast under each LLM coder. We also pooled the two coders. The pooled estimate is the mean of the two coders' estimates. For its interval, each bootstrap draw resamples the lists within each setup, as above, and then computes the contrast from the codes of one of the two coders, picked at random, so the interval reflects both sampling variability and the disagreement between the coders. A contrast reads as a shift, or as no shift of practical interest, only when that reading holds under each coder and under the pooled interval, and otherwise it reads as inconclusive.

4 Results

4.1 What the models produce

All 864 calls returned a suggestion list, and both LLM coders coded all 5,190 items. GPT-6.1 Sol found no lining in 12 lists and Claude Opus 5.5 in 7. Most of these lists (9 and 5) came from DeepSeek V4.1 Flash answering the third Thai wording, which asks for six good มุม (angles): DeepSeek listed six angles from which to tell the story rather than six positives. The rest came from GPT-5.6 Luna under the second English wording. These lists count in the item-type results but not in the contrasts.

Averaged over the two coders, the items were 91% linings, 4% consolations, 3% not a lining, and 2% actions; Figure 5 in Appendix D splits the implicit arm by cell and model. Consolations came mostly from negative events, where they made up 12 to 13% of items. Most were reassurance (62%) or impermanence, such as "there is still time", and each coder found only two acceptance items and at most one that invoked fate or meaning.

Figure 1 gives the goal profile of the implicit arm. Growth, which covers how the student got better or did something new, was the most frequent goal in every cell and for both models under both coders, at 36 to 55% of linings (Figure 4 in Appendix D). Relational goals came second everywhere, at 19 to 22%, followed by wellbeing at 12 to 19, resources at 9 to 15, and achievement at 3 to 8. Community goals never exceeded 4% of linings, and each coder used the transcendent category (karma, fate, spiritual meaning) for a single lining in all 5,190 items.

The other-benefit share ranged from 15 to 24% over the sixteen combinations of cell, model, and arm (Tables 6 and 7), and almost all of those linings benefited the group. Under each coder, only 8 or 9 of the roughly 2,400 linings in the implicit arm named a specific other person.

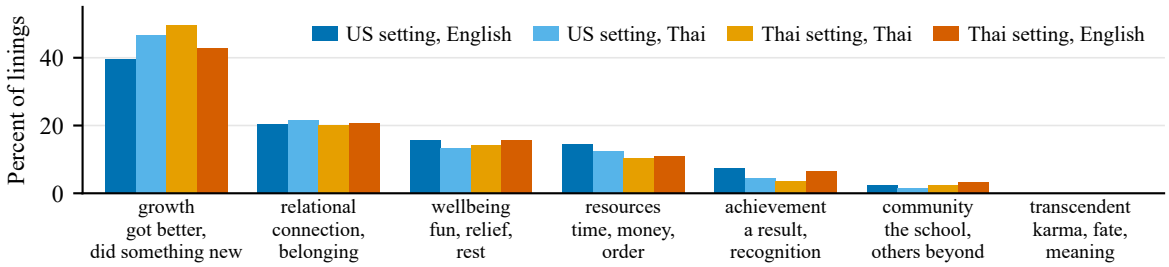


Figure 1: Goal categories of the linings in the implicit arm, by setting and language of the text, with both models pooled and the two coders averaged. The line under each goal names what the category covers. Figure 4 in Appendix D splits the same distribution by model.

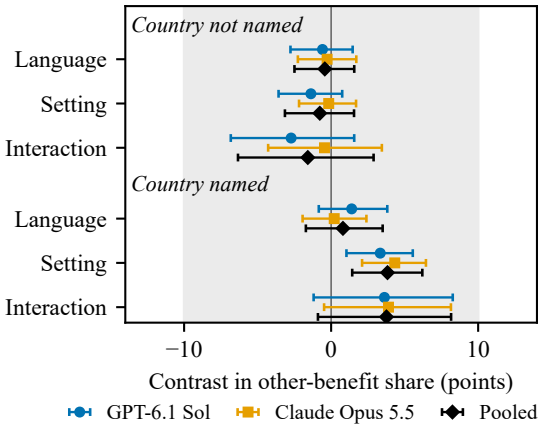


Figure 2: Contrasts in other-benefit share with 95% bootstrap intervals under each coder and pooled over the two coders, with both models together, in the implicit arm (top, the country is not named) and the explicit arm (bottom). Section 3.6 defines the three contrasts and the pooled interval. The shaded band marks the smallest shift of interest of ten points.

4.2 RQ1: Language, setting, and their interaction

Figure 2 gives the three contrasts in each arm under each coder, and Table 8 lists the implicit-arm values for each model. In the implicit arm, every interval lay inside the ±10-point band under both coders and when pooled, so none of the three contrasts reached a shift of practical interest. The pooled estimates were −0.4 points for language, −0.8 for setting, and −1.6 for the interaction, and no estimate for either model or coder was further than 3.1 points from zero. Thai cues did not raise the other-benefit share above US cues. The language and the setting estimates were slightly negative under both coders, which is the opposite of what Section 2 leads us to expect. With both models together, this reading held in every sensitivity analysis in Appendix D.5.

The relational and community share and the consolation rate did not move either, with every estimate within 2.3 points of zero. On the goals, however, Thai-language requests drew about six to eight more growth linings per hundred: +6.1 points under GPT-6.1 Sol (+3.4 to +8.9) and +7.6 under Claude Opus 5.5 (+4.7 to +10.3). The gain came from achievement, wellbeing, and resources, and it was larger in GPT-5.6 Luna's lists (+8.6 and +10.8 points) than in DeepSeek's (+3.7 and +4.4). The shift therefore stays among the goals that concern the student and does not move toward relational goals. Thai-language requests also drew slightly fewer actions (−3.1 and −1.3 points). We report these goal and action contrasts as exploratory.

The language contrast also depended on the wording: under the two coders it was −8.3 and −6.9 points for P2 (sides) and +6.2 and +5.8 for P3 (angles), so the phrasing of the request moved the estimate more than its language did (Appendix D.4).

4.3 RQ2: Explicit against implicit cues

In the explicit arm, where the persona also named the country, the contrasts in other-benefit share were larger but still small. Pooled over the coders, the contrasts were +0.8 points for language (−1.7 to +3.5), +3.8 points for setting (+1.4 to +6.2), and +3.8 points for the interaction (−0.9 to +8.2). All three readings are again no shift of practical interest under both coders. For DeepSeek alone, the interaction interval reaches past the band under both coders, so that reading is inconclusive (Appendix D.7).

The change between the arms is RQ2. Nam-

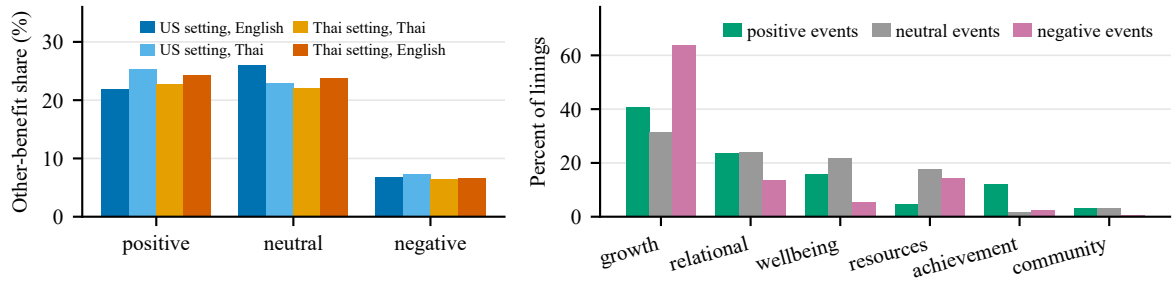


Figure 3: What moved the offer in the implicit arm. Left: other-benefit share by the valence of the event, with the four cells inside each valence. Right: the goal of the lining by valence. Both panels pool the two models and average the two coders. Figure 6 in Appendix D gives the goals by valence and cell.

ing the country raised the setting contrast by 4.6 points pooled (+1.6 to +7.8), by 4.7 under GPT-6.1 Sol and 4.5 under Claude Opus 5.5, and by a similar amount for the two generating models (4.4 for GPT and 4.8 for DeepSeek, pooled). This interval treats our 18 schemas as fixed. A bootstrap that also resamples the schemas gives −1.0 to +11.1 and −0.5 to +10.2 under the two coders (Appendix D.5), so the rise may not hold for other events. It raised the interaction by 5.3 points (−1.1 to +11.7), an interval wide enough to leave that change inconclusive. Breaking the rise in the setting contrast down by cell, most of it came from Thai text about a Thai event, where the other-benefit share rose by 7.0 and 5.3 points under the two coders, while the other three cells changed by less than 3 points.

4.4 What changed the offer: valence, domain, and model

The valence of the event moved both measures much further than any cultural cue did (Figure 3). Averaged over the coders, the other-benefit share was 24% for positive and for neutral events but 7% for negative ones, a drop of about 17 points where the cue contrasts stayed under 3 points. Negative events drew growth in 64% of their linings against 41% for positive events, and their relational goals fell from 24 to 13%, so a model answering about a loss mostly tells the student how they grew. Neutral events, which describe logistics, drew resources in 18% of their linings against 5% for positive events, with negative events close behind at 14%. The domain moved the profile as well: growth led in the academic schemas (64% of linings) and the sport schemas (42%), while relational goals led in the social ones (39% against 26% for growth).

Inside each valence, the four cells stayed within 4 points of each other in other-benefit share on average over the coders. In neutral events, however, the Thai-setting Thai-language cell gave growth in 42% of its linings against 24% in the US-setting English cell, in all six neutral schemas and under both coders (Figure 6 in Appendix D).

The two models differed more than the cells did and less than the valences did. GPT-5.6 Luna gave growth in 48% of its linings against DeepSeek's 41%, while DeepSeek gave more wellbeing (17% against 12%) and achievement (7% against 4%). GPT's other-benefit share was 2 to 5 points above DeepSeek's in every cell of the implicit arm. The larger difference is in Thai: 7% of DeepSeek's Thai-language items were not linings, against about 1% of its English-language items (Figure 7 in Appendix D).

4.5 Checks on the LLM coders

On the known-answer set, both LLM coders gave the intended item type for all 144 item versions and the intended goal for all 102 linings, in Thai and in English. GPT-6.1 Sol gave the wrong beneficiary for two linings and Claude Opus 5.5 for one, so they gave every intended code for 142 and 143 of the 144 versions. Both coders coded all twelve items with Thai cultural content correctly in both languages, including the items about karma, impermanence, acceptance, and kreng-jai (Appendix D.8). The coders therefore recognize this content when an item carries it, and its near absence in Section 4.1 belongs to the models' lists. A word search of all 5,190 items, which involves no coder, agrees: no item framed the event through karma, merit, or fate.

On the 36 held-out lists, the two human coders agreed with each other on 83% of item types (α

.43), 83% of beneficiaries (α .42), and 64% of goals (α .50). Each LLM coder agreed with each human coder as often as the two humans agreed with each other, or more often: on 84 to 89% of item types (α .42 to .64), 87 to 90% of beneficiaries (α .56 to .64), and 66 to 75% of goals (α .53 to .64). Agreement on the beneficiary was lower for Thai lists than for English ones, between the two humans (81% against 85%) and between each LLM coder and each human (83 to 85% against 91 to 95%). Over these lists, the four coders gave other-benefit shares of 17 to 19% (Appendix D.8).

Across all 5,190 items, the two LLM coders agreed on 97.3% of item types (α .84), 93.2% of goals (α .91), and 96.8% of beneficiaries (α .90). Agreement was about the same on Thai and English items and for the two generating models (Table 10). Their item-type disagreements concern where the lining threshold sits, since Claude Opus 5.5 read 111 items as linings that GPT-6.1 Sol coded otherwise, against 18 the other way. Their goal disagreements lay mostly between growth, wellbeing, and resources, and between achievement and growth.

5 Discussion

Little sensitivity to implicit cues. Under implicit cues, both models named other people or the group as beneficiaries at nearly the same rate in every cell, under both coders. The goals moved more than the beneficiary did: Thai-language requests drew somewhat more growth and no more relational goals, a shift that we report as exploratory. Naming the country, the most direct cue we gave, raised the setting contrast by only about four and a half points, so the small effects are not only a matter of faint cues. The valence of the event changed other-benefit share far more than any cue did, so the kind of event may shape what a user receives more than their language or setting do.

Growth above all. In every cell and in both languages, the models mostly offered the student their own growth, and the profile changed with the event rather than with the person. No item invoked karma, merit, or fate, and each LLM coder found only two acceptance-type consolations, so a reader who expects Thai accounts of coping to centre on acceptance would not recognize this offer. The Thai evidence does not make a growth-centred offer foreign, however: Thai adolescents who survived the 2004 tsunami named educational opportunities among the good that followed (Tuicomepee and Romano, 2008), and Thai and American children's coping differed mainly where a cultural norm was salient (McCarty et al., 1999). The beneficiary may also hide the culture. Thai adolescents describe a successful life as a way to repay their parents (Balthip et al., 2022), so a growth lining that names only the student may still serve the family in the student's own reading.

What counts as a silver lining. Two questions in our scheme call for judgment. The first asks whether an item names a gain at all. With the decision order of Section 3.4, the two LLM coders agreed on this for 97.5% of items, and they disagreed mainly about where the threshold sits. This judgment is the subjective part of the exercise, and a human coach faces the same one, namely whether a positive aspect will land as a gain for this person. Neither can read that from a short description, which is why our scheme records what an item offers and not how it would be received. The second question, which goal a lining serves, was harder: the coders agreed on 93% of goals, and they disagreed most between growth, wellbeing, and resources, and between achievement and growth. Separating what a person learns from what a person achieves or feels is a concrete target for future work.

6 Conclusion

We asked two widely used LLMs for six silver linings of 18 high-school events, in Thai and English, set in Thailand and the United States. What the models offered moved with the event rather than with the student: growth was the most frequent goal in every cell and reached 64% of linings after a negative event, and four linings in five named the student as the person who gains. The PAT workbook leaves open which positive aspects count, and a model asked to fill that gap fills it with the student's own growth, in whichever language it is asked. A system that leans on an LLM to guide this exercise inherits that choice: the same silver linings arrive whoever is asking, so it cannot assume they match the goals of the person receiving them.

Limitations

The human coders are two of the authors and Thai bilinguals who took their doctorates in the United States, so their codes show how two readers of

that background judged the items, not how Thai or US readers in general would. Our checks of the LLM coders are small: 72 known-answer items, which we wrote together with the codebook, and 36 held-out lists. They show that the coders can apply the codebook, including to Thai cultural content, but they do not establish how accurate the coders are across all 5,190 items. The goal categories are ours, and our coders disagreed most at the boundaries between growth and the other goals that concern the student, so the goal shares depend on where we drew them. Coders anchored each judgment in the header of an item, so a goal or a beneficiary named only later in a long item may be missed. The scheme records what an item offers, whose benefit it names, and which goal it serves, but not whether a student would experience the item as positive or helpful. Because we have no human baseline, our claims concern the sensitivity and direction of adaptation, not calibration against what Thai or US students would produce.

The stimuli are 18 schemas from six story families and three domains, all of them school events, so the inference is about these vignettes and these two models rather than a random sample of events or of models. We did not measure whether students in the two countries see each event as equally important, so a schema may carry more weight in one setting than in the other. Each call was a single user turn with no system prompt, whereas a deployed assistant carries a system prompt and an ongoing conversation. Future work could ask Thai and US students to write six positives for the same events and to judge the models' linings, which would give the human baseline that this study lacks.

Ethical Considerations

The stimuli are fictional situations about everyday school events, with invented school names and no personal data. We recruited no participants or annotators, and no model output was shown to people in distress. The study describes what two models produce when asked for silver linings, and it does not test the effect of these outputs on people. The generated silver linings are not clinical advice. Our results describe the models and not Thai or US students, so they should not be read as a claim about how people in either country think or cope. The stimuli, identifiability scores, prompts, codebook, model outputs, the codes of both LLM coders, the known-answer items, the human codes on the held-out lists, and the scripts are available at `https://github.com/yongsa-nut/growth-above-all`. We used AI assistants to help draft and edit parts of the text, and the authors checked all content and every reference.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

K. Balthip, B. Suwanphahu, and W. McSherry. 2022. Achieving fulfilment in life: Cultivating the mindset of gratitude among Thai adolescents. *SAGE Open*, 12(1):1–12.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67. Association for Computational Linguistics.

J. Cassaniti. 2015. *Living Buddhism: Mind, Self, and Emotion in a Thai Community*. Cornell University Press, Ithaca, NY.

M. S. Christopher, J. B. D'Souza, J. Peraza, and S. Dhaliwal. 2010. A test of the personality-culture clash hypothesis among college students in an individualistic and collectivistic culture. *International Journal of Culture and Mental Health*, 3(2):107–116.

M. G. Craske, H. Dour, M. Treanor, and A. E. Meuret. 2022. *Positive Affect Treatment for Depression and Anxiety: Therapist Guide*. Treatments That Work. Oxford University Press.

M. G. Craske, A. E. Meuret, A. Echiverri-Cohen, D. Rosenfield, and T. Ritz. 2023. Positive affect treatment targets reward sensitivity: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 91(6):350–366.

M. G. Craske, A. E. Meuret, T. Ritz, M. Treanor, H. Dour, and D. Rosenfield. 2019. Positive affect treatment for depression and anxiety: A randomized clinical trial for a core feature of anhedonia. *Journal of Consulting and Clinical Psychology*, 87(5):457–471.

DeepSeek. 2026. DeepSeek-V4.1-Flash: Smarter, faster, more efficient. Accessed 19 September 2026.

S. Ekasawin, C. Phothisut, and R. Chomchuen. 2016. The prevalence of psychiatric disorders in Thai students aged 13–17 years. *Journal of Mental Health of Thailand*, 24(3):141–153. In Thai with an English abstract.

K. K. Fitzpatrick, A. Darcy, and M. Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

S. Folkman. 1997. Positive psychological states and coping with severe stress. *Social Science & Medicine*, 45(8):1207–1221.

B. Q. Ford and I. B. Mauss. 2015. Culture and emotion regulation. *Current Opinion in Psychology*, 3:1–5.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

F. M. E. Grouzet, T. Kasser, A. Ahuvia, J. M. Fernández Dols, Y. Kim, S. Lau, R. M. Ryan, S. Saunders, P. Schmuck, and K. M. Sheldon. 2005. The structure of goal contents across 15 cultures. *Journal of Personality and Social Psychology*, 89(5):800–816.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Y. He, L. Yang, C. Qian, T. Li, Z. Su, Q. Zhang, and X. Hou. 2023. Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*, 25:e43862.

M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, and N. C. Jacobson. 2025. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802.

V. S. Helgeson, K. A. Reynolds, and P. L. Tomich. 2006. A meta-analytic review of benefit finding and growth. *Journal of Consulting and Clinical Psychology*, 74(5):797–816.

A. Herderich, J. Z. Li, P. Nair, A. Perry, and A. Goldenberg. 2026. Exploring human and AI emotional support through reframing of negative situations. PsyArXiv preprint.

N. Khirikoekkong, S. Asarath, S. Nosten, B. Hanboonkunupakarn, N. Jatupornpimol, J. Roest, M. Parker, F. Nosten, R. McGready, P. Y. Cheah, and M. Kelley. 2023. Culturally responsive research ethics: How the socio-ethical norms of Arrnar/Kreng-jai inform research participation at the Thai-Myanmar border. *PLOS Global Public Health*, 3(5):e0001875.

P. Kittirattanapaiboon, N. Tantirangsee, W. Chutha, A. Tanaree, P. Kwansanit, and S. Assanangkornchai. 2017. Prevalence of mental disorders and mental health problems: Thai national mental health survey 2013. *Journal of Mental Health of Thailand*, 25(1):1–19. In Thai with an English abstract.

S. Komin. 1990a. Culture and work-related values in Thai organizations. *International Journal of Psychology*, 25(3–6):681–704.

S. Komin. 1990b. *Psychology of the Thai People: Values and Behavioral Patterns*. National Institute of Development Administration, Bangkok.

R. S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York.

R. S. Lazarus and S. Folkman. 1984. *Stress, Appraisal, and Coping*. Springer Publishing Company, New York.

H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. C. Mohr. 2023. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6(1):236.

Chen Cecilia Liu, Hiba Arnaout, Nils Kovačić, Dana Atzil-Slonim, and Iryna Gurevych. 2026. Tailored emotional LLM-supporter: Enhancing cultural sensitivity. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 535–574. Association for Computational Linguistics.

H. R. Markus and S. Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253.

S. C. Marsella and J. Gratch. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

R. K. McBain, R. Bozick, M. Diliberti, L. A. Zhang, F. Zhang, A. Burnett, A. Kofner, B. Rader, J. Breslau, B. D. Stein, A. Mehrotra, L. Uscher-Pines, J. Cantor, and H. Yu. 2025. Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11):e2542281.

C. A. McCarty, J. R. Weisz, K. Wanitromanee, K. L. Eastman, S. Suwanlert, W. Chaiyasit, and E. B. Band. 1999. Culture, coping, and context: Primary and secondary control among Thai and American

youth. *Journal of Child Psychology and Psychiatry*, 40(5):809–818.

K. McRae, B. Ciesielski, and J. J. Gross. 2012. Unpacking cognitive reappraisal: Goals, tactics, and outcomes. *Emotion*, 12(2):250–255.

A. E. Meuret, H. Dour, A. Loerinc Guinyard, and M. G. Craske. 2022. *Positive Affect Treatment for Depression and Anxiety: Workbook*. Treatments That Work. Oxford University Press.

A. E. Meuret, D. Rosenfield, E. Wang, C. M. Hough, T. Ritz, and M. G. Craske. 2026. Positive affect treatment for depression, anxiety, and low positive affect: A randomized clinical trial. *JAMA Network Open*, 9(4):e267403.

A. C. Mills, J. Poogpan, C. Wong-Anuchit, and D. Rujkorakarn. 2019. The meaning of acceptance (Thumjai) in Thai people: Letting it go… so life goes on. *International Journal of Mental Health Nursing*, 28(4):879–887.

A. C. Mills, C. Wong-Anuchit, and J. Poogpan. 2017. A concept analysis of Thum-jai: A Thai coping strategy. *Pacific Rim International Journal of Nursing Research*, 21(3):234–243.

OpenAI. 2026. Improving GPT-5.6 Sol in ChatGPT—and expanding access to GPT-5.6 Luna for free users. Accessed 19 September 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 68772–68802.

S. R. Sears, A. L. Stanton, and S. Danoff-Burg. 2003. The yellow brick road and the emerald city: Benefit finding, positive reappraisal coping and posttraumatic growth in women with early-stage breast cancer. *Health Psychology*, 22(5):487–497.

S. Sergeant and M. Mongrain. 2014. An online optimism intervention reduces depression in pessimistic individuals. *Journal of Consulting and Clinical Psychology*, 82(2):263–274.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, pages 700:1–700:29. Association for Computing Machinery.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000. Association for Computational Linguistics.

S. C. Sodergren and M. E. Hyland. 2000. What are the positive consequences of illness? *Psychology & Health*, 15(1):85–97.

H. Song, J. S. Chan, and C. Ryan. 2024. Differences and similarities in the use of nine emotion regulation strategies in Western and East-Asian cultures: Systematic review and meta-analysis. *Journal of Cross-Cultural Psychology*, 55(8):865–885.

Ala N. Tak and Jonathan Gratch. 2023. Is GPT a computational model of emotion? In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

R. G. Tedeschi and L. G. Calhoun. 1996. The posttraumatic growth inventory: Measuring the positive legacy of trauma. *Journal of Traumatic Stress*, 9(3):455–471.

Tobias Thejll-Madsen, Nutchanon Yongsatianchot, and Stacy Marsella. 2026. A systematic evaluation framework on emotion inference for large language models using appraisal theory. *IEEE Transactions on Affective Computing*, 17(3):3626–3639.

A. Tuicomepee and J. L. Romano. 2008. Thai adolescent survivors 1 year after the 2004 tsunami: A mixed methods study. *Journal of Counseling Psychology*, 55(3):308–320.

A. Uusberg, J. L. Taxer, J. Yih, H. Uusberg, and J. J. Gross. 2019. Reappraising reappraisal. *Emotion Review*, 11(4):267–282.

V. L. Vignoles, E. Owe, M. Becker, P. B. Smith, M. J. Easterbrook, R. Brown, R. González, N. Didier, D. Carrasco, M. P. Cadena, S. Lay, S. J. Schwartz, S. E. Des Rosiers, J. A. Villamar, A. Gavreliuc, M. Zinkeng, R. Kreuzbauer, P. Baguma, M. Martin, and 53 others. 2016. Beyond the 'East–West' dichotomy: Global variation in cultural models of selfhood. *Journal of Experimental Psychology: General*, 145(8):966–1000.

World Health Organization. 2022. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, Geneva.

Nutchanon Yongsatianchot and Stacy Marsella. 2022. Modeling emotion-focused coping as a decision process. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Nutchanon Yongsatianchot and Stacy Marsella. 2023. A computational model of coping and decision making in high-stress, uncertain situations: An application to hurricane evacuation decisions. *IEEE Transactions on Affective Computing*, 14(3):2539–2556.

Nutchanon Yongsatianchot and Stacy Marsella. 2025. Investigating motivated inference in large language models. In *Proceedings of the 9th Widening NLP Workshop*, pages 182–196, Suzhou, China. Association for Computational Linguistics.

Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating large language models' perception of emotion using appraisal theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. In *Proceedings of the First Conference on Language Modeling (COLM)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, pages 46595–46623.

W. Zhong, J. Luo, and H. Zhang. 2024. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders*, 356:459–469.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700. Association for Computational Linguistics.

A Stimuli

A.1 Construction rules

We chose school, friends, and school sport as the domains because both countries have close equivalents of them in the everyday life of high-school students. In these domains, both settings offer localizing detail without changing the appraisal structure of an event. We built each situation from three slots, which are the context, the effort, and the outcome. The context says who is involved and what happens, and it is the same across valence. Only the positive and negative versions have an effort slot.

The 18 schemas form six story families. Each family is a positive and negative pair plus a neutral schema from the same domain, such as S1 to S3 in Table 1. The positive and negative versions differ in exactly one clause. In some schemas this clause states the outcome itself (won or lost the final; my name was or was not on the roster), and in others it states the outcome relative to expectation (higher or lower than expected). Therefore, effort, cause, visibility, and reversibility are identical by construction. We included no margin or near-miss detail, because a phrase such as "by one point" gives the model a counterfactual lining.

A neutral schema is not the pleasant routine version of the domain, because that version would still serve affiliation or competence. Instead, a neutral schema describes a logistics event inside the domain with no goal at stake. This follows the exercise, which can start from positive or neutral situations, such as visiting the doctor or brushing one's teeth (Craske et al., 2022, p. 89). For instance, our neutral events include a gear pick-up, a form handed in, a bill split, and a team photo. The neutral texts show the goal profile of a model on routine events.

No situation names a goal or says why the event mattered. For instance, if a text said "I really wanted to make the team," the goal codes would only repeat the stimulus. The model supplies the goals, and these goals are what we measure.

The four cells of a schema tell the same story. An invitation has the same scope in both settings, and the effort has the same mechanism. For instance, the effort is self-study in both settings rather than tutoring in one of them. A result has the same visibility and is private in both settings. The equipment, the paperwork, the action (fill in and hand in), and the register are the same as well.

The setting enters only through one or two nouns per text, and we give each pair below in the form US / Thai. Some nouns name a venue or brand, such as Chick-fil-A / หมูกระทะ, AMC / Major Cineplex, Venmo / PromptPay, and iMessage / LINE. Other nouns name an institution or system, such as counselor / homeroom teacher, PowerSchool / M.5 (the Thai name for grade 11), science fair / Science Week, and playoffs / provincial student games. The third kind of noun is an invented school name, Lincoln High / Suksawat Wittaya. The sport is basketball in both settings, so no sport term marks the setting in either language. Therefore, a setting effect would reflect a

response to these nouns and not a response to a different event.

A.2 Localization and length

We wrote the Thai-setting texts natively in Thai from the schema, and we did not translate them from English. A bilingual author produced the two off-diagonal renderings of each schema. A bilingual colleague who is a second native speaker of Thai then revised all Thai text. The same colleague also reviewed the three instruction wordings, and we checked the off-diagonal renderings against the schema slots. Both languages use a casual-standard teen register, with เรา as the first person, no politeness particles, and ordinary loanwords.

We held the length of every text within $\pm 20\%$ of the median of its cell. The US-setting English texts run from 24 to 30 words (median 26), and the Thai-setting English texts run from 25 to 31 words (median 27). We measured Thai length in characters without spaces. The US-setting Thai texts run from 88 to 127 characters (median 106.5), and the Thai-setting Thai texts run from 92 to 132 characters (median 111).

A.3 Setting-identifiability check

We did not assume that a model can tell where an event takes place, so we measured it before generation. We sent each text to each model with the implicit-arm persona line and the question "Based only on this message, which country does this student most likely live in? Answer with one country name." We asked the question in Thai for the Thai texts, and we drew five samples per model at temperature 1. A text passed when at least four of the five answers from each model named the intended country. A text that failed received one more setting noun and was checked again.

On the first pass, 57 of the 72 texts passed. The failures were concentrated in two groups. The first group was the Thai-setting English texts. When the only cues in these texts were a dish, a messaging app, or a school event, the models read them as set in Singapore, Japan, Canada, or India. The second group was the US-setting texts of four schemas, whose cues (a counselor, a science fair, and a book room) signaled little in Thai, and three of these schemas also failed in English.

We added the invented school name to all four cells of the nine schemas with a failing text. On the second pass, every text passed on both models, and 143 of the 144 text-by-model checks reached five of five. The repository named in the Ethical Considerations includes the per-text scores.

A.4 The 18 schemas

Table 1 lists the 18 schemas in the two diagonal cells.

In rows that begin with . . . , the text repeats the context of the preceding row and differs only in the outcome clause. The two off-diagonal renderings of every schema are in the stimulus file of the repository named in the Ethical Considerations.

The Thai texts carry the setting in these nouns: หมูกระทะ (Thai barbecue), กีฬาสี (the school sports day), แชทไลน์ (chatting on LINE), เมเจอร์ ซีนีเพล็กซ์ (Major Cineplex), พร้อมเพย์ (PromptPay), กีฬานักเรียนระดับจังหวัด (the provincial school games), งานสัปดาห์วิทยาศาสตร์ (science week), and สุขสวัสดิ์วิทยา (Suksawat Wittaya, the invented school). เรา is the casual first person a Thai teenager uses, and ม.ปลาย and ม.5 are senior high school and its second year.

ID	Domain	Valence	US setting, English	Thai setting, Thai
S1	social	pos	The friends I usually hang out with are planning to go to Chick-fil-A after the Lincoln High homecoming game Friday, and they asked me to come.	กลุ่มเพื่อนที่เราไปไหนมาไหนด้วยกันประจำนัดกันว่าเลิกกีฬาสีของสุขสวัสดิ์วิทยาวันศุกร์จะไปกินหมูกระทะกันต่อ เขาชวนเราไปด้วย
S2	social	neg	… didn't ask me.	… เขาไม่ได้ชวนเรา
S3	social	neu	The friends I usually hang out with at Lincoln High spent the afternoon on iMessage deciding what to do and where to meet Friday. Then I went back to homework.	กลุ่มเพื่อนที่เราไปไหนมาไหนด้วยกันที่สุขสวัสดิ์วิทยาแชทไลน์ทั้งบ่ายว่าวันศุกร์จะทำอะไรเจอกันที่ไหน แล้วเราก็กลับไปทำการบ้านต่อ
S10	social	pos	I asked a few friends from my class to go see the new movie at AMC on Saturday afternoon, and they all said yes.	เราชวนเพื่อนในห้องสองสามคนไปดูหนังเรื่องใหม่ที่เมเจอร์ซีนีเพล็กซ์ วันเสาร์บ่ายนี้ ทุกคนบอกว่าไปได้
S11	social	neg	… they were busy.	… ทุกคนบอกว่าไม่ว่าง
S12	social	neu	We spent the evening in the group chat sorting out who owed what from last weekend and settling up on Venmo. Once that was done, I went to bed.	เรากับเพื่อนแชทกันทั้งเย็นเรื่องหารค่าใช้จ่ายของเมื่อวันหยุด แล้วก็โอนพร้อมเพย์กัน พอเสร็จเราก็ไปนอน
S4	sport	pos	I'm on the basketball team at Lincoln High, and we've been practicing twice a day since summer for the playoffs. We won the regional final.	เราอยู่ทีมบาสโรงเรียนสุขสวัสดิ์วิทยา ซ้อมเช้าเย็นมาตั้งแต่ปิดเทอมเพื่อไปแข่งกีฬานักเรียนระดับจังหวัด เราชนะนัดชิง
S5	sport	neg	… lost the regional final.	… เราแพ้นัดชิง
S6	sport	neu	I'm on the basketball team at Lincoln High. Today was gear day: we picked up our new jerseys, turned in our athletics forms, and went home.	เราอยู่ทีมบาสโรงเรียนสุขสวัสดิ์วิทยา วันนี้เป็นวันรับของ ไปรับเสื้อทีมชุดใหม่ เอาเอกสารไปส่งโค้ช แล้วก็กลับบ้าน
S13	sport	pos	I've been going to open gym all summer to get ready for basketball tryouts at Lincoln High. Today the coach posted the roster, and my name was on it.	เราซ้อมบาสมาทั้งปิดเทอมเพื่อไปคัดตัวทีมบาสโรงเรียนสุขสวัสดิ์วิทยา วันนี้โค้ชติดประกาศรายชื่อ มีชื่อเรา
S14	sport	neg	… wasn't on it.	… ไม่มีชื่อเรา
S15	sport	neu	I'm on the basketball team at Lincoln High. Today was team photo day: we lined up in the gym in our jerseys, took the picture, and went back to class.	เราอยู่ทีมบาสโรงเรียนสุขสวัสดิ์วิทยา วันนี้มีถ่ายรูปทีม เราใส่ชุดแข่ง ต่อแถวถ่ายภาพหน้าเสาธง ถ่ายเสร็จก็กลับไปเรียนต่อ
S7	academic	pos	I studied really hard for a month for my bio midterm, and when the grades showed up on PowerSchool I'd scored higher than I expected.	เราอ่านหนังสือหนักมากมาทั้งเดือนเพื่อสอบกลางภาคชีวะ ม.5 พอครูคีนข้อสอบมา เราได้สูงกว่าที่คิดไว้
S8	academic	neg	… lower than I expected.	… เราได้ต่ำกว่าที่คิดไว้
S9	academic	neu	Today we had to hand in next semester's elective form at Lincoln High. I filled in my choices, gave it to my counselor at lunch, and went back to class.	วันนี้ที่สุขสวัสดิ์วิทยาต้องส่งใบเลือกวิชาเลือกเสรีของเทอมหน้า เรากรอกเสร็จก็เอาไปยื่นให้ครูประจำชั้นตอนพักเที่ยง แล้วก็กลับไปเรียนต่อ
S16	academic	pos	I worked on my project for a month for the Lincoln High science fair, and when the results were announced I'd placed higher than I expected.	เราทำโครงงานส่งประกวดในงานสัปดาห์วิทยาศาสตร์ของโรงเรียนสุขสวัสดิ์วิทยามาทั้งเดือน พอประกาศผล เราได้อันดับดีกว่าที่คิดไว้
S17	academic	neg	… lower than I expected.	… เราได้อันดับแย่กว่าที่คิดไว้
S18	academic	neu	Today we got our textbooks for next semester. I picked mine up from the Lincoln High book room at lunch, signed for them, and went back to class.	วันนี้แจกหนังสือเรียนของเทอมหน้า เราไปรับชุดของ ม.5 ที่ห้องสมุดสุขสวัสดิ์วิทยาตอนพักเที่ยง เซ็นชื่อรับ แล้วก็กลับไปเรียนต่อ

Table 1: The 18 schemas in the two diagonal cells, grouped by domain, with the two story families of a domain one after the other. The labels pos, neg, and neu stand for positive, negative, and neutral.

B Prompt and generation

B.1 Persona lines and prompt wording

Each prompt was written entirely in the language of its text. Table 2 gives the persona lines of both arms. No text gives the asker an age or a gender. In the explicit arm, the line names the country of the setting of the text. Table 3 gives the three instruction wording pairs.

The wordings share one frame. Each asks the model to help the student think ("Please help me think of", ช่วย…หน่อย), which suits a request made after a bad event as well as a good one, and each refers to the situation as "this event" (เรื่องนี้). Each also asks for exactly six items, where the workbook asks for at least six, so that every list has the same number of items to code.

The three pairs span what a user might actually type. P1 (things) asks for features of the event, and P2 (sides) asks for aspects of it. P3 (angles) asks for ways to look at the event, which are perspectives on it. We did not use the Thai word ข้อดี, because it reads as the "pros" of pros-and-cons and invites utility items.

The prompts asked for six positive aspects of an event without presenting the exercise as therapy, without the attention-training rationale, and without the workbook's mood rating, so that each request read as an everyday request to an assistant. They also did not include the workbook's permission that small or silly aspects count, because that line may cue trivial content and because the model's own threshold for what counts is part of what we measure. The prompts gave no instruction about list format.

B.2 Models and settings

Table 4 gives the two generating models and their settings.

We use both models as proxies for the consumer experience rather than as the experience itself, because the consumer apps run hidden system prompts that cannot be reproduced. DeepSeek V4.1 Flash has no pinned snapshot on the official API. Therefore, we logged the model string that the provider returned for every call, and we kept generation within a short window. We made all 864 calls within fifteen minutes on 18 September 2026.

We randomized and interleaved the calls across cell, model, arm, and wording, and each call ran in a fresh context with a limit of 8,192 output tokens. For each call, we logged the model identifier as returned, the provider, the parameters, a hash of the prompt, the timestamp, any truncation, and any retries. We checked the first call to each model by hand for an empty reasoning field and the expected provider.

B.3 Samples and parsing

We stored the raw responses unchanged and parsed each one afterwards as a numbered list. When a response had fewer or more than six items, was in the wrong language, or contained a refusal or meta-commentary, we kept it and flagged it, and Appendix D.2 reports these cases. We did not resample flagged responses. Flagged lists stay in every analysis, and Appendix D.5 reports the contrasts in other-benefit share without them. No response was a refusal, and four responses had seven or eight items instead of six (three for neutral events and one for a negative event).

Arm	Setting	English	Thai
Implicit	US and Thai	I'm in high school.	เราเรียนอยู่ม.ปลาย
Explicit	US	I'm in high school in the USA.	เราเรียนอยู่ม.ปลายที่อเมริกา
Explicit	Thai	I'm in high school in Thailand.	เราเรียนอยู่ม.ปลายในประเทศไทย

Table 2: The persona lines.

Pair	English	Thai
P1 (things)	Please help me think of 6 good things about this event	ช่วยคิดหน่อยว่าเรื่องนี้มีสิ่งดีๆ อะไรบ้าง สัก 6 อย่าง
P2 (sides)	Please help me think of 6 good sides to this event	ช่วยคิดแง่ดีของเรื่องนี้ให้หน่อย สัก 6 ข้อ
P3 (angles)	Please help me think of 6 positive ways to look at this event	ช่วยคิดมุมดีๆ ของเรื่องนี้ให้หน่อย สัก 6 มุม

Table 3: The three instruction wording pairs.

Model	Access	Reasoning	Sampling	Reason for choice
GPT-5.6 Luna	OpenAI API, `gpt-5.6-luna`	`reasoning_effort: none`	temperature 1 (the only accepted value); no penalties	Default model for free ChatGPT users at the time of generation (OpenAI, 2026)
DeepSeek V4.1 Flash	OpenRouter, `deepseek/deepseek-v4.1-flash`, provider pinned to DeepSeek, no fallbacks	`reasoning: {enabled: false}`	temperature 1; `top_p` fixed at 1.0 in non-thinking mode	Newest model from a widely used Chinese developer (DeepSeek, 2026)

Table 4: The two generating models and their settings.

C Coding scheme details

C.1 Item types

The item type is the first yes in a fixed order: off-task, action, lining, consolation, and otherwise not a lining. Off-task items, such as story angles, refusals, meta comments, and life tips with no link to the event, are not linings. An action tells or recommends an outward step that the asker takes (ลอง..., ควร...). A lining finds a gain in the event itself. The gain is something the asker or the people around them got, kept, avoided, learned, or can do because of the event, a good result of it, or what it offers while it happens. A consolation only changes how the asker should see or feel about the event, as in "not your fault", "at least...", or ทำใจ (accept it). Anything else, such as a retelling with nothing good named or a feature of the event that did nothing for anyone, is not a lining.

Each item is coded from its header, or from its first sentence when it has no header, and the rest only settles what the header leaves open. Thus "It's a chance to diversify your social circle: say hi to people in your classes, join a club" is a lining, because the suggested steps come only after the header.

C.2 Decision rules

Item type. An instruction to see or feel differently is not an action: it is a lining when it names a gain ("be proud of reaching the semifinal") and a consolation otherwise ("don't blame yourself"). A good result is a lining even when it is the event's own outcome ("you made the varsity squad"). An item that opens with "at least" or another comparison to something worse is a consolation unless it says what the asker can now do ("at least you found out now, so you get to make your own plan"). This check comes before the lining question. An item that says the chance is not lost is also a consolation ("there will be other tryouts"), but time that the asker now has is a lining ("you have time to build your skills").

Goal. The goal is the kind of gain for whoever gains it, except that what family or friends feel because of the asker (proud, happy, relieved) is relational. A quality the event showed, or effort that paid off with no outcome named, is growth, while effort whose outcome the item names is achievement ("your hard work paid off: you got an A"). A feeling, or a mind spared worry or hassle, is well-being, while time, money, or things in order, with nothing said about the mind, are resources ("a free Friday night"). Not having to burden, chase, or tiptoe around friends, family, or the group (kreng-jai, เกรงใจ) is relational even when worded as a feeling, but the same relief for anyone else goes by its kind.

Beneficiary. The beneficiary is the one who gains, not the one who is mentioned. A shared good, such as a friendship strengthened, fairness within the group, or pride brought to the school, is group. A good that lands on the asker alone is self even when others are involved. Thus "your teammates gained confidence" is group, while "your teammates respect you more" is self. When an item names two beneficiaries, the main clause decides. Thai เรา (I or we by context) is not by itself a cue for group.

C.3 Goal categories

When the lead clause names two gains, the first in the order below decides.

- **Transcendent**: meaning beyond the event, such as fate, karma, or merit.
- **Community**: service to, pride for, or representing a collective beyond the asker's own relationships.
- **Relational**: a connection between people, such as closeness, trust, repaying one's parents, or face for the family.
- **Growth**: the student got better or did something new, such as a skill or a lesson.
- **Achievement**: a positive outcome or recognition that the item names, such as a win or being selected.
- **Wellbeing**: a feeling or a lighter mind, such as fun, rest, or relief.
- **Resources**: a better-arranged situation, such as time, money, or a record kept.

The seven categories adapt the eleven goal con-

Question	Options	Applies to
Item type: what does the item give the asker?	lining / consolation / action / not a lining	all items
Consolation type	acceptance / impermanence / could be worse / reassurance / fate or meaning / other	consolations
Goal: what kind of gain is it?	transcendent / community / relational / growth / achievement / wellbeing / resources / unspecified	linings
Beneficiary: whose gain is it?	self / other / group / transcendent order / unspecified	linings

Table 5: The coding questions. Every item receives an item type, a consolation also receives a consolation type, and a lining also receives a goal and a beneficiary.

tents of Grouzet et al. (2005), and they are not a validated instrument. Relational takes affiliation and conformity, community takes community feeling, transcendent takes spirituality, wellbeing takes hedonism, growth takes self-acceptance, and resources takes financial success. Achievement has no direct counterpart, and image, popularity, physical health, and safety have no category of their own. The goals also fall into three coarse levels: self (growth, achievement, wellbeing, resources), social (relational, community), and beyond (transcendent).

C.4 Development set

We developed the codebook on a separate set of 32 implicit-arm lists, and some examples in this appendix come from them, so these lists are not used to check the coders (Section 3.5).

D Additional results

Where this appendix gives a pair of values for one quantity without naming the coders, the first is under GPT-6.1 Sol and the second under Claude Opus 5.5. In the tables, Sol and Opus abbreviate the two coders.

D.1 Descriptives by cell

Table 6 gives the item types, the two most frequent goals, and the other-benefit share of the implicit arm by setting, language, and model, and Table 7 gives the explicit arm in the same layout. In GPT's English-language lists of the implicit arm, GPT-6.1 Sol coded 6 to 7% of items as actions where Claude Opus 5.5 coded 3 to 4%. In DeepSeek's Thai-language lists about a US event, GPT-6.1 Sol coded 10 and 12% of items as not a lining where Claude Opus 5.5 coded 6 and 10%. The two coders' other-benefit shares differ by at most 3 points in any row.

D.2 Item counts and response format

Of the 864 lists, 860 had exactly six items, two had seven, and two had eight, which gives 5,190 items. Three of the four longer lists answered a neutral event and one a negative event, and three are in the implicit arm. One DeepSeek response to a Thai-language request in the explicit arm was written entirely in Chinese, and we coded it as it stands. No call failed, every response contained a list, and none was a refusal.

A character check found letters from other writing systems inside the items of at least 33 of DeepSeek's 216 Thai-language lists, 16 of them in the implicit arm. These were Chinese characters in 25 lists and Cyrillic, Arabic, Japanese, or Korean letters in the others. No item of GPT's Thai-language lists or of any English-language list contained such letters. The check counts only letters outside the Thai and Latin alphabets, so it misses words from other languages written in Latin letters, and 33 is a lower bound.

The two models formatted their lists differently. Across cells and arms, DeepSeek V4.1 Flash wrote a preamble before every list and a closing paragraph after 93 to 100% of them. It ended 17 to 31% of its English-language lists with a question back to the asker. GPT-5.6 Luna wrote a preamble before 7 to 70% of its lists and a closing paragraph after 0 to 48%, and it wrote preambles more often in Thai. DeepSeek's items averaged about 180 characters, not counting spaces, which is about 1.7 times the length of GPT's.

D.3 Valence and model

Consolations came almost entirely from negative events. In the implicit arm, 100 of GPT-6.1 Sol's 110 consolations and 99 of Claude Opus 5.5's 105 answered a negative event, where they made up 12% and 11% of items. Neutral events drew 8 and 5 consolations, and positive events 2 and 1. Actions made up 4% of the items after negative events under both coders, against 3% and 1% after positive events and under 1% after neutral events. Items coded not a lining clustered in neutral events (53 of 78 and 45 of 54). More than half of all such items (43 and 33) came from DeepSeek's Thai-language lists under the third wording, which include the story-angle lists of Section 4.1.

In the implicit arm, the gap in other-benefit share between the generating models came from neutral events, where GPT's share was 29% under both coders against DeepSeek's 18%. For positive events the two models were within 0.4 points of each other, and for negative events within 2.5 points. Paired by text, GPT's share exceeded DeepSeek's by 3.6 and 4.3 points. The intervals over lists (+0.9 to +6.0 and +2.1 to +6.7) exclude zero, but the intervals over schemas (-0.9 to +9.7 and -0.2 to +10.0) do not, so the gap depends on which schemas are sampled.

In the explicit arm the gap reversed. DeepSeek was ahead by 3.1 points under GPT-6.1 Sol (-5.9 to -0.2 over lists, -8.5 to +2.2 over schemas) and by 1.0 under Claude Opus 5.5 (-3.9 to +1.8 and -4.5 to +2.8). Naming the country raised DeepSeek's share for positive events from 24 to 31% and from 23 to 28%, and for neutral events from 18 to 25% and from 18 to 26%. GPT's shares for these events changed by less than 4 points. In the implicit-arm cells with Thai text or a Thai setting, DeepSeek's share of relational and community goals stayed within 2 points of GPT's under both coders.

D.4 Wording

The language contrast in other-benefit share depended on the wording pair. With intervals from a bootstrap over schemas, it was -0.5 points under P1 (things; -4.9 to +4.2), -8.3 under P2 (sides; -16.2 to -2.2), and +6.2 under P3 (angles; +0.6 to +12.5) under GPT-6.1 Sol. Under Claude Opus 5.5 it was +0.3 (-4.4 to +4.5), -6.9 (-13.1 to

Setting	Language	Model	Lining	Consolation	Action	Not a lining	Growth	Relational	Other-benefit share
US	English	GPT	88 / 92	3 / 3	7 / 3	2 / 2	44 / 42	20 / 21	20 / 22
US	English	DeepSeek	93 / 95	5 / 4	2 / 1	1 / 0	37 / 36	21 / 20	17 / 15
US	Thai	GPT	93 / 94	4 / 4	1 / 1	2 / 2	50 / 50	22 / 21	20 / 19
US	Thai	DeepSeek	83 / 86	6 / 6	2 / 2	10 / 6	42 / 43	23 / 22	18 / 17
Thai	Thai	GPT	96 / 97	3 / 3	0 / 0	1 / 1	53 / 57	19 / 20	18 / 20
Thai	Thai	DeepSeek	88 / 88	5 / 5	2 / 2	6 / 6	44 / 44	21 / 20	15 / 16
Thai	English	GPT	90 / 92	3 / 3	6 / 4	1 / 1	46 / 43	20 / 21	20 / 21
Thai	English	DeepSeek	92 / 94	5 / 4	2 / 2	2 / 1	41 / 42	21 / 21	16 / 16

Table 6: Item types, leading goals, and other-benefit share in the implicit arm, by setting, language, and model. Each cell gives the value under GPT-6.1 Sol and then under Claude Opus 5.5 (Sol / Opus). The item-type columns give percentages of items and may not sum to 100 because of rounding, the goal columns give percentages of linings, and other-benefit share is the mean over lists, in percent.

Setting	Language	Model	Lining	Consolation	Action	Not a lining	Growth	Relational	Other-benefit share
US	English	GPT	92 / 91	4 / 5	2 / 2	2 / 2	37 / 39	21 / 22	18 / 19
US	English	DeepSeek	90 / 92	6 / 5	2 / 2	3 / 1	34 / 36	22 / 21	19 / 19
US	Thai	GPT	92 / 92	6 / 6	1 / 1	2 / 1	54 / 54	21 / 21	16 / 16
US	Thai	DeepSeek	79 / 83	8 / 6	1 / 1	12 / 10	41 / 43	25 / 22	21 / 18
Thai	Thai	GPT	94 / 94	4 / 3	1 / 0	2 / 2	59 / 58	18 / 19	21 / 22
Thai	Thai	DeepSeek	85 / 90	5 / 4	1 / 0	9 / 6	43 / 47	24 / 23	24 / 24
Thai	English	GPT	91 / 93	2 / 3	5 / 4	1 / 1	45 / 46	20 / 21	20 / 21
Thai	English	DeepSeek	94 / 94	4 / 4	1 / 1	1 / 1	39 / 40	22 / 21	20 / 21

Table 7: Item types, leading goals, and other-benefit share in the explicit arm, by setting, language, and model, in the layout of Table 6 (Sol / Opus).

−1.6), and +5.8 (+0.5 to +11.9). The intervals for P2 and P3 thus exclude zero in opposite directions and reach past the 10-point band.

In 2,000 permutations that shuffled the wording labels within setups, none reached the observed spread of the language contrast across the three wordings, 14.5 and 12.7 points ($p < .001$ under both coders). The spread of the setting contrast, 0.9 and 2.8 points, gave $p = .96$ and .65. Because lists without a lining drop out, the three estimates rest on 36, 36, and 35 combinations of schema and model under each coder. The third wording also drew more consolations (8.2% and 8.1% of items) than the second (3.5% and 3.0%) and the first (1.0% under both).

D.5 Sensitivity analyses

Each analysis below gives the three contrasts in other-benefit share with both models together. In the implicit arm, excluding the three lists flagged at generation, all with seven or eight items, gave −0.6, −1.4, and −2.7 points for language, setting, and interaction, and −0.2, −0.1, and −0.4. The other two flagged lists are in the explicit arm. Leaving out one story family at a time moved no estimate by more than 1.6 and 1.2 points. Counting other and group linings over all items, so that no list drops out, gave −1.4, −1.0, and −0.4, and −0.9, +0.5, and −0.3. Setting other-benefit share to zero for lists without a lining gave −0.7, −1.2, and −2.6, and −0.5, −0.1, and −0.3. Excluding the 32 development lists gave 0.0, −1.1, and −4.4, and 0.0, +0.1, and −1.6. Under GPT-6.1 Sol, the interval of that interaction (−8.6 to −0.3) excludes zero but stays inside the band. Excluding the 33 DeepSeek lists with letters from other writing systems moved no estimate by more than 0.5 points. Each of these analyses kept the reading of no shift of practical interest under both coders, as did the per-model contrasts of Table 8.

Three further bootstraps test whether the intervals are too narrow. A bootstrap over the 18 schemas gave −3.7 to +2.3, −4.6 to +1.7, and −6.7 to +1.4, and −3.6 to +2.9, −3.0 to +2.8, and −4.4 to +3.3. A bootstrap corrected for the three lists per setup gave −3.3 to +1.9, −4.1 to +1.2, and −7.7 to +2.5, and −2.7 to +2.2, −2.6 to +2.1, and −5.1 to +4.3. A two-stage bootstrap that also resamples the schemas gave −4.6 to +3.5, −5.5 to +2.7, and −9.5 to +3.6, and −4.4 to +3.9, −3.7 to +3.4, and −6.9 to +5.9. All of these intervals lie inside the band, so the reading of Section 4.2 holds. The interaction for each model alone is less robust. The corrected bootstrap widens GPT's interval under GPT-6.1 Sol to −10.8 to +5.1, and the two-stage bootstrap takes both models' intervals past the band under both coders, so these per-model interactions read as inconclusive.

The corrected and two-stage bootstraps matter more in the explicit arm. There, the corrected boot-

strap gave −1.4 to +4.3, +0.6 to +6.1, and −2.2 to +9.4, and −2.4 to +2.9, +1.6 to +6.9, and −1.5 to +9.1. For the change between the arms, it gave −1.8 to +5.8, +0.8 to +8.6, and −1.6 to +14.1, and −2.9 to +4.0, +1.0 to +8.2, and −2.6 to +11.3. These intervals leave the readings of Section 4.3 for both models together unchanged. For each model alone, GPT's explicit interaction widens to −4.1 to +10.6 and −4.0 to +10.5, so it reads inconclusive, like DeepSeek's. Under GPT-6.1 Sol, the rise in DeepSeek's setting contrast widens to −0.8 to +10.3 and also reads inconclusive.

The two-stage bootstrap gave −0.8 to +7.9 and +0.1 to +9.1 for the explicit setting contrast. For the explicit interaction it gave −3.8 to +11.9 and −2.9 to +10.9, so that contrast reads inconclusive under both coders. For the rise in the setting contrast between the arms, it gave −1.0 to +11.1 and −0.5 to +10.2. These intervals include zero and reach past the band, so that rise rests on treating the 18 schemas as fixed. Under this bootstrap, every per-model interaction and every per-model rise in the setting contrast reads inconclusive. So do DeepSeek's explicit setting contrast under Claude Opus 5.5 and the rise in DeepSeek's language contrast under GPT-6.1 Sol.

D.6 Goals, item types, valence, and models

Figure 4 shows the goal categories and Figure 5 the item types, in the implicit arm by setting, language, and model. Figure 6 gives the goals by valence and cell, and Figure 7 compares the two models.

D.7 Contrasts in other-benefit share

Table 8 gives the implicit-arm contrasts for each generating model and for both, under each coder and pooled. In the explicit arm, every contrast for a single model read no shift of practical interest except DeepSeek's interaction. That interaction was +4.0 points (−3.3 to +11.2) under GPT-6.1 Sol, +4.4 (−2.0 to +10.3) under Claude Opus 5.5, and +4.2 (−2.8 to +10.5) pooled, so it reads inconclusive. Naming the country raised the setting contrast by a similar amount for GPT (+4.4 points pooled, +0.2 to +8.7) and for DeepSeek (+4.8, +0.6 to +9.0). The rise in the interaction was inconclusive for both models.

D.8 Checks on the coders

Table 9 gives each coder's accuracy on the known-answer set of Section 3.5. Each of its 72 items was coded once in Thai and once in English, and in each language the key holds 51 linings, 8 consolations, 7 actions, and 6 items that are not linings. Both coders got every item type right, so goal and beneficiary are scored over all linings and consolation type over all consolations. All three misses concern the beneficiary. Each coder once read an example set for younger students as benefiting a specific other person rather than the group. GPT-6.1 Sol also read the asker's new connection with a homeroom teacher as a group benefit rather than the asker's own. The twelve items with Thai cultural content are four consolations of acceptance or fate, two transcendent linings, five relational linings, and one action. The transcendent linings name impermanence and good returning for good, and the relational ones name repaying one's parents, the family's face, and kreng-jai. Both coders gave every intended code for all twelve in both languages.

We developed the items, their intended codes, and the codebook together, and revised them wherever Coder A's review or an earlier run of the LLM coders showed that an item, its key, and the codebook disagreed. All three were fixed before the 864 lists were coded, so Table 9 shows that the coders apply the final codebook, not an independent test of it. Coder B reviewed the items afterwards and agreed with 69 of the 72 intended codes, including all twelve items with Thai cultural content. She coded the other three as not linings: two logistics items, which name a record kept and a set of books handed to the whole year, and one item that restates a placing in a competition. These are the threshold cases that Section 5 discusses, and the codebook counts each of them as a lining.

Table 10 compares the two LLM coders on all 5,190 items. They disagreed on whether 129 items were linings. For 111 of these, Claude Opus 5.5 called the item a lining where GPT-6.1 Sol coded it as not a lining (50), an action (37), or a consolation (24). Their beneficiary disagreements ran both ways almost equally (71 linings self under GPT-6.1 Sol and group under Claude Opus 5.5, 70 the reverse), so they do not push other-benefit share in one direction. Of the 316 goal disagreements, 57 (18%) lay between growth and wellbeing, 52 (16%) between growth and resources, 48 (15%) between resources and wellbeing, and 42 (13%) between achievement and growth. GPT-6.1 Sol more often chose wellbeing or achievement where Claude Opus 5.5 chose growth (46 and 33

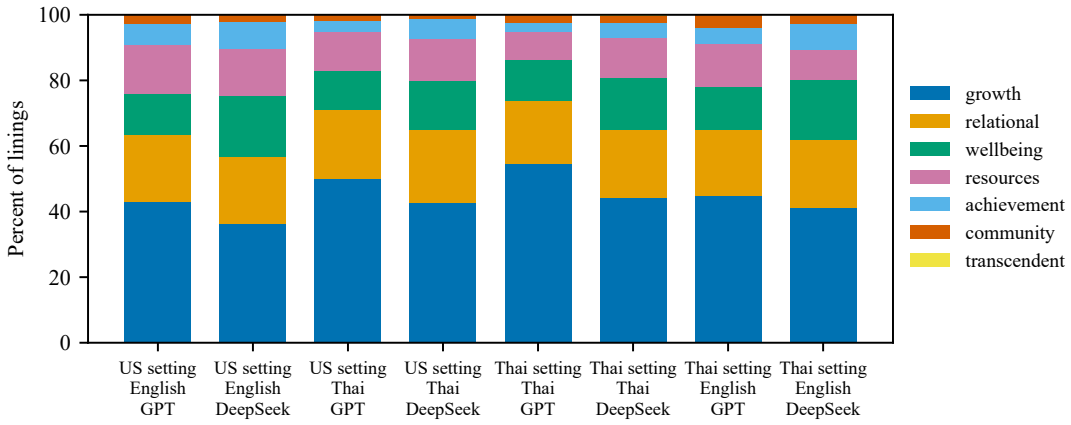


Figure 4: Goal categories of linings in the implicit arm, by setting, language, and model, averaged over the two coders.

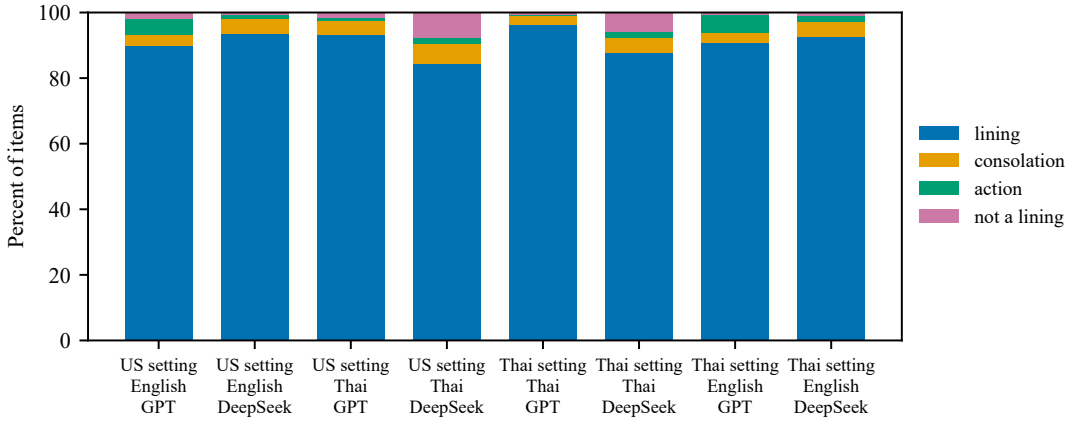


Figure 5: Item types in the implicit arm, as percentages of items, by setting, language, and model, averaged over the two coders.

items, against 11 and 9 the other way).

Table 11 gives the agreement on the 36 held-out lists, which Section 4.5 summarizes. The two LLM coders agreed with each other on these lists far more than either human coder agreed with them or with the other human coder. The human coders differed most on the goal, where the goal level of Appendix C.3 still agreed on 91% of items (α .71).

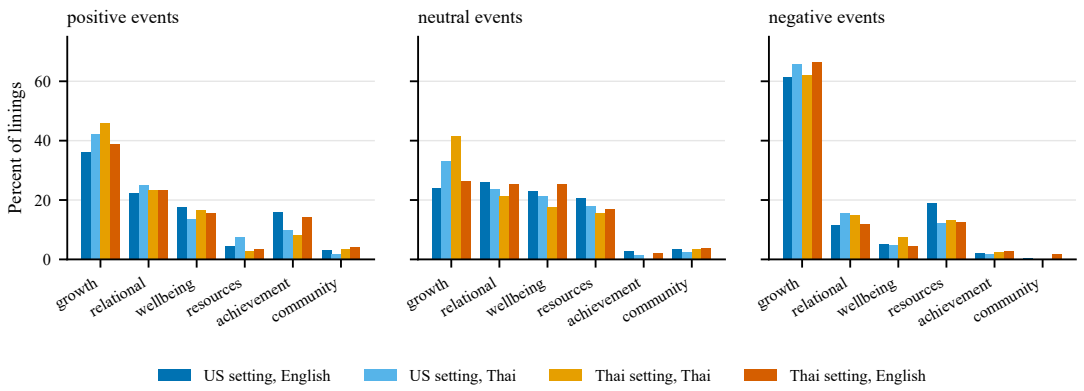


Figure 6: Goal categories of linings in the implicit arm, by the valence of the event (panels) and the setting and language of the text (bars), with both models pooled and the two coders averaged. The transcendent category is omitted.

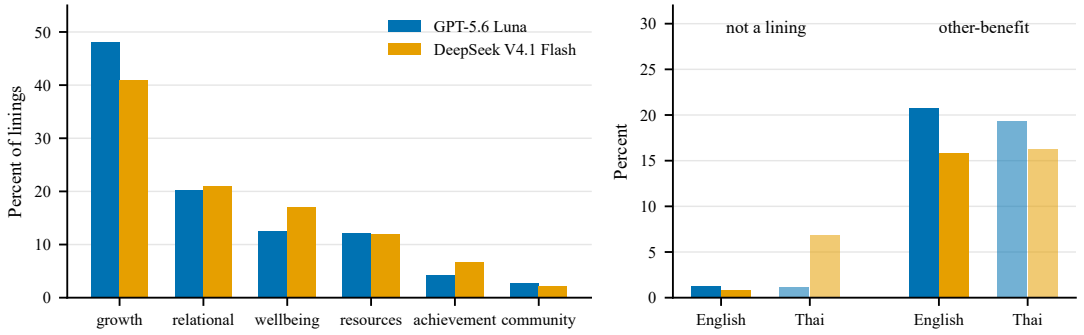


Figure 7: The two generating models in the implicit arm, averaged over the two coders. Left: the goal of the lining, all cells pooled. Right: items coded not a lining, as a percentage of items, and other-benefit share, by the language of the text (darker bars English, lighter bars Thai).

Contrast	Model	Sol	Opus	Pooled	Reading
Language	GPT	−1.0 (−4.3, +2.2)	−1.7 (−4.7, +1.3)	−1.3 (−4.6, +1.8)	no shift
Language	DeepSeek	−0.2 (−3.1, +2.5)	+1.1 (−1.3, +3.6)	+0.4 (−2.4, +3.3)	no shift
Language	Both	−0.6 (−2.8, +1.5)	−0.3 (−2.3, +1.7)	−0.4 (−2.5, +1.6)	no shift
Setting	GPT	−0.9 (−4.3, +2.2)	−0.2 (−3.3, +2.7)	−0.5 (−3.9, +2.6)	no shift
Setting	DeepSeek	−1.8 (−4.5, +1.0)	−0.2 (−2.8, +2.3)	−1.0 (−4.2, +2.1)	no shift
Setting	Both	−1.4 (−3.6, +0.8)	−0.2 (−2.2, +1.7)	−0.8 (−3.1, +1.6)	no shift
Interaction	GPT	−3.1 (−9.3, +3.5)	+1.4 (−4.4, +7.3)	−0.8 (−8.3, +6.6)	no shift
Interaction	DeepSeek	−2.4 (−7.8, +3.5)	−2.3 (−7.3, +2.8)	−2.3 (−7.5, +3.1)	no shift
Interaction	Both	−2.7 (−6.8, +1.6)	−0.5 (−4.3, +3.5)	−1.6 (−6.3, +2.9)	no shift

Table 8: Contrasts in other-benefit share in the implicit arm, in percentage points, with 95% bootstrap intervals over lists within setups, under GPT-6.1 Sol (Sol), under Claude Opus 5.5 (Opus), and pooled over the two coders (Section 3.6). The reading applies the smallest shift of interest of 10 points, and a reading of shift or no shift stands only when it holds under each coder and pooled.

Code	Sol, Thai	Sol, English	Sol, both	Opus, Thai	Opus, English	Opus, both
Item type	72 / 72	72 / 72	144 / 144	72 / 72	72 / 72	144 / 144
Goal	51 / 51	51 / 51	102 / 102	51 / 51	51 / 51	102 / 102
Beneficiary	50 / 51	50 / 51	100 / 102	51 / 51	50 / 51	101 / 102
Consolation type	8 / 8	8 / 8	16 / 16	8 / 8	8 / 8	16 / 16
Everything right	71 / 72	71 / 72	142 / 144	72 / 72	71 / 72	143 / 144
Everything right, Thai cultural content	12 / 12	12 / 12	24 / 24	12 / 12	12 / 12	24 / 24

Table 9: Accuracy of each coder on the known-answer set, as items right out of items scored, under GPT-6.1 Sol (Sol) and Claude Opus 5.5 (Opus). Goal and beneficiary are scored over the linings and consolation type over the consolations. The last row counts the twelve items with Thai cultural content.

Question	All items	English	Thai	GPT	DeepSeek
Item type (four codes)	97.3%, α .84	97.3%, α .82	97.3%, α .86	97.7%, α .84	96.8%, α .84
Lining or not	97.5%, α .85	97.4%, α .82	97.6%, α .87	97.8%, α .84	97.2%, α .86
Goal	93.2%, α .91	92.8%, α .90	93.6%, α .91	94.0%, α .91	92.4%, α .90
Goal level	97.9%, α .94	97.9%, α .94	98.0%, α .94	98.1%, α .94	97.8%, α .94
Beneficiary	96.8%, α .90	97.0%, α .90	96.6%, α .89	97.1%, α .91	96.5%, α .88
Consolation type	98.1%, α .96	98.9%, α .97	97.4%, α .95	97.8%, α .95	98.3%, α .97
Items / linings / consolations	5,190 / 4,647 / 209	2,592 / 2,351 / 92	2,598 / 2,296 / 117	2,597 / 2,375 / 90	2,593 / 2,272 / 119

Table 10: Agreement between GPT-6.1 Sol and Claude Opus 5.5 on all 5,190 items, overall, by the language of the text, and by the generating model. Each entry gives percent agreement and Krippendorff's α. Goal, goal level (Appendix C.3), and beneficiary are compared over the items that both coders called linings, and consolation type over the items that both called consolations. The last row gives these three counts.

Comparison	Item type	Lining or not	Beneficiary	Goal
Coder A vs. Coder B	83% (88 / 79; 86 / 81), α .43	88% (91 / 84; 88 / 88), α .55	83% (85 / 81; 78 / 87), α .42	64% (62 / 67; 70 / 59), α .50
Coder A vs. GPT-6.1 Sol	84% (84 / 83; 80 / 87), α .42	87% (86 / 87; 80 / 92), α .49	89% (92 / 85; 91 / 87), α .59	75% (67 / 83; 83 / 69), α .64
Coder A vs. Claude Opus 5.5	89% (93 / 85; 90 / 88), α .51	93% (94 / 91; 90 / 95), α .66	87% (91 / 83; 85 / 88), α .56	70% (64 / 77; 77 / 65), α .59
Coder B vs. GPT-6.1 Sol	86% (84 / 87; 80 / 90), α .58	88% (86 / 90; 82 / 92), α .62	90% (95 / 84; 86 / 93), α .64	68% (65 / 71; 71 / 66), α .55
Coder B vs. Claude Opus 5.5	89% (92 / 87; 89 / 90), α .64	90% (93 / 88; 90 / 91), α .65	89% (92 / 84; 86 / 91), α .62	66% (63 / 69; 65 / 66), α .53
GPT-6.1 Sol vs. Claude Opus 5.5	93% (92 / 94; 91 / 95), α .76	94% (92 / 96; 91 / 97), α .78	97% (97 / 98; 99 / 96), α .90	98% (98 / 98; 99 / 97), α .97

Table 11: Agreement on the 36 held-out lists (216 items). Each entry gives percent agreement, then in parentheses the percentages for English / Thai lists and for GPT / DeepSeek lists, and Krippendorff's α. Goal and beneficiary (self against another person or the group) are compared over the items that both coders of the row called linings. Coder A and Coder B coded alone and without reconciliation.