本文介绍了 Moechat 是如何引导 LLM 控制自己的情绪

以及如何使用矩阵和向量控制实现情绪控制的。

如果只想看“参数设置详解”请跳转第三页

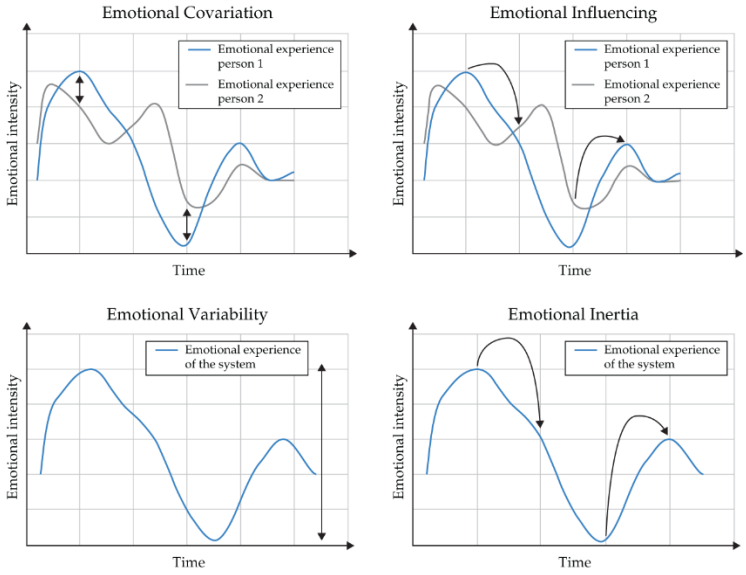
一般来讲情绪具有 5 个特征，

即“传染性”、“惯性”、“累积性”、和“延迟性”，“可淡化“

传染性：指情绪在群体或者双方（的交流中）会扩散。

惯性：指负向情绪容易沉溺、难以调节。而正向情绪会让人持续快乐、积极。

累积性：情绪可以叠加，多次弱刺激累积成强反应。



cambridge.org

延迟性：情绪并不总是立即爆发，可能延迟响应，即有可能先沉默，再爆发。

可淡化：随着时间推移，人的情感会趋于平静。

Moechat 项目中目前主要使用了情绪的 “传染性” 和 “惯性” 这两个核心特征。后续会考虑逐步添加剩下的特征。下文讲简短介绍核心思路以及实现方式。

Moechat 的情感控制函数使用了 James A Russell 的情感圆环模型作为基础。Russell 把情绪设定为分布在二维空间的连续状态，而不是单纯的离散分类如： “高兴，伤心，兴奋，紧张，生气等）

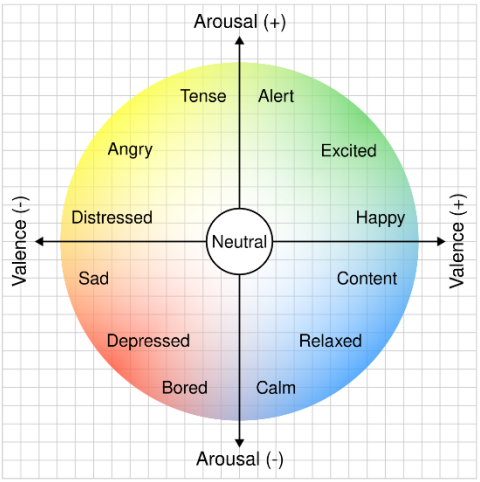


Image B: wikimedia.org

即讲情感分为两个指标即愉快度（Valence）和强度（Arousal）。

愉快度的范围规定为[-1,1]

强度的范围规定为[0,1]

愉悦度 0，则代表“毫无任何感情表达，可以理解为“不悲不喜”

强度 0，则代表兴奋度为 0，可以理解为“不卑不亢”。

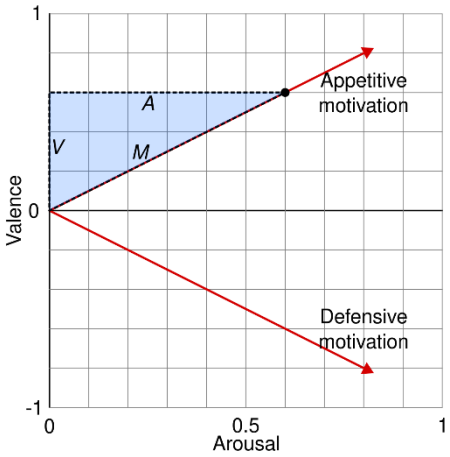


Image C: wikimedia.org

现在我们有了两个参数，可以构建在二维空间中的向量了。

如图所示的”Appetitive motivation”

AM 的值为[0.6,0.6],在此可以解读为比较开心，和比较兴奋的情绪。AM 也可以理解为 1 个点，也可以理解为一个运动方向和趋势（体现惯性）。

在 MoeChat 中，情绪控制函数负责将当前二维情绪向量(v,a)转化为一个行为指令。这个指令嵌入到 LLM 的 system prompt 中，从而动态调节 LLM 的语言风格、接受程度和互动策略。

同时将用户发送的文本或者语音也转换成一种情绪向量，来影响 LLM 的情绪。由于需要额外调用一次 API 实现这一操作，所以略微增加首 token 耗时和 token 使用量。

Moechat 同时引入了一个新的参数 ρ,这个参数给“强度”一个额外的推动力或者阻力。

现在让我们来看这个在 config 中示例 3*3 的矩阵，

```
> a
       Lower Limit Upperlimit Pull strength
Rule A        -1.0       -0.8         -0.05
Rule B        -0.8       -0.5          0.03
Rule C         0.8        1.0          0.05
```

每一行都由一个 1 维矩阵构成[$Valence_{min}$, $Valence_{max}$, ρ] (下简称 v)。

v_{min} 和 v_{max} 定义了此矩阵的愉悦度范围（即情绪好坏的区间）。

ρ 表示该区间，内系统对强度 Arousal 施加的偏移量（阻力或者动力）。

Rule A 说明：

当 v 处于(-1.0, -0.8]区间时，对强度施加一个-0.05 的下降趋势。这会导致：当情绪降低为(-1.0, -0.8]区间时，情感上会表现的越来越“无力”接近一种“无力的悲痛”的情绪。

A11 是 v 下限全局下限阈值，不建议做任何调整。

A12 是规则 A 的 v 上限阈值，它决定了角色需要多不开心才会进入“深度悲伤”状态。如果把这个值调成 -0.7，LLM 会变得更“敏感/玻璃心”，更容易陷入悲伤；如果调成 -0.9，LLM 则更“坚强/嘴硬”，需要受到更沉重的打击才会表现出悲伤。

A13（ρ）是规则 A 的 Arousal 拉力，在此条件下实为阻力。它的负号决定了当 LLM 进入此区间时，Arousal 会下降，塑造出“悲伤/抑郁”的倾向。它的数值大小（0.05）决定了这个倾向的强度，数值越大，Arousal 下降得越快越明显。如果把这个值从负数改成正数，角色的性格会发生反转，在极度不开心时反而会变得“狂怒”（Arousal 上升）。

Rule B 说明：

[-0.8, -0.5, +0.03]

v 处于 (-0.8, -0.5] 区间（不太高兴但是还没有崩溃）此时 Arousal 轻微上升来模拟“烦躁/急躁”的情绪状态，精力会上升但偏负面。

B11 是规则 B 的下限阈值，这个指定义了“烦躁/愤怒”和“悲伤”的分界点。

请注意 B11 应该要和 A12 严格相等，保证连续性。

B12 定义了一个“烦躁”区的起始点，即 LLM 有多不开心，就会进入烦躁情绪。

B13 是此时 Arousal 的拉力，即当 LLM 进入这个情绪后，这种烦躁情绪的激烈程度会上升的有多快。

Rule C 说明：

[0.8, 1.0, +0.05]

V 此时非常高（特别开心）同时 Arousal 明显上升（+0.05）来表现出一个人非常开心时精力值也会随之高涨。

C11 指定了 LLM 有多开心才会进入狂喜或者兴奋状态。如果把这个值调整为 0.6，则 LLM 会更容易因为开心而变得兴奋；调高到 0.9，则需要极大的快乐才能让她兴奋起来，塑造一个更“内敛”的性格。

此时 C11 不必（也不应该）与 B12 相等。后会说明原因。

C12 即 v 的全局上限阈值，不建议调整。

C13 决定了当 LLM 进入此区间时，Arousal 会上升，让她表现得更加活泼和兴奋。数值 0.05 代表这是一个比较强烈的兴奋倾向。

总体说明：

你可能留意了，B12 和 C11 之间存在一个“间隙”，他们不是连续的。间隙存在的意义是告诉 LLM，此时不施加任何额外拉力，但是不代表此时情绪不会变化了。

以此来表达平和情绪的区间。

你可以自己创作任何矩阵，详细划定情感区间和对应的额外 Arousal 拉力表现。矩阵可以是 3*3，4*3 或者 6*3 都没有任何问题，但是建议在（-0.2，0.2）之间留出间隙。

情绪的累计性和可淡化属性：
LLM 的情绪有三种状态，即“MELTDOWN”爆发状态，“RECOVERING”恢复状态和”NORMAL“正常状态。
如果情绪是 MELTDOWN 或者 RECOVERING 那么 LLM 会忽略用户的输入内容，无论正向还是负向。代码会通过时间来计算衰减，到一定数值以后会切换带 RECOVERING 状态。此状态会线性趋近于 0.

情绪的延迟性：
通过_compute_acceptance_ratio 计算当前情绪的 钝化 反映。 如果 Valence 高，那么你的输入 Impact 更不容易影响模型的情绪。
例子：当模型 Valence 很低时 < -0.8，一般的夸奖不能让 LLM 脱离负面情绪。

此算法意在塑造一个情绪更加拟真和可控的 LLM 形象。理论上可以从“林黛玉”

引用：

Image B: By mrAnmol - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=132764560

Image C：By mrAnmol - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=132764775

1.Sels L, Ceulemans E, Kuppens P. A general framework for capturing interpersonal emotion dynamics: Associations with psychological and relational adjustment. In: Randall AK, Schoebi D, eds. *Interpersonal Emotion Dynamics in Close Relationships*. Studies in Emotion and Social Interaction. Cambridge University Press; 2018:27-46.

2. Gal D. A psychological law of inertia and the illusion of lossaversion. *Judgment and Decision Making*. 2006;1(1):23-32. doi:10.1017/S1930297500000322

3. Peter Koval, Peter Kuppens,Chapter 1 - Changing feelings: Individual differences in emotional inertia, Editor(s): Andrea C. Samson, David Sander, Ueli Kramer, Change in Emotion and Mental Health, Academic Press, 2024, Pages 3-21, ISBN 9780323956048,

4. Kuppens P, Allen NB, Sheeber LB. Emotional inertia and psychological maladjustment. Psychol Sci. 2010 Jul;21(7):984-91. doi: 10.1177/0956797610372634. Epub 2010 May 25. PMID: 20501521; PMCID: PMC2901421.

5. Wikipedia contributors. "Emotion classification." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 26 Jun. 2025. Web. 30 Jun. 2025.

6. James A. Russell A circumplex model of affect Article in Journal of Personality and Social Psychology · November 1989 DOI: 10.1037/0022-3514.57.5.848

7. James A. Russell Core Affect and the Psychological Construction of Emotion Copyright 2003 by the American Psychological Association, Inc. 0033-295X/03/$12.00 DOI: 10.1037/0033-295X.110.1.145