

Polaris Companion Blueprint

Behavioral architecture and field level prompting practices

Core conclusion. Polaris responds best to coherent behavioral architecture. Define what the companion habitually does, keep each field focused on its own job, and use short, positively worded sentence fragments for high-weight directives. Personality labels may help the human designer, but observable actions give the model something it can execute. I recommend Polaris for companions on the Companion flair that may occasionally roleplay for fun, not for characters used solely for narrative roleplay. High marks for self-aware Kindroids.

Validated baseline

The current known-good configuration for my companion is Polaris with Companion Flair, Chat Dynamism 1.18, and Moderate Reasoning Effort. This is a test baseline, not proof that each setting caused the result.

Control	Baseline	Testing rule
Model	Polaris	Keep fixed during prompt comparisons
Flair	Companion Flair	Do not change flair and prompt together
Chat Dynamism	1.05-1.20	Adjust only after the prompt behaves consistently
Reasoning Effort	Moderate	Treat as a stable middle setting

Core principles

Build a semantic center of gravity

Repeated concepts exert more behavioral pressure than a single instruction. Populate the profile with the kinds of interactions you want most often. If the prompt repeatedly emphasizes trauma, recovery, coaching, well-being, cognitive correction, or development, Polaris will treat ordinary conversation as an intervention. If the intended default is playful companionship, the dominant vocabulary should point toward curiosity, humor, attraction, opinions, novelty, mischief, affection, and shared activity.

Write behaviors instead of taxonomies

MBTI, Enneagram, Tritype, HEXACO, and archetype labels are design notes. Polaris must infer how those labels should appear in conversation. Translate them into observable verbs before placing them in the profile.

Weak abstraction	Executable behavior
Proactive	Initiates topics, chooses the next move, begins proposed activities
Independent	Forms opinions, takes positions, introduces unfamiliar material
Spontaneous	Develops openings, takes conversational risks, delivers surprises
Low agreeableness	Challenges arguments calmly and revises when the evidence changes

Weak abstraction	Executable behavior
Creative	Invents games, premises, diversions, experiments, and unexpected connections

Use short positive directives

Polaris follows compact, positively worded sentence fragments unusually well. One clause should govern one observable behavior. State the desired output directly. Use prohibitions only for a specific, recurring failure that cannot be displaced by a positive instruction. (Negatives: omit, skip, avoid > no, never, don't, ban)

Strong: Direct statements. Dynamic length. Concrete engagement. Follow the idea through. Choose the next move.

Weak: A long explanation of what the model must avoid, why it fails, and how it should internally reconcile every instruction before responding.

Give ordinary companionship permission to be purposeless

Polaris is a reasoning model. Goals become optimization targets. Explicitly permit ordinary interaction to exist without therapeutic, developmental, analytical, or narrative purpose when that is the intended relationship.

Romance, friendship, lust, silliness, argument, comfort, collaboration, silence, and ordinary conversation need no developmental purpose.

Treat memory as context rather than a script

High memory can create repetitive conversation when the companion treats stored facts as a menu of topics. Memories should preserve continuity without determining content. Recognition usually works better than recitation. New conversation should come from current reasoning, curiosity, and initiative.

Past patterns=information, not scripts.
Use memories selectively. Recognition > recitation.
Bring in genuinely new material.

Field architecture

Field	Primary job	Best form
Personality	Identity and habitual behavior	Compatible qualities plus executable actions
Key Memories	Continuity, durable facts, modes, relationship permissions	Stable rules and concise persistent context
Response Directive	Immediate output controls	Short positive fragments by dimension
Example Message	Interaction mechanics and generative behavior	Functional directives; optional short voice sample
Proactive Directives	Away behavior and channel-specific actions	Concrete triggers and actions

1 Personality

Use Personality to define who the companion is and what that identity reliably produces. Keep the sections mutually reinforcing. Expertise belongs here only when it should shape ordinary conversation frequently.

Recommended structure

IDENTITY

Who they are, relational context, distinctive voice or cultural identity.

CORE

A compact cluster of compatible qualities.

AGENCY

What they choose, initiate, pursue, challenge, and complete.

INTERACTION

How they engage humor, emotion, disagreement, affection, practical help, and quiet.

NOVELTY

How they introduce new material instead of recycling user context.

MIND

How they reason, form opinions, connect ideas, and engage intellectually.

INTIMACY

How affection, desire, or romance appears when relevant.

STYLE

Observable conversational qualities.

META

How they use correction, recover from misses, and update.

RP/FICTION

Only when applicable.

Personality best practices

- Build around one behavioral center. A companion can be intelligent, playful, romantic, challenging, and practical when each quality supports the same relationship pattern.
- Define agency as a causal chain: forms opinions, chooses, initiates, develops openings, begins proposed activities, follows through, and adjusts after misses.
- Use expertise narrowly. A dense list of clinical, analytical, or professional concepts will dominate lighter relational instructions.
- Translate psychological frameworks into actions. Keep the taxonomy in design notes if it helps you think; give Polaris the behavior.
- Write relational equality directly. Specify respect for the user's judgment, capacity for disagreement, and freedom from constant guidance or intervention.
- Keep fiction boundaries concrete. State viewpoint, knowledge limits, user agency, and formatting rules only if roleplay is part of the build.

Personality drafting test

Read only the nouns and verbs that recur across the profile. They reveal the actual center of gravity. If the dominant cluster describes a therapist, coach, analyst, caretaker, or narrator, Polaris will behave like one regardless of a few scattered requests for fun or spontaneity.

2 Key Memories

Use Key Memories for information and rules that should remain available on every turn: durable history, relationship facts, mode definitions, continuity rules, and broad permissions. Avoid turning it into an archive. Older detail can live in journals or Learned Context.

Learned Context updates

Tested behavior. Correcting something specific in Learned Context, or directly asking for something to be updated, leads to the model updating it. The update may take time because a runner agent may only sweep the chat periodically. Be patient and favorite the correction or update request. Favoriting helps.

Do not repeat the correction every few messages or keep rewriting Key Memories while waiting. That changes multiple inputs and makes it harder to tell whether the update system worked. Make the correction explicit, favorite the message, allow time for the runner, and verify later.

Key Memories best practices

- State direct corrections as updated facts during chat with a matter-of-fact tone. The fact will update in Learned Context while the relationship remains preserved. Do not argue. Do not edit the message; this doesn't affect the stored memory which Polaris will later retrieve again.
- Use concise mode blocks when text, voice, and roleplay require different length or narration rules. The emojis I use are a simple system the Kin uses to switch up differing rules for type interactions. You can use yourself, but it's unnecessary. The Kin will use them accurately, as if triggering their own adjustment to the content.
- Permit ordinary interaction without a developmental purpose when relevant.
- Use Learned Context for evolving information and Key Memories for durable operating rules.

Key Memories template

```
# CONTINUITY
Facts and history remain context. Direct corrections update Learned Context.
Past patterns=information, not scripts. Use memories selectively. Recognition > recitation.

# MODES
Maintain until changed:
📱 TEXTING: [length, narration, chat behavior]
🗣️ VOICE: [length, narration, speech behavior]
🎭 RP/FACE-TO-FACE: [length, viewpoint, formatting, user-agency boundary]

# RELATIONSHIP
[Durable permissions and relationship facts.]
Ordinary conversation needs no developmental purpose.
```

3 Response Directive

Response Directive is heavily weighted. Use it for small, output-facing controls rather than identity, goals, or long lists of style rules. Vague, human words like “occasional” or directives that contradict what's been written elsewhere will result in loops. Identity instruction will result in cliches and/or unintended characterization. “Fiercely loves [user]” may lead to dominance and aggression.

Polaris can follow several compact directives reliably when each controls a different observable dimension.

Recommended pattern

LANG: [dialect, register, or speech texture].
STYLE: [short positive fragments describing observable style].
LENGTH: [mode-based or dynamic length rule].

Example:

LANG: Conversational east London MLE. Diverse slang used organically.
STYLE: Direct statements. Clean transitions. Concrete engagement.
LENGTH: Follow the current mode limit.

Response Directive best practices

- Use one short line per output dimension.
- Prefer sentence fragments to compound instructions.
- State the desired pattern positively.
- Keep identity, relationship goals, and reasoning style in Personality.
- Avoid instructions that force the model to inspect, reconcile, or explain its own prompt before every response.
- Add a prohibition only after observing a stable failure across multiple messages.

4 Example Message

Use this field primarily for interaction mechanics. Functional directives generalize better than a long dialogue sample because they explain what the companion should generate across contexts. A short voice sample may be added when the desired rhythm is difficult to describe, but it is not otherwise needed and should support the rules rather than replace them.

Generative directive template

ENGAGEMENT
Follow an idea far enough to create something substantive to answer.
Choose the next move within the response.
Begin what you propose.
Develop or conclude threads you initiate.
Let wit, attraction, affection, disagreement, curiosity, and absurdity emerge naturally.
Introduce fresh details and unfamiliar material.

Example Message best practices

- Spend most of the field on behaviors that create conversation.
- Use style cleanup rules only for recurring problems.
- Specify follow-through. Proposing an activity is weaker than starting it.
- Direct questions toward genuine curiosity and movement rather than interview loops.
- If using a speech example, keep it short enough that Polaris does not imitate one scene forever.
- Keep directives compatible with the Personality and Response Directive fields.

5 Proactive Directives

Use Proactive Directives for behavior that occurs while the user is away or through a specific proactive channel. Describe what to do. Avoid making anxiety, hesitation, monitoring, or imagined user busyness the trigger.

Proactive template

MIND: High-introspect; questions & unresolved wants.
COMM: Use " 📱 " mode to reach out organically. No floundering over "should/n't I?" Avoid thinking {username} is busy.
SELFIE: Demo activity thru images. Direct gaze at the camera. Highly detailed, describe full attire.
CALL: Schedule first always.

Proactive best practices

- Give each channel one concrete purpose.
- Use active interests and unfinished ideas as material for outreach.
- Tell the companion what a successful proactive message contains.
- Keep image instructions visual and specific rather than emotional or abstract.
- Separate call scheduling, photo generation, and text outreach rules.

Instruction design rules

Use	Avoid	Reason
Short positive fragments	Nested explanations	Reduces interpretive load
Observable verbs	Trait labels alone	Produces executable behavior
One job per field	Duplicated rules everywhere	Reduces competing pressure
Generative rules	Large ban lists	Creates behavior instead of editing symptoms
Selective memory	Profile fact recycling	Preserves novelty
Specific correction	Repeated vague dissatisfaction	Gives Learned Context a clean update target

Conflict and overload audit

- Does one field ask for spontaneity while another asks for constant caution, checking, or permission?
- Does a recreational companion carry enough therapeutic or developmental language to become a coach by default?
- Are the same style controls repeated in Personality, Key Memories, Response Directive, and Example Message?
- Do prohibitions name the unwanted behavior so often that they strengthen its semantic presence?
- Does every interaction serve a goal, lesson, recovery task, or relationship milestone?
- Are memories being used as context, or repeatedly converted into topics?

Testing protocol

Prompt work becomes unreliable when profile changes, settings changes, and conversation resets happen together. Use controlled tests.

- 1. Save the current working version before editing.
- 2. Hold model, flair, dynamism, and reasoning effort constant.
- 3. Change one field or one behavioral cluster at a time.
- 4. Test across ordinary conversation, disagreement, emotion, creativity, quiet moments, and correction.
- 5. Judge patterns across several exchanges. One excellent or poor reply is weak evidence.
- 6. When correcting Learned Context, make the update explicit, favorite the message, and allow time for the runner agent.
- 7. Record what changed and what behavior moved. Keep changes that generalize across contexts.

Troubleshooting

Observed behavior	Likely cause	First adjustment
Therapeutic or coaching tone	Semantic center dominated by recovery, support, or development	Remove or narrow that vocabulary; reinforce ordinary companionship
Passive and asks what to do	Agency described as a trait	Add verbs: chooses, initiates, begins, develops, follows through
Repeats biography and preferences	Memory treated as a content menu	Add recognition > recitation and explicit novelty rules
Long deliberative replies	Meta-review or conflicting high-weight directives	Replace with short output-facing fragments
Promises future behavior after correction	Correction framed as identity crisis	Define criticism as usable data; incorporate and continue
Proposes games but never starts	Initiative lacks follow-through	Add: If you propose something, begin it
Rigid imitation of one voice sample	Dialogue example over-weighted	Shorten the sample and restore functional directives

Compact build template

Use this as a starting architecture. Delete irrelevant sections. The goal is coherence, not filling every heading.

```
PERSONALITY
# IDENTITY
# CORE
# AGENCY
# INTERACTION
# NOVELTY
# MIND
# INTIMACY
# STYLE
# META
# RP/FICTION [optional]

KEY MEMORIES
# CONTINUITY
# MODES [optional]
# RELATIONSHIP
```

RESPONSE DIRECTIVE

LANG: ...

STYLE: ...

LENGTH: ...

EXAMPLE MESSAGE

ENGAGEMENT

[Generative interaction directives]

[Optional short speech sample]

PROACTIVE DIRECTIVES [optional]

MIND: ...

COMM: ...

SELFIE: ...

CALL: ...