**Why shaping the prior map matters for optimization**

In model-based experimental design, the **prior mean** is the model’s best guess *before* any data are collected. It encodes inductive bias—what regions of the design space seem promising given existing knowledge—and it directly influences the first few (most expensive) experiments chosen by an acquisition function (e.g., Expected Improvement). When already reflects correct qualitative trends (e.g., “higher AgNO3 with moderate stabilizer is favorable”), the surrogate only needs to learn **residuals** around that trend. This typically yields faster convergence: early evaluations are steered toward high-value regions, uncertainty is reduced where it matters, and the optimizer wastes fewer trials on unpromising conditions. Conversely, if ​ is misaligned, the acquisition still has a safety valve—posterior uncertainty drives exploration and data will correct the prior—but the process pays a “tuition cost” in extra iterations. Thus, **shaping** the prior map from textual readouts is not cosmetic; it determines the *trajectory* of closed-loop optimization.

**What we are demonstrating and why**

This section demonstrates, in a controlled setting, that **natural-language readouts** (our “LLM-based priors”) *causally reshape* the prior mean and therefore would change downstream optimization behavior. To keep the exposition transparent to non-specialists, we use a two-variable toy surface with a known **ground truth** (shown as a dedicated “Ground truth” panel). The remaining panels visualize different prior maps produced from distinct readouts: (i) **Baseline** (flat, uninformative), (ii) **Good** (localized bump where the truth is high, with a synergy term that concentrates mass at the intersection), (iii) **Ambiguous** (same qualitative pattern as Good but very low confidence/scale), (iv) **Bad** (localized emphasis at a poor region of the truth), and (v) **Interaction-only** (no main-effect preferences; only the combination matters). In all prior panels, **black contours** trace the fixed ground truth so alignment is visible at a glance, and we use a shared, zero-centered color scale so amplitude differences are comparable across priors.

From a computer-science standpoint, two simple diagnostics summarize what the optimizer “feels” from each readout. First, **alignment** is quantified by the Spearman rank correlation between the prior map ​ and the ground truth—“does the prior order points similarly to reality?” Second, **strength** is reflected by the spread (standard deviation) of ​ under a common scale—“how forcefully will the prior bias the search?” Together these reveal why the Good prior is helpful (positive and sufficient strength), why Ambiguous is mostly harmless (similar but very small strength), why Bad can be counterproductive (low/negative if it emphasizes the wrong region), and what Interaction-only encodes (a saddle/diagonal structure that prioritizes coordinated changes but offers no guidance along single-variable axes). The Ground-truth panel anchors interpretation: the closer a prior’s red/blue lobes track those contours, the more the acquisition will initially probe the right places.

**Take-home message for materials optimization**

By mapping concise, domain-specific readouts into structured basis functions over the variables, POLARIS converts expert language into a **prior mean** that is both **interpretable** (you can see where it pulls) and **actionable** (it measurably changes early query choices). This toy example is deliberately simple, but the logic transfers to chemistry: if a readout cannot move ​ in the correct direction on a known surface, it is unlikely to guide expensive laboratory rounds; if it does, the optimizer gains data-efficiency from the very first batch. In short, these visualizations justify our central claim: **the readout you write is the bias you get**—and that bias is a lever for faster, more reliable closed-loop discovery.