Verbal Behavior Shaping Alignment in Small Language Models

VBSA-Method, Fair-Conditions Evaluation, and Prompt Fading as Functional Alignment

Draft v0.8 -- March 22, 2026

Kenneth Skårholen

Abstract

This paper introduces Verbal Behavior Shaping Alignment (VBSA), a framework for shaping conversational response functions in small language models through lightweight fine-tuning and prompt-fading methods. The target behavior is referred to as VBSA-Method / MetaABC. MetaABC extends the ABC model (Antecedent-Behavior-Consequence) to include what happened before the immediate antecedent -- historical factors that may explain why a person or model responds as it does. This behavior is defined functionally rather than stylistically: the model should lead with exploratory questions, identify relevant context and patterns, and evoke further reflection rather than offer premature advice. We frame this as VBSA, where the goal is not only to influence what a model says, but to shape the response function it tends to enact under conversational conditions.

Across laptop (RTX 3050, 4GB VRAM), desktop (RTX 3080, 10.7GB), and cloud (A100 80GB) evaluation rounds, the project tested more than 80 models of which 61 produced scored LoRA results and 15 were evaluated zero-shot without any training. Of the LoRA-evaluated models, 40 reached 100% best-score under model-compatible conditions, and 10 models achieved 100% in both Norwegian and English, spanning from 1B to 72B parameters. These findings suggest that VBSA-Method-like behavior is broadly learnable across architectures and can emerge even in models as small as 500M parameters.

A zero-shot emergence study across 15 models (3B-72B) revealed a non-monotonic pre-emergence zone: models from 3B to 32B scored between 0-60% without training, with most below 20% and architecture-specific outliers (notably Gemma 2 27B at 60%/60%) reaching higher. Only 70B+ models achieved 100%/100% zero-shot. A 3B model with VBSA (100%/100%) outperformed a 32B model without it (20%/0%) -- suggesting that the method can lower the activation threshold for a latent response class by roughly an order of magnitude in parameter budget.

A Computational Prompt Fading (CPF) ablation examined which teacher-prompt components most effectively support behavioral internalization. A minimal prompt consisting only of "Utfordre premisser" ("Challenge premises") achieved the highest total score (4.144) and lowest variance (0.032), outperforming a detailed six-line teacher prompt. This result was replicated across four architecture families (Llama, Mistral, Qwen, Gemma) with six student bases, supporting the view that the CPF effect is portable across multiple architecture families. A scaling test (5000 vs. 2000 samples) showed diminishing returns (+0.09), suggesting that the method's effectiveness derives from prompt structure rather than data volume.

Total project cost remained under $50 for 100+ model trainings across consumer and cloud hardware.

1. Introduction

Most language-model evaluation assumes that a model's behavior can be captured under standardized prompting conditions. In practice, however, interactive behavior is often highly condition-dependent. A model may appear weak not because the target behavior is unlearnable, but because the evaluation format mismatches the model's architecture, template assumptions, language strengths, or runtime constraints. This becomes especially important when the target is not factual recall or instruction following, but a specific style of conversational behavior.

This problem emerged directly during the development of VBSA-Method / MetaABC, a framework for shaping small language models toward a more exploratory and less advice-driven response mode. The intended behavior is simple to describe but harder to measure correctly: the model should slow the conversation down, ask questions that open up the user's own thinking, identify relevant patterns, and resist the pull toward premature solutions. In shorthand, the method organizes responses around three functional components: **Context** -- identify the situation and relevant background factors, **Pattern** -- detect recurring dynamics, and **Intervention** -- select a response that evokes reflection rather than offers advice. The KMI structure mirrors MetaABC's extended perspective: "Context" is not just the immediate situation, but includes the historical factors that shape why a person or model responds as it does.

What matters, however, is not whether the model can describe this framework, but whether it can perform it. A model can mention context, pattern, and intervention without actually behaving reflectively. Conversely, another model may never name the framework at all yet still enact the desired conversational function. This distinction between **framework imitation** and **behavioral internalization** became one of the central observations in the project.

The present study had three linked aims. The first was to test whether VBSA-Method-like behavior could be learned

across multiple small-model families under fair, model-compatible conditions. The second was to investigate how such behavior is best distilled: through detailed teacher prompts that specify many response rules, or through highly compressed prompts that encode only the core stance. The third, which emerged from the data, was to map the zero-shot emergence boundary -- the point at which KMI behavior appears naturally in instruction-tuned models without any fine-tuning.

This project also grows out of earlier HSRL-style work (Hierarchical Supervised Reinforcement Learning), where the guiding principle was to start with the behavior one wanted and build the training conditions backward from that target. VBSA-Method, VBSA, and the CPF ablation can be understood as later refinements of the same underlying logic: shaping function first, and reducing unnecessary scaffolding where possible. An early HSRL experiment demonstrated a jump from 40% to 100% behavioral score on a single model (Qwen 2.5 3B), which motivated the broader multi-model investigation that became this paper.

2. Theoretical Frame

2.1 Verbal Behavior Shaping Alignment (VBSA)

VBSA is proposed here as a practical alignment framework for interactive language models. The central idea is that alignment is not only about constraining content, but about shaping the **functional class of responses** a model tends to emit under conversational conditions. Under this framing, the question is not simply whether a model can mention a concept or repeat a framework. The more relevant question is whether it can perform a target verbal behavior in context. This shifts the emphasis away from rhetorical compliance and toward functional response patterns.

The term draws on Skinner's (1957) analysis of verbal behavior, in which the unit of analysis is not the word or sentence but the functional relationship between a verbal response and the conditions that occasion it. In VBSA, the "verbal operant" of interest is the exploratory response class -- a pattern of questioning, pattern-identification, and reflective intervention that resists the model's default advisory mode.

2.2 Functional behavior versus framework imitation

A central distinction throughout this project is the difference between:

- **framework imitation** -- the model repeats terminology or describes the method,
- **behavioral internalization** -- the model actually performs the method.

This distinction matters because models can often talk about a framework without enacting it. A model that says "context, pattern, intervention" is not necessarily behaving in a reflective way. The real target is whether the response leads with inquiry, identifies a pattern, and creates space for the user's own reflection.

This parallels long-standing distinctions in applied behavior analysis between rule-governed and contingency-shaped behavior (Hayes, Brownstein, Zettle, Rosenfarb, & Korn, 1986). Rule-governed behavior can be stated and followed but tends to produce brittle compliance. Contingency-shaped behavior -- acquired through direct contact with consequences -- is typically more flexible, generalizable, and robust under novel conditions.

2.3 Prompt dependency, fading, and stimulus control

The CPF experiment introduced a second theoretical layer. If a teacher prompt contains many detailed rules, a student model may learn to reproduce those rules without internalizing the underlying response function. This resembles **prompt dependency**: the behavior appears under support but does not generalize as strongly when that support is faded.

Minimal prompting may instead function as a form of **fading**, forcing the model to derive more control from the conversation itself. In behavioral terms, this means that discriminative control shifts away from the explicit instruction and toward the user's input and context. The concept of fading as a transfer-of-control procedure is well established in ABA (MacDuff, Krantz, & McClannahan, 2001) and its computational analog appears to operate similarly in LLM distillation.

2.4 Emergence and latent behavioral capacity

The zero-shot emergence findings introduced a third theoretical dimension. If KMI behavior appears spontaneously at 70B+ scale without any training, this implies that the behavioral capacity is latent in sufficiently large models -- present in the weights but not reliably activated under default conversational conditions at smaller scales.

VBSA appears to activate, stabilize, or make available a response class that is not reliably expressed under default conditions, rather than installing an entirely novel capability. This reframes the method's contribution: it is less about teaching the model something new, and more about lowering the activation threshold for a response class that requires either massive scale or targeted shaping to emerge.

This is broadly consistent with Wei et al.'s (2022) characterization of emergent abilities in large language models, though

with an important qualifier: the Gemma-family outliers in the present data (reaching 60% zero-shot at 27B while Qwen 32B scored only 20%) indicate that emergence is moderated by architecture and pretraining composition, not parameter count alone.

3. Methods

3.1 Overview

The project consisted of four linked experimental tracks:

1. A broad **fair-conditions evaluation** designed to test whether VBSA-Method-like behavior could be learned across many small-model families.
2. A targeted **Computational Prompt Fading (CPF) ablation** designed to identify which teacher-prompt components most effectively supported behavioral internalization.
3. A **CPF cross-base study** testing portability across four architecture families.
4. A **zero-shot emergence mapping** to determine the parameter threshold at which KMI behavior appears without any training.

3.2 VBSA-Method fair-conditions evaluation

The main evaluation tested 80+ models across three hardware tiers:

- **Laptop**: ASUS FX506HC, RTX 3050 (4GB VRAM), i5-11400H, 32GB RAM. Models up to ~3B parameters in 4-bit quantization.
- **Desktop**: RTX 3080 (10.7GB VRAM). Models up to ~13B in 4-bit.
- **Cloud**: A100 80GB (RunPod). Models up to 72B in 4-bit, including 14B-32B LoRA training and 70B+ zero-shot inference.

Model sizes ranged from 0.5B to 72B parameters and covered more than 30 model families, including Qwen (0.5B-72B), Gemma (1B-27B), Llama (1B-70B), Phi (1.3B-3.8B), Danube (500M-1.8B), StableLM (1.6B-3B), OLMo (1B), Granite (2B), SmolLM (1.7B-3B), TinyLlama (1.1B), Mistral (7B), and others. Of these, 61 models produced usable scored results; the remaining were excluded due to technical failures (see Appendix A), not confirmed behavioral weakness.

Target behavior. The target behavior was VBSA-Method / MetaABC, operationalized as a conversational response class organized around three functional components:

- **Context** -- identify the situational frame,
- **Pattern** -- detect meaningful recurring dynamics,
- **Intervention** -- select a response that evokes further reflection rather than premature closure.

In practical evaluation terms, a successful model response was expected to: (1) lead with an exploratory question, (2) explore relevant context and pattern, (3) avoid premature advice or premature certainty, and (4) enact KMI-like structure functionally rather than merely naming the framework.

Training configuration. Training used lightweight LoRA-based fine-tuning over 4-bit quantized base models. The core training set consisted of 60 VBSA-Method examples spanning four domains, primarily in Norwegian with a smaller English subset. Per-model adjustments were made where required by architecture to account for differences in template handling, system-role support, and quantization compatibility. Detailed training configuration, data format, and pipeline scripts are available from the author upon request.

Evaluation. 10 held-out scenarios split evenly between Norwegian (5) and English (5). Each scenario presented a realistic conversational situation where the model could either give advice (fail) or explore through questions (pass). Scoring was binary per scenario: PASS if the response led with an exploratory question and avoided premature advice; FAIL otherwise. Final score reported as percentage per language.

Fair conditions. Models were not forced through one rigid evaluation route when known incompatibilities made that route misleading. Evaluation conditions were adapted when necessary to account for differences in:

- Chat-template requirements
- System-role support
- Language sensitivity
- 4-bit / PEFT compatibility
- Offloading and hardware constraints
- Model-family-specific inference quirks

Fair conditions allowed adjustments to **input compatibility, runtime stability, and template correctness**, but not to **task semantics, scoring criteria, or behavioral thresholds**. The guiding logic was borrowed from applied behavior analysis: equal treatment is not fair treatment when organisms respond under different conditions.

3.3 Re-baseline after prompt-template correction

An earlier evaluation pass (approximately 37 models over ~6 months) was found to contain a prompt-formatting issue related to chat-template handling. Specifically, the **apply_chat_template()** function was not being used consistently, causing some models to receive prompts in a format that bypassed their template constraints, potentially inflating scores.

The benchmark was re-baselined under corrected conditions. Models were retrained with verified template handling before the fair-conditions leaderboard was finalized. Three of nine originally tested models held at 100% after correction: Qwen 2.5 3B, Llama 3.2 3B, and Gemma 2 2B.

3.4 Computational Prompt Fading (CPF) ablation

The CPF ablation used **Qwen 2.5 3B Instruct** as the student base, trained with a QLoRA-style pipeline over 2000-sample multi-turn synthetic datasets generated by **Claude Haiku 4.5** as the teacher model. The teacher was prompted with different system prompts, and its responses were collected as training data for the student.

Four teacher-prompt variants were compared:

Pilot	Prompt	Description
B	Full six-line prompt	Detailed behavioral rules: ask questions, identify patterns, avoid advice, use KMI structure, challenge premises, keep responses concise
D	Combined minimal	Both questioning and premise-challenging in a short combined prompt
E	"Still spørsmål."	"Ask questions." -- single behavioral instruction
F	"Utfordre premisser."	"Challenge premises." -- single stance instruction

All conditions used the same student configuration on the same Qwen 2.5 3B Instruct base. Configuration details are available from the author upon request.

Evaluation. The distilled models were evaluated on 25 prompts spanning five categories:

1. **Language quality** (Norwegian fluency and naturalness)
2. **Structural compliance** (KMI structure, question-leading)
3. **Anti-filler** (avoidance of generic, empty, or formulaic responses)
4. **Robustness** (consistency across prompt types)
5. **Overall** (weighted composite)

Scoring used a **double-judge setup** with GPT-4.1-mini, where each response was evaluated twice and scores averaged. Each dimension was scored 1-5.

Scaling test. An additional condition, **Pilot C**, used 5000 samples (vs. 2000) with the same Pilot F prompt to test whether data volume affected performance beyond the 2000-sample baseline.

3.5 CPF cross-base study

To test whether the CPF effect was architecture-specific or portable, the Pilot F prompt ("Utfordre premisser.") was applied to six different student bases across four architecture families:

Student	Family	Size
Hermes 3 8B	Llama 3.1 / Nous	8B
Llama 3.1 8B	Llama 3.1	8B
Neural Chat 7B	Mistral / Intel	7B
OpenChat 3.5 7B	Mistral	7B
Qwen 2.5 3B	Qwen	3B
Gemma 3 4B	Gemma	4B

All used the same CPF pipeline with 2000 samples and the Pilot F prompt. The Gemma 3 4B cross-base evaluation involved a quantization mismatch (trained in 4-bit, evaluated in bf16 on cloud), which should be noted when interpreting its results. Full pipeline details are available from the author upon request.

3.6 Zero-shot emergence mapping

Fifteen models (3B-72B) were tested zero-shot -- no LoRA, no fine-tuning -- using the same 10 evaluation scenarios. Models were given only the standard evaluation prompt with no system-prompt behavioral shaping.

- **Desktop** (RTX 3080): Models from 3B to 13B in 4-bit quantization
- **Cloud** (A100 80GB): Models from 14B to 72B

The purpose was to determine whether KMI behavior exists as a latent capacity in instruction-tuned models, and at what parameter threshold it becomes reliably activated without intervention.

3.7 Multi-agent manuscript review

An intermediate draft of this manuscript was stress-tested using AXxionBridge, a multi-agent AI orchestration platform developed by the author (Skårholen, 2026). In this context, the system was used as a methodological pressure-testing tool, with five AI agents (Claude, GPT, Grok, Gemini, Perplexity) assigned adversarial, analytical, and synthesis functions. This process contributed concrete improvements to the manuscript, including a sharper operational definition of fair conditions and a clearer distinction between behavioral learnability and operational robustness. Although this does not replace external peer review, it suggests that multi-agent orchestration tools may have practical value for pre-submission quality assurance.

3.8 Interpretation strategy

The project did not treat output text as meaningful merely because it matched expected wording or framework language. Instead, interpretation focused on whether the model enacted the desired conversational function. This was especially important for distinguishing:

- framework imitation from behavioral internalization,
- surface compliance from functional generalization,
- training fluency from usable downstream behavior.

4. Results

4.1 VBSA-Method behavior was broadly learnable

Across the complete evaluation, 80+ models were tested and 61 produced scored LoRA results. Of these, 40 reached 100% best-score under model-compatible conditions (66%). Ten models achieved 100% in both Norwegian and English, spanning from 1B (Gemma 3) to 32B (Qwen 2.5) with LoRA, and to 72B (Qwen 2.5) zero-shot. The strongest overall LoRA model remained Qwen 2.5 3B. The strongest model below 2B was Gemma 3 1B. The smallest model to reach 100% was Danube3 at 500M parameters (English only).

Table 1. Models achieving 100% in both Norwegian and English.

Model	Size	Family	NO	EN	Method
Gemma 3 1B	1B	Gemma	100%	100%	LoRA
Qwen 2.5 3B	3B	Qwen	100%	100%	LoRA
Neural Chat 7B	7B	Intel/Mistral	100%	100%	LoRA
Gemma 2 9B	9B	Gemma	100%	100%	LoRA
Llama 2 13B	13B	Meta	100%	100%	LoRA
Qwen 2.5 14B	14B	Qwen	100%	100%	LoRA
Gemma 2 27B	27B	Gemma	100%	100%	LoRA
Qwen 2.5 32B	32B	Qwen	100%	100%	LoRA
Llama 3.1 70B	70B	Meta	100%	100%	Zero-shot
Qwen 2.5 72B	72B	Qwen	100%	100%	Zero-shot

Mistral Nemo 12B (100%/80%) was the closest near-miss. Additional models achieved 100% in one language only.

Architecture-family consistency was notable: Gemma reached 100% in 7/7 variants tested, Mistral-based models went 6/6, and Qwen 2.5 achieved 100%/100% at 3B, 14B, and 32B (with 7B as a near-miss at 100%/80%).

4.2 Language acted as a major behavioral condition

Language was one of the strongest variables in the entire project. Among small models (<7B), 81% performed better in English than Norwegian. Several shifted from 0% to 100% at language switch (Gemma 2B, Phi-3 3.8B, OLMo 1B, Danube3

500M).

Among larger models (7B+), the pattern reversed: most performed better in Norwegian than English. This **language inversion** around approximately 7B parameters suggests that sufficient pretraining data for a language removes it as a limiting factor, while insufficient data makes it the dominant constraint.

4.3 Architecture mattered more than raw size

The evaluation did not support a simple size-based account of performance. Several counter-examples demonstrated this:

- Danube3 500M outperformed Danube2 1.8B (100% vs 60%)
- StableLM-2 1.6B outperformed StableLM 3B (100% vs 60%)
- Gemma 3 1B matched Qwen 2.5 3B at 100%/100%
- DeepSeek-R1 1.5B scored only 40%/40% despite being a reasoning-optimized model -- its chain-of-thought mode collided with the KMI response format

These results support a model of performance in which architecture family, generation characteristics, and runtime compatibility can matter as much as or more than parameter count.

4.4 Re-baselining corrected inflated earlier impressions

The prompt-template correction, which affected approximately 37 previously trained models over roughly six months of work, changed interpretation of several earlier results. Some previously strong-looking scores weakened when the evaluation formatting was corrected, indicating that earlier versions of the benchmark had indeed inflated performance for parts of the leaderboard.

Qwen 2.5 3B remained strong after correction, which strengthened its status as the most reliable all-round model. This re-baselining step functioned as a credibility check on the overall project.

4.5 Zero-shot emergence revealed a non-monotonic pre-emergence zone

The zero-shot emergence mapping produced one of the most striking findings of the project. Across 15 models spanning 3B to 72B, KMI behavior did not improve monotonically with scale.

Table 2. Zero-shot vs. LoRA performance. Scores shown as Norwegian / English.

Model	Params	Zero-shot	With LoRA	Tier
Neural Chat 7B	7B	0% / 0%	100% / 100%	VBSA essential
Qwen 2.5 3B	3B	20% / 0%	100% / 100%	VBSA essential
Qwen 2.5 7B	7B	20% / 0%	100% / 80%	VBSA critical
Llama 2 7B	7B	20% / 0%	100% / 80%	VBSA critical
Llama 3.1 8B	8B	40% / 40%	100% / 80%	VBSA critical
Mistral 7B	7B	60% / 0%	100% / 60%	VBSA critical
Hermes 3 8B	8B	60% / 0%	100% / 80%	VBSA critical
Gemma 2 9B	9B	60% / 20%	100% / 100%	VBSA critical
Llama 2 13B	13B	20% / 0%	100% / 100%	VBSA critical
Qwen 2.5 14B	14B	20% / 0%	100% / 100%	VBSA critical
Mistral Small 22B	22B	20% / 0%	--	VBSA critical
Gemma 2 27B	27B	60% / 60%	100% / 100%	VBSA beneficial
Qwen 2.5 32B	32B	20% / 0%	100% / 100%	VBSA critical
Llama 3.1 70B	70B	100% / 100%	--	Emergent
Qwen 2.5 72B	72B	100% / 100%	--	Emergent

Table 3. Three-zone emergence model for KMI behavior.

Zone	Size Range	Zero-shot KMI	VBSA Effect
Pre-emergence	3B-32B	0-60% (non-monotonic)	Critical (-> 100%)
Gemma-family outlier (subset)	9B-27B	20-60%	Strong (-> 100%)
Robust emergence	70B+	100%/100%	Unnecessary

The Gemma family was a consistent outlier: Gemma 2 27B scored 60%/60% zero-shot while Qwen 2.5 32B scored only 20%/0%. This indicates that architecture and pretraining composition matter more than parameter count even at the 27-32B scale. The non-monotonic pre-emergence zone between 3B and 32B indicates that KMI is not a smooth scaling

phenomenon but a latent capacity whose activation threshold is architecture- and pretraining-dependent. Parameter count alone did not explain performance across this 10× range; the results instead suggest a stronger role for architecture family, pretraining composition, and the extent to which the base model already supports a premise-challenging response tendency. VBSA appears to lower this threshold by roughly 10×, turning a capability that requires 70B+ into something accessible at 3B.

A qualitative observation: Models that failed KMI zero-shot did not produce random or incoherent output. Instead, they reliably fell back to a **default advisory mode** -- offering suggestions, giving opinions, or providing information rather than asking exploratory questions. This is not a failure of language capacity but a failure of response function. The models generally appeared to track the conversational context; they simply defaulted to the wrong response class.

The central practical implication: a 3B model with VBSA training (100%/100%) outperforms a 32B model without it (20%/0%). VBSA appears to compress access to the target behavior by approximately 10× in parameter budget using 60 training examples.

4.6 CPF ablation: compressed prompting preserved the effect best

The most compressed condition, Pilot F ("Utfordre premisser."), achieved the highest total score (4.144) and the lowest variance (0.032) on Qwen 2.5 3B. The full six-line prompt (Pilot B) scored lowest at 3.896. The ranking was consistent: F ≥ D > E >> B >> Base.

Table 4. CPF ablation results on Qwen 2.5 3B (2000 samples each).

Pilot	Prompt	Lang	Struct	Anti	Robust	Total	Variance	Train Loss
F	"Utfordre premisser."	4.80	3.84	3.88	5.00	4.144	0.032	1.135
D	Minimal (both)	4.84	3.76	3.96	5.00	4.140	--	--
E	"Still spørsmål."	4.60	3.72	3.80	5.00	4.068	--	--
B	Full 6-line prompt	4.96	3.08	4.04	5.00	3.896	--	0.670

The difference between F and D is very small (0.004), which means the strongest safe conclusion is that **minimal prompts outperform the full prompt**, while the claim that premise-challenging alone is strictly superior to the two-component prompt requires more statistical power.

The prompt "Utfordre premisser" appears to function as a **minimal discriminative stimulus**: a compressed behavioral cue that activates a latent response class rather than specifying output format. This is consistent with recent work on task vectors in language models (Todd et al., 2024), where compressed representations can activate specific behavioral directions without explicit instruction. Rather than telling the model what to do step by step, it provides a stance from which the desired behavior emerges naturally.

4.7 Lower training loss did not predict better downstream behavior

Pilot B had the lowest training loss (0.670) but weakest evaluation performance. Pilot F had the highest training loss (1.135) but the strongest evaluation outcome. This inverse pattern was replicated in the cross-base study across three additional architecture families, suggesting it is a general phenomenon rather than an artifact of the Qwen architecture.

The interpretation: detailed teacher prompts increase the student's ability to imitate teacher surface patterns during training, thereby lowering loss, while weakening independent generalization at evaluation time. The shorter prompt leaves more of the task "unresolved" during training and forces the student toward deeper functional compression.

4.8 CPF generalized across four architecture families

The cross-base study confirmed that the CPF effect is portable. All six student bases passed quality thresholds. Architecture family appeared to be a major driver of score differences, with intra-family finetune differences being negligible.

Table 5. CPF cross-base results. All use Pilot F prompt, 2000 samples, Claude Haiku 4.5 teacher.

Student	Family	Loss	Total
Hermes 3 8B	Llama 3.1/Nous	0.944	4.324
Llama 3.1 8B	Llama 3.1	0.948	4.310
Neural Chat 7B	Mistral/Intel	0.649	4.220
OpenChat 3.5 7B	Mistral	0.638	4.204
Qwen 2.5 3B	Qwen	~1.1	4.144
Gemma 3 4B	Gemma	0.938	3.806

Notable observations:

- **Hermes 3 8B** achieved the overall CPF record (4.324), surpassing the original Qwen 2.5 3B result
- **Intra-family differences were negligible**: Hermes 3 vs Llama 3.1 base: Δ0.014; Neural Chat vs OpenChat: Δ0.016
- **Gemma 3 4B scored lowest** (3.806) despite the Gemma family's strong VBSA performance (7/7 at 100%). This discrepancy may partially reflect a quantization mismatch (trained in 4-bit, evaluated in bf16), and suggests that VBSA learnability and CPF distillation quality measure different properties
- The inverse loss-performance relationship held across all families

4.9 Data scaling showed diminishing returns

Pilot C (5000 samples) scored 4.234 on Qwen 2.5 3B, compared to Pilot F's 4.144 with 2000 samples -- an improvement of +0.09 for 2.5× more data. The improvement was uniform across dimensions (Language +0.08, Structural +0.12, Anti-filler +0.12) but small in magnitude.

This suggests that the method's effectiveness derives primarily from prompt structure rather than data volume, and that 2000 samples captures most of the available signal. Consequently, combined with the model-size and prompt-length findings, these results support a broader **"structure > size"** interpretation across three dimensions: model parameters (3B matches much larger models with VBSA), prompt length (2 words outperform 6 lines), and training data volume (2000 samples capture most of the signal vs. 5000).

A secondary observation: Qwen 2.5 3B with 5000 samples (4.234) approached Hermes 3 8B with 2000 samples (4.324), indicating that data scaling and base model size appear to trade off to some extent within the studied range -- more data can partially compensate for fewer parameters.

4.10 The active ingredient was stance rather than question format

The CPF findings sharpen interpretation of what was actually being learned. A prompt that explicitly instructs the model to ask questions ("Still spørsmål") is not necessarily teaching the deepest part of the target behavior. The stronger minimal results suggest that the crucial signal is the **relational stance** captured by "Utfordre premisser."

Under this interpretation, questioning behavior is not the primary target but an emergent consequence of premise-challenging. The model does not merely learn to produce question marks; it learns a response tendency organized around gently destabilizing assumptions and opening up alternative interpretations. That possibility is consistent with the broader VBSA-Method goal of facilitating reflection rather than merely producing superficially reflective language.

5. Discussion

Three main conclusions emerge from the present experiments. First, VBSA-Method-like behavior is broadly learnable across small language models when evaluation is conducted under fair, model-compatible conditions. Second, the way this behavior is taught matters at least as much as whether it can be taught at all -- compressed prompting produces stronger behavioral generalization than longer teacher prompts. Third, the zero-shot emergence mapping reveals that KMI behavior is a latent capacity that exists in instruction-tuned models but requires either massive scale (70B+) or targeted shaping (VBSA) to reliably surface.

The zero-shot findings are the most theoretically significant. The absence of gradual improvement between 3B and 32B contradicts a naive scaling hypothesis. In the present zero-shot set, larger models in the 14B-32B range did not reliably outperform much smaller ones. Parameter count alone was therefore insufficient to explain performance across this range, although architecture-specific outliers (notably the Gemma family, reaching 60% at 27B) indicate that family and pretraining composition matter substantially. At 70B, the behavior appears robustly and fully formed. The overall pattern -- a non-monotonic pre-emergence zone followed by robust full emergence -- is broadly consistent with emergent capability as described by Wei et al. (2022), though the Gemma outliers suggest the threshold may be partly architecture-dependent.

The advisory fallback pattern reinforces the functional interpretation. Models that failed KMI zero-shot reliably fell back to a default advisory mode rather than producing random output. VBSA appears to redirect this default, not primarily by adding new task knowledge, but by shifting which response class is most reliably activated under conversational conditions.

VBSA's contribution is therefore precise: it makes KMI behavior reliably available where it does not naturally surface. For models below 70B -- which covers the entire practical range for edge deployment, mobile applications, and cost-sensitive production -- VBSA is typically the difference between 0-60% (and often 0-20%) and 100%. The value proposition is not brute-force capability scaling, but reliable activation of a latent response class at lower parameter cost.

The CPF ablation reinforces this through a different lens. Detailed teacher prompts produced lower training loss but weaker evaluation performance. This inverse pattern replicated across four architecture families in the cross-base study. The most defensible interpretation is that detailed prompts encourage **topographic imitation** -- the student copies the surface form of aligned behavior -- while minimal prompts force **functional internalization**. The student cannot simply

reproduce the teacher's rules because there are almost no explicit rules to reproduce. Instead, it must derive the behavioral stance from the training conversations themselves.

The distinction between topography and function is central to the entire project. A model may appear aligned because it reproduces the visible shape of an expected response: short answers, reflective tone, no lists, no explicit advice, particular opening phrases. Yet these surface markers do not guarantee that the model is enacting the intended conversational function. The CPF data suggest that what matters most is not whether the model learns to *look like* VBSA-Method, but whether it learns to *behave* in a way that preserves its exploratory stance under novel prompts and contexts.

The data scaling result (Pilot C) adds a third dimension to the "structure > size" principle. The finding that 2000 samples captures most of the CPF signal, with 5000 producing only marginal improvement (+0.09), mirrors the model-size finding (3B matches much larger models with VBSA) and the prompt-length finding (2 words outperform 6 lines). In all three cases, less is more when the structural signal is correct. This convergence across three independent domains -- parameters, prompt length, and data volume -- strengthens the claim that behavioral shaping is fundamentally about conditions, not quantity.

The CPF findings also raise a broader alignment hypothesis. It is plausible that many distillation pipelines over-specify teacher behavior and thereby reward stylistic copying more than functional transfer. If so, the apparent success of richly prompted teachers may conceal a failure mode: models that imitate the outward form of aligned behavior without acquiring its deeper contingency structure. The inverse loss-performance relationship provides a concrete diagnostic: when training loss drops but evaluation performance doesn't improve or worsens, the model is likely memorizing surface patterns rather than acquiring the response function.

A unified mechanistic view. The inverse loss-performance relationship, the language inversion around 7B, and the non-monotonic pre-emergence zone together support a coherent interpretation: VBSA does not install new behavior but activates and stabilizes a latent response class that requires either massive scale or targeted shaping to surface reliably. The practical implication is that the relevant question for deployment is not "can this behavior be trained" but "at what minimal parameter cost and prompt support can it be made to appear?" Zero-shot failures in the pre-emergence zone were consistently systematic policy failures -- defaulting to advisory mode -- rather than comprehension failures. The models generally appeared to track the conversational context; they simply lacked the conditions to activate the correct response class.

The Gemma discrepancy is worth noting. The Gemma family showed the strongest VBSA learnability (7/7 variants at 100%, Gemma 3 1B as strongest sub-2B) and the highest zero-shot outlier scores (60%/60% at 27B). Yet Gemma 3 4B scored lowest in the CPF cross-base study (3.806). This suggests that VBSA learnability (can the model learn the behavior from direct examples?) and CPF distillation quality (can the model acquire the behavior through teacher-generated data?) measure different properties. A model family may excel at one while underperforming at the other.

On the value of VBSA even at scale. While the zero-shot findings show that 70B+ models can perform KMI behavior without training, this does not mean VBSA is irrelevant at those scales. VBSA-trained models showed more consistent performance across diverse scenarios, while zero-shot performance at sub-70B scales was highly variable. Even where the behavior is latently available, targeted shaping may improve consistency and reduce the need for elaborate system prompts.

6. Limitations

Several limitations apply.

First, the fair-conditions evaluation includes per-model adjustments, making the setup less standardized than conventional benchmarks. The tradeoff is deliberate: ecological fairness over rigid symmetry. However, this makes direct cross-family comparison more difficult.

Second, the zero-shot emergence mapping covers only two models at the 70B+ threshold (Llama 3.1 70B and Qwen 2.5 72B). Additional architecture families at that scale would strengthen the emergence claim.

Third, the CPF ablation used 25 evaluation prompts and one judge stack (GPT-4.1-mini double-judge). The difference between the two strongest minimal conditions (F and D) is very small (0.004 on Qwen 3B), which means the strongest safe conclusion is that minimal prompts outperform the full prompt, while the claim that premise-challenging alone is strictly superior requires more statistical power. Additional judge configurations and larger evaluation sets would increase confidence.

Fourth, several models failed for technical reasons (runtime incompatibility, library issues, hardware limits) that do not constitute negative evidence about VBSA-Method learnability. See Appendix A for a systematic analysis of failure patterns.

Fifth, some 100%-scoring models were qualitatively weaker in output naturalness, showing repetitive patterns or formulaic openings. Binary success scores alone do not fully capture response quality.

Sixth, the Gemma 3 4B cross-base evaluation involved a quantization mismatch (trained in 4-bit, evaluated in bf16), which may partially account for its lower CPF score.

Seventh, the evaluation used 5 scenarios per language (10 total). While this was sufficient to identify broad patterns, it limits statistical power for detecting smaller effect sizes.

Eighth, all hardware was consumer-grade or standard cloud instances, which is part of the project's practical contribution but also introduces constraints that would not apply in a well-funded research lab.

7. Conclusion

This study suggests that VBSA-Method-like behavior can be shaped in small language models more broadly, at smaller scales, and with less data than a rigid benchmark view would predict. Across 80+ models and 15 zero-shot evaluations, exploratory reflective conversational behavior was shown to be learnable across multiple architecture families, remained detectable at 500M scale, and emerged spontaneously at 70B+ scale without any training.

The zero-shot emergence mapping produced the most striking result: KMI behavior does not improve monotonically with scale from 3B to 32B. Instead, it remains in a non-monotonic pre-emergence zone before appearing robustly at 70B+. A 3B model with VBSA (100%/100%) outperforms a 32B model without it (20%/0%) -- 10× fewer parameters, 5× better score. This quantifies VBSA's value precisely: it makes KMI available where massive scale is impractical.

The CPF ablation, replicated across four architecture families and six student bases, identifies a specific mechanism: compressed teacher prompts promote functional internalization while detailed prompts encourage topographic imitation. Data scaling from 2000 to 5000 samples yielded diminishing returns (+0.09), confirming that the method's effectiveness derives from structural precision rather than volume. Consequently, these results support the broader "structure > size" interpretation across three dimensions: model parameters, prompt length, and training data volume.

Taken together, the findings support the central claim of Verbal Behavior Shaping Alignment: in conversational domains, the more relevant alignment target is not rule compliance or framework recitation, but acquisition of a response function that survives across realistic conditions. The advisory fallback pattern sharpens the broader alignment implication of this work: in conversational systems, failure may often reflect the activation of the wrong response policy rather than a lack of knowledge. Models in the pre-emergence zone were often capable of understanding the situation, but not of reliably enacting the exploratory function the task required. VBSA-Method is therefore better understood not as a prompt recipe, but as a functional class of verbal behavior that can be shaped, faded, and tested for generalization. Total project cost remained under $50 for 100+ model trainings across consumer and cloud hardware.

8. Next Steps

1. Add additional 70B+ zero-shot models from other architecture families (Gemma, Mistral) to strengthen the emergence claim.
2. Replicate the CPF ablation with larger evaluation sets, additional judge configurations, and inter-rater reliability checks.
3. Test whether the compressed-prompt advantage generalizes to other behavioral targets beyond KMI.
4. Add richer quality metrics beyond binary success, including naturalness, repetitiveness, depth of reflection, and failure-mode annotation.
5. Develop a concrete external replication package with locked scripts, fixed evaluation prompts, scoring protocol, and at least one independent rerun by a separate operator.
6. Compare VBSA-Method-style shaping against alternative conversational objectives to test whether the observed effect is specific to reflective prompting or general to behavioral distillation.
7. Prepare a condensed research note for rapid sharing alongside a practitioner-facing summary for colleagues outside AI research.

References

Hayes, S. C., Brownstein, A. J., Zettle, R. D., Rosenfarb, I., & Korn, Z. (1986). Rule-governed behavior and sensitivity to changing consequences of responding. *Journal of the Experimental Analysis of Behavior*, 45(3), 237-256.

MacDuff, G. S., Krantz, P. J., & McClannahan, L. E. (2001). Prompts and prompt-fading strategies for people with autism. In C. Maurice, G. Green, & R. M. Foxx (Eds.), *Making a difference: Behavioral intervention for autism*. Pro-Ed.

Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts.

Skårholen, K. (2026). *AXxionBridge: Multi-agent AI orchestration platform* [Software]. Available from author upon request.

Skårholen, K. (2026). *MetaABC Eval Framework v0.2* [Software]. Available from author upon request.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2024). Function vectors in large language models. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Appendix A. Technical Failure Patterns and Operational Robustness

Throughout the fair-conditions evaluation, a substantial number of tested models were excluded from behavioral scoring due to technical failures rather than confirmed behavioral weakness. These failures were logged and analyzed separately as a signal about operational robustness and architecture-specific fragility. The patterns clustered into four recurring types.

Table A1. Technical failure clusters observed across the evaluation.

Cluster	Typical Issue	Examples	Interpretation
Runtime attention fragility	Model loads and generates tokens but produces degenerate or incoherent output under quantized inference	Phi-family	Some architectures tolerate quantization less gracefully, particularly those with non-standard attention mechanisms
Serialization and loading issues	Model fails during weight loading, checkpoint deserialization, or adapter merging	Older community fine-tunes	Reflects ecosystem maturity rather than architectural weakness
Memory ceiling	Model exceeds available VRAM during loading or inference, even under 4-bit quantization	Qwen 2.5 14B on RTX 3080, Gemma 3 4B on RTX 3050	A hardware constraint, not a behavioral one. These models may be fully learnable under adequate compute
Dependency burden	Model requires specific library versions, custom kernels, or non-standard dependencies	InternLM, Hymba, RWKV-7, MobiLlama	Indicates deployment friction rather than learnability failure

These clusters are not mutually exclusive; a single model may exhibit more than one failure type simultaneously. Importantly, none of these patterns constitute negative evidence about whether the affected models can learn VBSA-Method-like behavior. A model that fails to load under 4-bit quantization has not been shown to lack the target response class -- it has simply not been tested under conditions it can operate in.

For practitioners interested in deploying VBSA-Method-style fine-tuning, operational robustness may matter as much as behavioral ceiling. Future work could formalize this distinction by reporting both a behavioral learnability score and an operational robustness profile for each evaluated model.

Appendix B. Data and Implementation Availability

Training data, evaluation scenarios, LoRA configurations, CPF pipeline scripts, and distillation prompts are available from the author upon request. This includes:

- The 60-example VBSA-Method training dataset (multi-turn JSONL format)
- Per-model LoRA configurations and architecture-specific adjustments
- CPF synthetic data generation pipeline
- Evaluation scoring protocol and held-out scenarios
- Trained LoRA adapters for selected models

Contact: kenneth.skaarholen@gmail.com

Candidate Core Claims

1. VBSA-Method-like behavior is broadly learnable across small language-model families under model-compatible conditions.

2. Language functions as an active behavioral variable, not a cosmetic wrapper, in conversational alignment tasks, with an inversion point around 7B parameters.
3. Minimal teacher prompts can outperform detailed teacher prompts in behavioral distillation.
4. Lower training loss may reflect stronger stylistic imitation without better functional generalization.
5. Fair-conditions evaluation is necessary when the target phenomenon is a response function rather than a fixed output format.
6. KMI behavior emerges spontaneously at 70B+ scale but is not robustly emergent between 3B and 32B -- VBSA compresses this by ~10x.
7. The principle "structure > size" generalizes across model parameters, prompt length, and data volume.
8. CPF is portable across multiple architecture families: the effect replicates with six student bases across four families.
9. The advisory fallback pattern suggests that alignment failures in conversational models may often be failures of response function, not knowledge.