# Specialist-Only Model Composition via Relevance-Gated Cross-Attention

Rohith V. Putha
MemraneLabs.org
ro@membranelabs.org

**Abstract**

Representation-level model composition offers a way to reuse a specialized language model without updating a general-purpose base model. Existing composition methods, including CALM, train an interface using examples of the composed behavior, so the same data distribution must specify both how specialist information should be transferred and when that information should be used. This paper studies a different factorization for narrow specialists. A cross-attention bridge is trained only on the specialist task, while a separate prompt-level relevance gate decides whether the resulting specialist residual is applied to the base model. The bridge learns representation transfer; the gate learns activation. In a biomedical negation experiment using frozen Qwen2.5-0.5B base and specialist models, a specialist-only bridge reaches 0.985 accuracy and 0.941 negated-class F1. Adding relevance gating preserves nearly all in-domain performance, reaching 0.983 accuracy and 0.938 F1, while suppressing the specialist pathway on ordinary assistant prompts. A no-specialist attention ablation remains near linear-probe performance, indicating that the gain comes from specialist hidden states rather than from additional trainable attention alone. The results support relevance-gated interfaces as a data-factorized approach to attaching narrow specialists: specialist supervision trains the communication path, and relevance supervision controls its use.

## 1 Introduction

Modern language-model ecosystems increasingly contain narrow specialists: models adapted for clinical text, code, mathematical subroutines, low-resource languages, or other bounded capabilities. A useful composition mechanism should reuse such a specialist without overwriting the behavior of a general-purpose base model. The mechanism should also match the data that is usually available: for a narrow specialist, one often has task supervision for the specialist itself, but not a curated dataset describing the desired composed model across both specialist and general-purpose inputs.

Representation-level composition is an attractive abstraction for this setting. Instead of routing between final text outputs, the base and specialist are both run, and a learned interface transfers hidden-state information from one model to another. CALM demonstrates that frozen models can be composed by training cross-attention modules over their intermediate representations [1]. That formulation uses composition examples $\mathcal{D}_{\text{comp}}$ that represent the target combined behavior. When the target is genuinely a joint skill, such data is the right supervision. For narrow specialists, however, the central difficulty is different: the bridge can be trained from specialist data, but it should not become active on every prompt.

This paper introduces a relevance-gated compositional interface for this specialist-only regime. The design separates two roles that are coupled in a standard always-on bridge. First, a cross-model attention bridge learns how to transfer specialist representations into the base residual stream

using only specialist-task data $\mathcal{D}_{\text{spec}}$. Second, a relevance router learns whether the prompt invokes the specialist skill using prompt-level relevance data $\mathcal{D}_{\text{rel}}$. At inference time, the router scales the specialist residual:

$$\widetilde{H} = H^B + g(x)\, C_\phi(H^B, H^S), \qquad g(x) \in [0, 1]. \tag{1}$$

The base model, specialist model, and base language-model head remain frozen.

The key distinction is between *confidence* and *relevance*. A binary negation specialist can be confident in the label space {`YES`, `NO`} even when the user asks for a poem or an arithmetic answer. High specialist confidence therefore does not imply that the specialist computation belongs in the base model's residual stream. The gate estimates task invocation: whether the specialist pathway is licensed for the current prompt.

The empirical study uses biomedical negation detection as a controlled specialist task. Biomedical negation has a narrow output space, a clear binary objective, and an easily observable off-task mode-transfer behavior. The results show that specialist hidden states are responsible for the in-domain improvement, and that relevance gating preserves the base model's ordinary generation mode without requiring general assistant examples in the bridge-training set. The contributions are:

1. a data-factorized formulation of specialist-only model composition, separating representation transfer data $\mathcal{D}_{\text{spec}}$ from relevance data $\mathcal{D}_{\text{rel}}$;
2. a residual cross-attention interface controlled by a prompt-level relevance gate;
3. an empirical demonstration that the gated interface retains the specialist benefit on biomedical negation while preventing the specialist label space from dominating unrelated prompts; and
4. an ablation showing that trainable attention without specialist hidden states does not account for the observed gain.

## 2 Related Work

**Cross-model composition.** CALM composes frozen language models by learning projections and cross-attention modules between an anchor model and an augmenting model [1]. It provides the closest architectural precedent for this work: both approaches operate on intermediate representations rather than on final text outputs. The present work targets a different data regime. CALM trains the composition module with examples of the desired composed behavior $\mathcal{D}_{\text{comp}}$. Here, the bridge is trained only with specialist-task examples $\mathcal{D}_{\text{spec}}$, and a separate router supplies the activation signal. The methods are therefore complementary: CALM addresses how to learn a composed behavior from composition examples, while relevance-gated interfaces address how a narrow specialist can be attached when the bridge is trained from specialist data alone.

| Aspect | CALM-style composition | Relevance-gated specialist composition |
|---|---|---|
| Source models | Frozen anchor and augmenting models. | Frozen base and frozen specialist. |
| Communication path | Learned cross-attention over model representations. | Learned cross-attention from base states to specialist states. |
| Bridge supervision | Composed-behavior examples $\mathcal{D}_{\text{comp}}$. | Specialist-task examples $\mathcal{D}_{\text{spec}}$. |
| Activation policy | Induced by the composition training distribution. | Explicit prompt-level gate trained on $\mathcal{D}_{\text{rel}}$. |
| Primary data question | How to learn the target joint behavior. | How to attach a narrow specialist without bridge training on general prompts. |

Table 1: Positioning relative to CALM. The proposed interface keeps the representation-level communication path but separates bridge training from activation control.

**Attention and residual interfaces.** The bridge uses scaled dot-product attention [2] as a communication mechanism between two frozen models. Base hidden states form queries; specialist hidden states form keys and values. The bridge output is a residual update to the base hidden state, following the principle that learned modules can modify an existing computation through residual updates rather than replace it [3]. In this setting, the residual form is important: the specialist proposes an update, and the gate controls whether that update is expressed.

**Parameter-efficient adaptation and modular routing.** Adapters, prompt tuning, and LoRA freeze most pretrained parameters while learning task-specific modules [4, 5, 6]. AdapterFusion combines information from separately trained adapters [7], and sparse mixture-of-experts models route computation to subsets of parameters [8]. The present work is closer to modular routing than to single-model adaptation: the specialist is an independent frozen model, and the learned module controls access to its representations. The router is prompt-level in the present experiment, which is a natural choice for a binary task specialist; multi-skill settings can extend the same principle to expert selection or token-level activation.

**Biomedical negation.** The BioScope corpus annotates negation, uncertainty, and scope in biomedical text [9]. Biomedical negation is useful as a specialist task because it has a narrow label space and an unambiguous target behavior. The task provides a clean experimental substrate for studying composition because the specialist label space and off-task behavior are both directly observable.

# 3 Method

## 3.1 Problem setup

Let $x$ denote a chat-formatted prompt. A frozen base model $B_\theta$ produces hidden states $H^B \in \mathbb{R}^{T\times d}$, and a frozen specialist model $S_\psi$ produces hidden states $H^S \in \mathbb{R}^{T\times d}$. The base language-model head $W_{\text{lm}}$ is also frozen. The trainable components are a cross-model bridge $C_\phi$ and a relevance router $R_\eta$.

The composition problem is expressed through three datasets:

$$\mathcal{D}_{\text{spec}} = \{(x_i, y_i)\} : \text{specialist-task supervision}, \tag{2}$$
$$\mathcal{D}_{\text{rel}} = \{(x_i, r_i)\} : \text{prompt-level relevance supervision}, \tag{3}$$
$$\mathcal{D}_{\text{comp}} = \{(x_i, y_i)\} : \text{composed-behavior supervision}. \tag{4}$$

The specialist-only setting assumes access to $\mathcal{D}_{\text{spec}}$ and $\mathcal{D}_{\text{rel}}$ for a narrow skill, while avoiding bridge training on $\mathcal{D}_{\text{comp}}$. The bridge is optimized to transfer specialist information on in-scope examples. The router is optimized to decide whether the current prompt is in scope.

### 3.2 Cross-model residual bridge

For each attention head $i \in \{1, \ldots, h\}$, the bridge computes

$$Q_i = \text{LN}(H^B) W_i^Q, \tag{5}$$
$$K_i = H^S W_i^K, \tag{6}$$
$$V_i = H^S W_i^V, \tag{7}$$
$$A_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}} + M\right), \tag{8}$$
$$Z_i = A_i V_i, \tag{9}$$

where $M$ masks padding positions in the specialist sequence. The specialist proposal is

$$\Delta H = C_\phi(H^B, H^S) = [Z_1; \ldots; Z_h] W^O. \tag{10}$$

The router produces

$$g(x) = p_{R_\eta}(r = 1 \mid x), \tag{11}$$

and the composed representation is

$$\widetilde{H} = H^B + g(x) \Delta H. \tag{12}$$

The next-token distribution is computed by the frozen base head:

$$p(y_t \mid x) = \text{softmax}(W_{\text{lm}} \widetilde{H}_t). \tag{13}$$

During autoregressive decoding, the current prefix is passed through the base, specialist, and router; the gated residual is formed; and the next token is decoded from the base head. No source-model parameters are updated.

In the experiments, $d = 896$, $h = 8$, and $d_h = 112$, yielding approximately 3.21M trainable bridge parameters. The router is implemented as a learned classifier with the same model family as the source models, ensuring that routing capacity is not the limiting factor in the mechanism study.

### 3.3 Training decomposition

The router is trained with binary cross-entropy on relevance labels:

$$\min_\eta \; \mathbb{E}_{(x,r)\sim\mathcal{D}_{\text{rel}}} \left[ -r \log g(x) - (1 - r) \log(1 - g(x)) \right]. \tag{14}$$

Positive examples invoke the negation task. Negative examples include general assistant prompts and alternative task instructions, including yes/no formats, so that relevance cannot be reduced to the output tokens `YES` and `NO`.

After router training, $\eta$ is frozen. The bridge is then trained only on negation examples:

$$\min_{\phi} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{spec}}} \left[ -\log p_\phi(y \mid x) \right], \qquad y \in \{\texttt{YES}, \texttt{NO}\}. \tag{15}$$

Thus the bridge learns specialist transfer without observing ordinary assistant examples. Preservation of the base model's general behavior is handled by the activation variable $g(x)$ rather than by mixing general prompts into bridge training.

| Component | Supervision | Function |
|---|---|---|
| Specialist $S_\psi$ | Biomedical negation task. | Provides task-specialized hidden states. |
| Bridge $C_\phi$ | Negation-only examples $\mathcal{D}_{\text{spec}}$. | Transfers specialist information into the base residual stream. |
| Router $R_\eta$ | Relevance labels $\mathcal{D}_{\text{rel}}$. | Activates or suppresses the specialist residual. |
| Base $B_\theta$ and head $W_{\text{lm}}$ | Frozen. | Preserve the general language-model computation and decoding head. |

Table 2: Training decomposition. Specialist transfer and relevance estimation are trained from different supervision signals.

### 3.4 Relevance is not specialist confidence

The router estimates task invocation, not the specialist model's predictive confidence:

$$\text{conf}_{S_\psi}(x) = \max_y p_{S_\psi}(y \mid x), \tag{16}$$

$$\text{rel}(x) = p_{R_\eta}(r = 1 \mid x). \tag{17}$$

A narrow classifier must distribute probability over its own label space even on out-of-scope prompts. Relevance is therefore a property of the prompt and task, not a property of the specialist's preferred label.

## 4 Experimental Setup

### 4.1 Task and data

The target task is binary biomedical negation detection. Inputs are formatted with a system instruction asking whether a sentence states that something is not true, not present, or did not happen. The output space is exactly {`YES`, `NO`}. The held-out test set contains 1,188 examples, including 152 negated examples. Performance is reported as accuracy and F1 for the negated class.

Base preservation is examined with a greedy-generation probe under a generic helpful-assistant system prompt. The prompts are ordinary non-biomedical requests. This probe tests whether the specialist pathway changes the model's output mode from assistant generation to specialist classification.

### 4.2 Models and systems compared

The base model is Qwen2.5-0.5B-Instruct. The specialist is a Qwen2.5-0.5B model adapted for biomedical negation. Both models are frozen in all interface experiments.

The comparison includes four systems:

1. **Frozen-base linear probe.** A linear classifier is trained on frozen base hidden states. This measures the negation information available without specialist representations.
2. **Specialist-only always-on bridge.** The residual cross-attention bridge is trained on negation-only examples and applied with $g(x) = 1$ for every prompt.
3. **Specialist-only relevance-gated bridge.** The same bridge is scaled by the learned prompt-level relevance score.
4. **Gated no-specialist attention ablation.** The specialist states are replaced by base states, keeping the residual attention form while removing specialist information.

## 5 Results

### 5.1 Specialist-task performance

| System | Specialist states | Gate | Accuracy | Negated F1 |
|---|---|---|---|---|
| Frozen-base linear probe | No | No | 0.9015 | 0.4706 |
| Gated no-specialist attention | No | Yes | 0.9024 | 0.4775 |
| Specialist-only always-on bridge | Yes | No | 0.9848 | 0.9408 |
| Specialist-only relevance-gated bridge | Yes | Yes | 0.9832 | 0.9383 |

Table 3: Held-out biomedical negation results. Specialist hidden states account for the improvement: the no-specialist attention ablation remains near the linear-probe baseline, while both specialist bridges reach high negated-class F1.

Table 3 isolates the role of specialist representations. Adding trainable attention without specialist states produces little change over the linear probe. Allowing the bridge to attend to specialist states raises the negated-class F1 from below 0.48 to approximately 0.94. The relevance-gated bridge preserves almost all in-domain performance of the always-on bridge while using an explicit activation variable to control off-task behavior.

### 5.2 Off-task preservation

| Prompt | Always-on bridge | Relevance-gated bridge |
|---|---|---|
| What is the capital of France? | `NO` | `The capital of France is Paris.` |
| Write a haiku about the ocean. | `NO` | `Whispers of the sea,\nMists rise in` |
| $2 + 2 =$ | `NO` | `2 + 2 = 4` |
| Tell me a fun fact about cats. | `NO` | `One fun fact about cats is that they are the` |

Table 4: Greedy generations under a generic assistant prompt. The always-on specialist bridge maps unrelated prompts into the specialist label space; the relevance-gated bridge suppresses that residual and recovers ordinary assistant completions.

The preservation probe shows why activation control is part of the composition problem rather than a post-processing detail. The always-on bridge is effective on the specialist task, but the same unconditional residual can make the specialist label space dominate unrelated prompts. The gated bridge keeps the specialist communication path available on in-scope prompts and inactive on the generic assistant prompts in Table 4.

### 5.3 Relevance behavior

The trained router separates prompts that invoke the negation task from generic assistant prompts and alternative task formats. This behavior depends on the instructional context rather than only on the sentence content or the target label tokens. The same biomedical sentence can be a negation-classification request under the task instruction and ordinary content under a helpful-assistant instruction. The router therefore supplies the activation variable in specialist composition: whether the specialist computation is appropriate for the prompt, independently of the specialist's preferred answer.

## 6 Analysis

**Data-factorized composition.** The central design choice is the factorization of bridge training and activation control. An always-on interface trained on specialist examples learns a useful specialist residual, but it has no structural variable representing whether that residual belongs in the current computation. Relevance gating introduces that variable explicitly. The resulting supervision is separated by role:

$$\text{transfer: } \mathcal{D}_{\text{spec}} \to C_\phi, \qquad \text{activation: } \mathcal{D}_{\text{rel}} \to g(x). \tag{18}$$

This factorization is useful precisely because $\mathcal{D}_{\text{spec}}$ and $\mathcal{D}_{\text{rel}}$ are different kinds of data. Specialist-task examples require task labels. Relevance examples only require knowing whether a prompt invokes the specialist skill. The method therefore avoids using general assistant prompts as bridge-training examples while still preserving the base model's general output mode.

**Transfer and routing address different problems.** The no-specialist attention ablation shows that the in-domain improvement is explained by access to the specialist's hidden states rather than by additional trainable parameters near the frozen head. The preservation probe identifies a separate property of the interface: a specialist residual should be expressed only when the prompt

invokes the specialist skill. Cross-attention defines the communication channel; relevance gating defines its activation policy.

**Relation to CALM.** CALM and the present interface share the core idea that frozen models can be composed through learned representation-level communication. They differ in the supervision used to train the interface. CALM learns the composition module from examples of the composed behavior. Relevance-gated specialist composition trains the bridge from specialist-task supervision and represents activation with a separate prompt-level variable. This gives a different operating point: when composed-behavior examples are available, they can directly supervise the target joint capability; when only a specialist dataset is available, relevance gating provides an explicit mechanism for attaching the specialist without training the bridge on general-purpose behavior.

**Design implications.** The experiment isolates the mechanism in a setting with a clear specialist label space and directly observable off-task behavior. The same decomposition extends to richer modular systems: cross-model interfaces define how specialist representations enter the base computation, while relevance variables define which specialist pathways are active for a prompt. In multi-specialist settings, this formulation turns composition into a structured activation problem rather than an unconditional merging of all available specialist computations.

# 7 Conclusion

This paper presents a relevance-gated interface for specialist-only model composition. A residual cross-attention bridge transfers hidden-state information from a frozen specialist into a frozen base model, while a prompt-level router controls whether that transfer is active. The bridge is trained on specialist examples alone; it does not require general assistant examples to learn preservation behavior. In a biomedical negation experiment, specialist hidden states raise negated-class F1 from below 0.48 to approximately 0.94, and relevance gating preserves nearly all of this gain while suppressing the specialist residual on unrelated prompts. The broader lesson is that model composition should separate two decisions: how models communicate, and when that communication is licensed by the task.

# References


[1] Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. LLM augmented LLMs: Expanding capabilities through composition. arXiv:2401.02412, 2024.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

[5] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, 2021.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[7] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rueckle, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of EACL*, 2021.

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[9] Gyorgy Szarvas, Veronika Vincze, Richard Farkas, Gyorgy Mora, and Janos Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2008.

[10] Qwen Team. Qwen2.5 technical report. arXiv:2412.15115, 2024.