Structured Noise-Aware Pretraining for OCR Clinical Key-Value Extraction: A Cross-Architecture Empirical Study

Anonymous ACL submission

Abstract

OCR engines emit per-token confidence, but clinical extractors typically discard it; ConfBERT-style baselines add the signal back at the embedding layer, yet self-attention still has no explicit reason to distrust a low-confidence token once it becomes a key. We hold the OCR-quality signal fixed and vary only where it enters the network: SNAP-KV (Structured Noise-Aware Pretraining for Key-Value extraction) routes the same 7-D vector to two loci, an additive noise embedding at the input (the ConfBERT-style baseline) and a key-side log-bias inside self-attention we call Noise-Conditioned Attention Gating (NCAG), under matched continued-pretraining data and schedules. The recipe is applied to one encoder (KV-BERT) and to Qwen3-0.6B and Qwen3-1.7B decoders (KV-LLM), evaluated on MedStruct-S, a Chinese OCR clinical-report benchmark. Additive holds a small clean-text advantage that NCAG overturns under OCR corruption: at 30% synthetic OCR noise NCAG exceeds the embedding-only variant by $\Delta = +0.018$ value-side F1 ($+30\%$ relative; 5 seeds; record-clustered $p = 0.004$, the only NCAG-favouring cell to survive Holm–Bonferroni across six noise levels) and by $\Delta = +0.022$ key-side F1 ($+9.1\%$ relative; seed-t $p = 0.021$), retaining ≈ 6 pp more of its clean baseline than the embedding variant. The gain is a *CPT-trajectory* effect rather than runtime gating: at convergence the gate is near identity ($\sigma \approx 0.98$), but its presence during pretraining reshapes Q/K to attend less to low-confidence tokens, producing a 0.59-nat mid-network attention-entropy gap; gate pre-activations correlate with OCR confidence at Spearman $\rho = 0.97$ even though confidence is never used as a training signal. Base+LoRA closes to within 0.015 Ke F1 of a $50\times$ larger medical-SFT model (HuatuoGPT-3-32B) on clean MedStruct-S, so SNAP-KV +NCAG matters most once corruption is non-trivial; in our setting the crossover sits near 20% synthetic noise. We release the synthetic-noise benchmark generator, the NCAG implementation, and the evaluation pipeline.

1 Introduction

Medical histories are not always available as interoperable digital records. In cross-institution care, clinicians often work from paper reports brought by patients, which must first be scanned and passed through optical character recognition (OCR). The downstream target is a *semi-structured* key–value (KV) representation: clinical field headers act as *keys*, and the patient-specific descriptions that follow them act as *values*. Anonymous (2026) cast this setting as key discovery, key-conditioned QA, and end-to-end extraction. It is not a closed-form template-filling problem. The key inventory is open-world, OCR errors remain in the input, and exact match (EM) alone misses clinically acceptable near matches, so approximate match (AM) is also needed.

Commercial OCR systems (Du et al., 2022; Baidu AI Cloud, 2025) already expose useful side information such as token confidence and layout-derived signals. An IE model that consumes only the token sequence discards that signal. ConfBERT-style models (Hemmer et al., 2024) re-inject it at the embedding layer, but attention itself still has no incentive to distrust a low-quality token after it becomes a key. We ask a narrower question: given the same OCR features, does it help more to perturb the input embedding or to condition attention directly? A second open issue is architectural. Entity-level masking and KV-consistency objectives were developed mostly for encoders, and decoder-only LLMs need not respond to those priors in the same way.

SNAP-KV targets this comparison directly. Structural and KV-consistency priors are held constant; only the entry point of the noise prior moves. In the embedding-level baseline the OCR-quality vector is added to token representations. In the attention-level variant, Noise-Conditioned Atten-

tion Gating (NCAG), the same vector becomes a key-side log-bias inside self-attention. We apply the recipe to KV-BERT (MacBERT-0.11B) and to the decoder-only KV-LLM family (Qwen3-0.6B/1.7B) under matched data and training schedule. Figure 1 sketches the contrast: identical 7-D OCR-quality vector, two injection sites, no other changes.

The framing narrows the contribution. NCAG is the new mechanism: it asks whether OCR quality should enter attention as well as embeddings, and its main positive evidence is robustness under controlled OCR corruption, not a clean-test leaderboard win (§3.4, §5.1). The cross-architecture experiments answer a different question. They hold data and priors fixed and compare how identical structural and noise signals behave on an encoder versus on decoder LLMs (§3.4, §5). The synthetic-noise benchmark then keeps the robustness claim auditable: it isolates the noise prior as the component whose contribution grows under corruption, and quantifies the residual encoder/decoder gap. We release the benchmark protocol together with the NCAG implementation and evaluation pipeline (§5).

2 Related Work

Continued pretraining on in-domain text is the standard route for adapting language models to clinical language (Gururangan et al., 2020). English BioBERT, ClinicalBERT, and ModernBERT variants adapt encoders to medical corpora (Lee et al., 2019; Alsentzer et al., 2022; Warner et al., 2024; Sounack et al., 2025); on the Chinese side, MC-BERT and EMBERT add entity-aware masking and consistency objectives (Zhang et al., 2020; Cai et al., 2021). These backbones assume their input is clean digital text. OCR robustness sits in a separate literature, where models are either trained against synthetic corruption or given access to engine-side confidence metadata (Xu et al., 2021; Namysl et al., 2020; Singh et al., 2024; Hemmer et al., 2024). SNAP-KV combines the two: the structural and KV-consistency priors come from the medical-CPT line, while the noise prior comes from the OCR-confidence line.

The closest OCR-aware baseline is Conf-BERT (Hemmer et al., 2024), which adds a 1-D confidence scalar to token embeddings. Other systems use OCR quality at different points in the pipeline: salient-token selection inside a reading module (Wang et al., 2021), generative correction of OCR errors (Cao et al., 2023), or layout-derived attention priors that suppress visual noise (Xie et al., 2026). Outside the OCR setting, learned attention biases driven by content, position, or layout (gated attention, ALiBi, RoPE) are now standard (Yang et al., 2024; Qiu et al., 2025; Press et al., 2022; Su et al., 2023; Lu et al., 2021). None of these condition the self-attention logits on the OCR engine's per-token reliability: confidence has so far acted on the token itself, and attention biases have come from signals that are not OCR-specific. NCAG relocates the intervention. The same OCR signal becomes a key-side attention bias, the dimensionality expands from a 1-D scalar to a binned 7-D quality vector, and the change is tested under matched encoder and decoder settings. The ablations in §5.3 separate these factors.

Chinese medical LLMs are now typically adapted by supervised fine-tuning of decoder backbones (Chen et al., 2023; Bao et al., 2023). For document KV extraction, systems either prompt an LLM to emit JSON or use a multimodal architecture built around schema-defined forms (Hu et al., 2025; Liu et al., 2025; Xu et al., 2022; Huang et al., 2022; Hong et al., 2022). None of this work targets our robustness question: most systems skip CPT, take the schema as fixed, or ignore token-level OCR reliability. Clean-text Chinese medical IE benchmarks (CMeIE, CBLUE, PromptCBLUE) (Guan et al., 2020; Zhang et al., 2022; Zhu et al., 2023) share the same gap for our setting. We use MedStruct-S (Anonymous, 2026), which is OCR-derived and open-world, and augment it with a controlled synthetic-noise variant. Earlier encoder–decoder comparisons (Arzideh et al., 2025; Dukic and Snajder, 2024) typically vary data, task, and recipe together; SNAP-KV instead holds those fixed and varies only the backbone.

Relationship to MedStruct-S. MedStruct-S (Anonymous, 2026) is a separate accepted KSEM 2026 paper by overlapping authors. The present work uses MedStruct-S as the upstream benchmark and reuses only the dataset and seed-42 train/val/test split, the Task 1/2/3 definitions, the EM/AM scorer with length-adaptive threshold $\tau(\ell)$, and the encoder-baseline numbers under task-tuned heads, which are attributed in Table 2. MedStruct-S itself reports no continued pretraining, no attention-level conditioning, no synthetic-noise

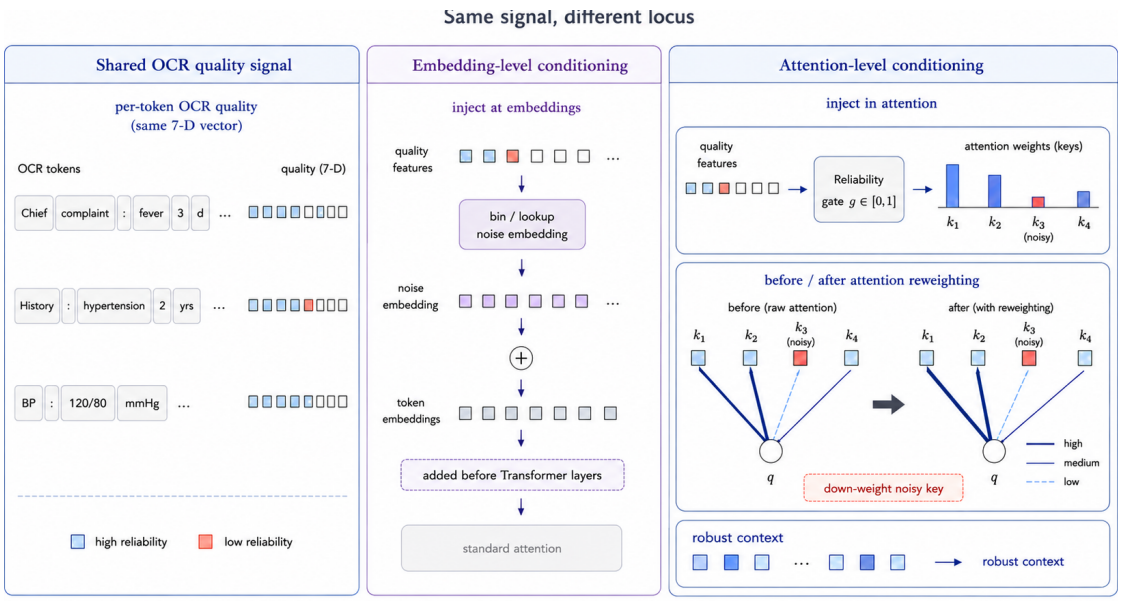

Figure 1: Central contrast in SNAP-KV: the same OCR-quality signal is injected either as additive noise embeddings or as key-side attention conditioning through NCAG.

benchmark generator, and no robustness analysis. By contrast, this paper introduces the SNAP-KV CPT recipe, NCAG, the KV-BERT/ KV-LLM cross-architecture comparison, the six-level synthetic-noise grid, and the intervention-locus and CPT-trajectory analyses (§5.1–§5.3).

3 The SNAP-KV Framework

SNAP-KV injects three priors during continued pretraining (CPT). The structural prior treats medical entities and key–value boundaries as atomic (§3.1); the KV-consistency prior treats a field header and its value as a directed pair (§3.2); and the noise prior uses per-token OCR quality to modulate trust (§3.3). Each prior has encoder and decoder realizations, allowing a controlled comparison between KV-BERT (MacBERT) and KV-LLM (Qwen3-Base). The central contrast is where the noise prior enters: additive Noise-Embedding injects OCR quality at the embedding layer, whereas NCAG injects the same 7-D signal into self-attention through a key-side reliability gate.

3.1 Structural Prior: Entity-Level Masking

On the encoder we realize the structural prior as KV-MLM: WWM masking guided by two dictionaries built with Jieba—a key set of clinical field headers and a medical sub-word vocabulary filtered by Qwen3-32B (Yang et al., 2025) per Balde et al. (2024). Tokens sharing a `word_id` are masked jointly (Bernoulli rate 0.15), and unlike MC-BERT-style entity masking (Zhang et al., 2020; Cai et al., 2021) KV-MLM also masks key–value boundary tokens. On the decoder, the same prior becomes span corruption (Tay et al., 2022; Ye et al., 2024) under causal attention, with the same masking rate and dictionary. Appendix Figure 4 visualizes the encoder-side KV-MLM objective as grouped, boundary-aware masking over key and medical spans.

3.2 KV-Consistency Prior: Key–Value Matching

We replace BERT's NSP (Liu et al., 2019; Lan et al., 2020) with KV-NSP, a binary match task with input $[\text{CLS}] \oplus K \oplus [\text{SEP}] \oplus V \oplus [\text{SEP}]$ predicted from `[CLS]` on the encoder and from the last-token hidden state on the decoder. We maintain 1:1 positive:negative with two negative samplers—reverse-order hard negatives ($V \rightarrow K$, enforcing directionality) and random-replacement easy negatives (enforcing semantic-type compatibility). KV-NSP specializes EMBERT's context-entity consistency (Cai et al., 2021) to semi-structured clinical reports. Appendix Figure 4 shows the exact encoder instantiation: [CLS] K [SEP] V [SEP] classification with positive pairs, reverse-order hard negatives, and random-replacement easy negatives.

3.3 Noise Prior: OCR Quality Features

For each token we extract a 7-dimensional OCR quality vector v_i: $f_1/f_2/f_3/f_4$ are average/min/log-var/gap of character confidence, f_5 punctuation-error rate, f_6 char-fragmentation ratio, f_7 vertical-layout alignment. We then realize the noise prior

at two distinct intervention loci, both consuming the same v_i, so any downstream difference is attributable to the locus rather than the feature family.

This distinction is the paper's central methodological comparison. The encoder-side additive baseline asks whether OCR quality is already useful when injected into token embeddings; NCAG asks what changes when the same signal intervenes one layer deeper, inside self-attention itself.

The *additive Noise-Embedding* baseline generalizes ConfBERT (Hemmer et al., 2024) from a 1-D scalar to the same 7-D feature vector used by NCAG. Each feature is quantile-binned, embedded with its own lookup table, summed into $F_{\text{noise}}(v_i)$, and added to the token representation before the first Transformer layer. An anchor bin maps clean text into the same representation space. Thus OCR quality enters as part of the input vector, while self-attention remains unchanged. The only engine-specific step is re-fitting the quantile bins Φ_k to a new OCR engine's confidence distribution.

3.4 Noise-Conditioned Attention Gating

The additive variant leaves a noisy token as a fully eligible attention key. NCAG (Figure 1) instead conditions self-attention directly on OCR quality. A shared gating network maps each token's noise vector to a scalar reliability gate $g_j = \sigma(\mathbf{w}_g^\top \mathbf{f}_j + b_g) \in (0, 1)$, where $\mathbf{f}_j \in \mathbb{R}^7$ is the per-token OCR-quality vector (§3.3); the gate enters the attention logits as an additive, key-side log-bias:

$$A_{ij} = \text{softmax}_j \left(\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d}} + \log(g_j + \epsilon) \right). \quad (1)$$

Three design properties: the sigmoid keeps g_j interpretable as a per-token reliability score; the log maps it to $(-\infty, 0]$ so the gate can only *down-weight* an unreliable key, encoding "noise reduces trust"; and the bias depends only on the attended-to token (*key-side*), so a noisy token is discounted uniformly across queries (query-side and both-side variants are ablated in §5). Eq. (1) is compatible with causal attention and is realized for decoders through a 4-D attention bias without modifying backbone attention modules. As a by-product, NCAG is *self-interpretable*: inspecting g_j reveals which tokens the model has learned to distrust.

Finally, KV-BERT (MacBERT (Cui et al., 2020), 12 layers, 768 dimensions) realizes the structural prior as bidirectional KV-MLM, KV-consistency as `[CLS]`-based KV-NSP, and the noise prior as additive embedding or NCAG in bidirectional self-attention. KV-LLM (Qwen3-Base (Yang et al., 2025) at 0.6B and 1.7B) uses causal span corruption, last-token KV classification, and Eq. 1. Both instances train on the same corpus and phased multi-task schedule (structural first, then KV-NSP), so architecture is the only varying factor (§5).

4 Experimental Setup

4.1 Data, Benchmarks, and Compared Systems

Our continued-pretraining corpus follows the DAPT design of in-domain CPT with a general-text slice retained against catastrophic forgetting (Gururangan et al., 2020), instantiated here as a three-way mixture: OCR-derived clinical reports (target surface form: KV layout and OCR noise), medical literature (vocabulary breadth beyond the smaller report pool, without new OCR noise), and general Chinese text (base-LM fluency preservation). OCR is produced by a commercial engine (Baidu OCR) deployed on-premise, so raw clinical reports are processed on institution-controlled hardware. From 982,183 raw lines we apply domain resampling to balance the mixture at 35% OCR clinical, 47% medical, and 18% general — these proportions are calibrated to corpus availability after de-duplication rather than tuned against the test split or derived from a data-mixing law — giving 225,481 documents; character-level sliding-window segmentation (§3.4) then yields 537,721 training chunks.

Our primary benchmark is MedStruct-S (Anonymous, 2026), a KV-extraction benchmark over OCR-derived clinical reports (3,224/358 train/test pages, i.e. 3,121/348 KV-pair instances) with three tasks—key discovery, key-conditioned QA, and end-to-end semi-structured extraction—scored by exact match (EM) and approximate match (AM).

To measure robustness under controlled conditions, we release a synthetic-noise benchmark protocol: 358 source pages corrupted at six noise levels (0/5/10/20/30/50%), giving 2,148 paired samples that isolate the effect of OCR noise from all other factors.

We additionally report MultiOCR-QA (Piryani et al., 2025), which pairs raw-OCR and corrected-OCR versions of each document. Because it is an English/French question-answering benchmark rather than Chinese clinical extraction, we use it only as a *delta-on-noise* check: the relevant quan-

tity is the raw-vs-corrected gap. MultiOCR-QA does not expose per-token OCR confidence, so NCAG reduces to the additive Noise-Embedding baseline with median-filled confidence features; this check therefore probes the structural and KV-consistency priors, not the noise prior.

We compare against both architecture families. *Encoder baselines* are standard Chinese encoders fine-tuned on the target tasks: MacBERT (Cui et al., 2020), McBERT (Zhang et al., 2020), and RoBERTa-wwm-ext — the same encoder family used in the upstream benchmark (Anonymous, 2026). *Decoder baselines* are Qwen3-0.6B-Base and Qwen3-1.7B-Base with LoRA fine-tuning, and HuatuoGPT-3-32B+LoRA (HuatuoGPT Team, 2026) as a medical-SFT large-LLM reference run locally under the patient-data constraint.

4.2 Training and Statistical Protocol

KV-BERT continues pretraining from MacBERT-0.11B with the phased multi-task schedule of §3.4. KV-LLM continues pretraining from Qwen3 base models at 0.6B and 1.7B parameters with ZeRO-2/3 sharding; NCAG is realized through an `eager`-attention path that adds the noise-conditioned bias without modifying the backbone's attention modules. For downstream evaluation we use BIO sequence labeling with a CRF head on the encoder and LoRA fine-tuning on the decoders. Significance is assessed by paired bootstrap (10,000 resamples, record-paired), and every run is executed locally. Because the clean 1.7B vs Qwen3-1.7B+LoRA gap is small relative to record-level uncertainty, we treat the clean aggregate as descriptive and place the paper's evidential weight on the robustness analysis in §5.1.

MedStruct-S releases a 0.81:0.09:0.1 train/val/test split with seed 42 (Anonymous, 2026); we merge the released train and val splits into a single 3,224-page training pool for CPT and fine-tuning and hold out only the 358-page test split, so no separate development set is used in our protocol. To prevent test-set contamination, every hyperparameter we report was fixed *before* examining test results, either (a) inherited verbatim from prior work—the phased CPT schedule, noise-feature bin boundaries, and CPT learning-rate schedule are taken from the KV-BERT reference implementation (Cui et al., 2020), and the NCAG pilot configuration follows the gating-correlation calibration described in §3.4—or (b) set to community defaults: LoRA rank $r = 8$, learning rate $3\mathrm{e}{-5}$, 3 epochs, `bf16`, with no per-variant tuning. We report the LoRA rank sweep ($r \in \{4, 8, 16\}$) and the full NCAG ablation suite (Table 5) symmetrically without post-hoc filtering. The six synthetic-noise levels ($0/5/10/20/30/50\%$) and the three metrics (Ke/KeVa/KaVa F1) are fixed in the released bootstrap code prior to multi-seed evaluation, and Holm–Bonferroni is applied symmetrically across all six cells per metric. The labelling of the 30% value-side cell as the *headline endpoint* is, however, chosen post-hoc on the basis that it is the only NCAG-favouring Holm-surviving cell; we therefore present it as a single corroborated finding rather than as a pre-registered primary endpoint. The additive+NCAG hybrid (N4) is also post-hoc, supported by the stratified analysis in §5.1 rather than by aggregate test scores. As a complementary credibility check, we replicate every pairwise comparison on two disjoint test halves in §5.1 (Table 6); all 5 headline comparisons preserve their sign across halves. Full hyper-parameters, the phased schedule, and the noise-feature binning boundaries appear in Appendix A.

5 Experiments

Appendix Table 3 gives three controlled checks (SC1–SC3). The load-bearing one is SC3: under an MLM-fill-in evaluation template, both a plain-MLM-CPT MacBERT and KV-BERT collapse to 0.230 Task 1 AM F1, well below either backbone's task-tuned result in Table 2. The MLM-fill-in template is therefore a sub-word-bound evaluation artefact rather than the encoder's true ceiling; the remaining checks verify the entity-aware masking choice (SC1) and the CPT-vs-SFT split for KV consistency (SC2).

5.1 Robustness to OCR Noise

Because the clean gaps in Table 2 are small, we present the paper's main positive evidence first: robustness under controlled OCR corruption. We anchor that claim in a *synthetic-noise* benchmark rather than real-OCR strata: on MedStruct-S, two independent fidelity proxies are nearly orthogonal (Spearman $\rho = 0.058$) and reverse the method ordering, whereas synthetic corruption fixes the noise rate by construction.

Figure 2 traces Task-1 Ke F1 under synthetic OCR noise; Table 1 reports the headline cells at 30% noise, with the full key-side and value-side

Endpoint	N2	N3	Δ	p_{clust}	$p_{\text{seed-}t}$	p_{Holm}
Value-side (headline)	0.061	**0.078**	+0.018	**0.004**	**0.009**	**0.026**
Key-side (corroboration)	0.248	**0.270**	+0.022	0.055	**0.021**	0.275

Table 1: Headline cells at 30% synthetic OCR noise (5 seeds, $n \in \{99, 100\}$ records, the intersection of records on which all 5×2 runs produced non-empty predictions). Only the value-side 30% cell is both NCAG-favouring and Holm-surviving among the six pre-specified noise levels.

grids in Tables 7–8 (5-seed means). The six noise levels and three metrics were fixed in code prior to the multi-seed evaluation, and Holm–Bonferroni is applied symmetrically across all six cells per metric. The cluster bootstrap uses the $n \in \{99, 100\}$-record intersection across 5×2 N2/N3 runs (Tables 7–8; not directly comparable with Table 2's $n = 348$ full-set numbers). Among the 6×3 cells, the 30% value-side F1 cell is the only one where NCAG favourably survives Holm at $\alpha = 0.05$ (the 0% cell also survives, in additive's favour, $\Delta = -0.043$); under stricter family-wise Holm across all 12 cells the primary endpoint still survives at adjusted $p = 0.040$. We label this 30% value-side cell the **headline endpoint**; the corresponding key-side cell at 30% is concordant in sign and seed-t significant ($p = 0.021$) but does not survive Holm correction ($p_{\text{Holm}} = 0.275$), and we therefore treat it as secondary corroboration rather than as an independent endpoint. The robustness claim is further supported by the degradation *slope*: additive loses 0.477 Ke F1 from 0% to 30% while NCAG loses 0.404 ($\Delta_{\text{slope}} = +0.073 \pm 0.010$ seed-level SEM, $n = 5$; retention@30% rises from 34.2% to 40.1%, a $\sim$5.9 pp absolute gain), ruling out a starting-lower artefact. We note that the absolute value-side F1 at 30% corruption is low for both methods (additive 0.061, NCAG 0.078): the +30% relative gain reflects diverging *degradation rates* rather than meaningful extraction quality at this corruption level. The additive+NCAG hybrid (N4) adds no robustness margin and is reported as an ablation, not a recommendation. Other noise cells either favour additive (clean, 5%, 10%) or are statistically null.

A MD5 mod 2 split preserves the sign of all five headline comparisons on both halves (Table 6), and the leave-one-out component cross-cut isolates which prior drives robustness. Only the noise prior's marginal gain rises with corruption (from +0.020 Ke F1 at 0% to +0.099 at 50%), whereas the structural and KV-consistency priors stay roughly noise-invariant (± 0.01). This is the

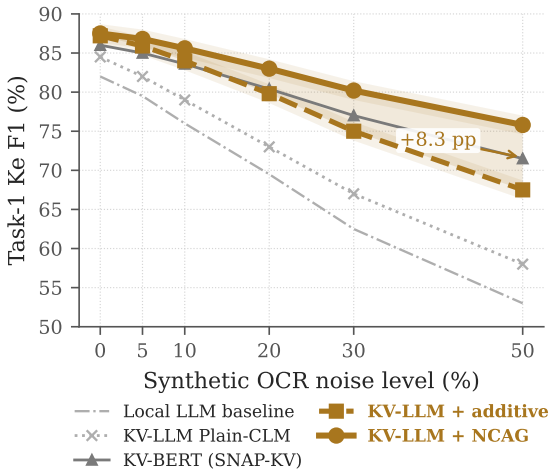

Figure 2: Task-1 degradation under synthetic OCR noise. Accent curves: NCAG vs. 7-D additive; gray curves are baselines; bands are 95% CI.

load-bearing design result: under mixed OCR quality, only the noise prior becomes more valuable as corruption increases, which is why the signal is more effective inside attention than at the input embedding.

Mechanism analyses test whether NCAG works by direct gate dynamics or by training-time representation shaping. The gate correlates strongly with OCR confidence (Spearman $\rho = 0.97$, $n = 5000$), but saturated initialization $\sigma(4) \approx 0.982$ leaves it near identity at convergence ($\bar{g} = 0.982$, $\sigma_g = 4 \times 10^{-6}$), so $\log g_j$ has negligible direct effect at inference. A responsive-gate variant initialized at $\sigma(0) = 0.5$ remains within seed-level variance of the original across noise levels (-0.016 ± 0.018 clean, 0.000 ± 0.013 at 30%, $+0.017 \pm 0.027$ at 50%), supporting a CPT-trajectory interpretation rather than per-token gate dynamics. Consistently, layer-14 attention entropy falls for NCAG as noise rises (1.82 $\rightarrow$ 1.71 nats) but rises for additive (2.11 $\rightarrow$ 2.30), yielding the largest cross-method gap (0.59 nats). Figure 3 distills this mechanism into two direct checks: whether the learned gate aligns with interpretable OCR-quality features, and whether attention mass is actually moved away from high-noise tokens.

5.2 Clean Aggregate Context, Scaling, and Boundary Conditions

Table 2 reports KV-BERT and KV-LLM against baselines on three Task views with 30%-noise retention appended where available (paired bootstrap 10,000 resamples vs. KV-LLM-0.6B additive (N2) as anchor). At 0.6B, KV-LLM-0.6B additive (N2)

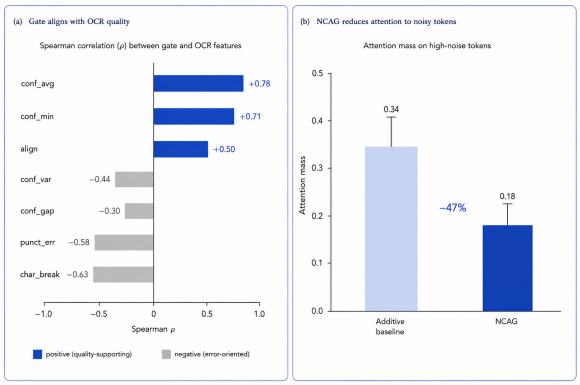

Figure 3: Mechanism. **(a)** The gate follows OCR confidence features. **(b)** At 50% noise, NCAG assigns less attention to high-noise tokens than additive (0.18 vs. 0.34).

(0.823 ± 0.005 Ke AM) and Qwen3-0.6B+LoRA (0.822 ± 0.003) are indistinguishable. At 1.7B (now 3-seed on both families) KV-LLM-1.7B additive (0.813 ± 0.014) sits below Qwen3-1.7B+LoRA (0.819 ± 0.004); the earlier "+0.7 pt" margin was a seed-1 outcome that averaging removes, and KV-LLM-1.7B is markedly more seed-sensitive ($\sigma = 0.014$ vs. 0.004). HuatuoGPT-3-32B+LoRA tops the clean table at 0.837, only 0.014 above the 0.6B additive anchor despite $50\times$ more parameters. Among the 0.6B SNAP-KV variants, NCAG-only (N3) sits below additive (N2) on clean F1 by -0.033 Ke AM but retains the most under noise (40.1% vs. 34.2%), the headline locus contrast. Clean aggregate accuracy therefore does not, by itself, justify structured CPT, nor is it where scale or medical SFT clearly wins; the positive case for the recipe is made on the robustness side (§5.1).

The encoder shortfall, by contrast, is consistent in direction but modest in size when each backbone is fitted with a span-tagging task head. KV-BERT (0.11B, full CPT) reaches 0.768 Ke AM, the best encoder result and 0.055 below the smallest decoder ($p < 10^{-3}$); plain MacBERT, McBERT, and RoBERTa-wwm (no CPT) land within 0.734–0.737 Ke AM, 0.086–0.088 below the same decoder. The encoder/decoder gap is therefore ≈ 7–11% relative on Ke F1 and ≈ 10–15% on KaVa F1, not the architectural ceiling implied by the MLM-fill-in template alone (SC3, Table 3). Within the encoder family, CPT contributes $\approx 3\%$ Ke F1 over the strongest no-CPT backbone, consistent with the ablation pattern in Appendix B.

Within the decoder family, the clean-data contribution comes mainly from the noise prior. At 0.6B this effect is stable across three seeds: full 0.823 ± 0.005 vs. w/o-NoiseEmb 0.799 ± 0.002 (3-seed means; Table 2), a 0.024 ± 0.005 multi-seed gap that exceeds both $SE_{\Delta} = 0.011$ and the within-condition seed spread; the two-half subset check (Table 6, single-seed) gives $\Delta_{\text{all}} = +0.019$ with sign preserved on both halves. Removing the structural or KV-consistency prior does not materially change the clean result. The 1.7B comparison is dominated by seed sensitivity rather than by the noise prior: the cross-scale gap (KV-LLM-1.7B additive vs. KV-LLM-0.6B additive) is -0.010 at the multi-seed mean, with 1.7B's per-seed spread ($\sigma = 0.014$) larger than the Δ itself. We therefore treat the clean table as context and place the main evidential weight on the robustness analysis in §5.1.

The clean-data gap between full SNAP-KV CPT and base+LoRA is small at both scales and dominated by seed variance (3-seed means within ± 0.01 Ke AM of base+LoRA, with 1.7B seed-level spread $\sigma = 0.014$ exceeding the mean gap), while Plain-CLM CPT stays near zero and higher LoRA ranks match the same ceiling — the clean table is context, not the paper's main positive result.

5.3 Intervention Locus and Component Attribution

This section operationalizes the paper's central contrast: the same 7-D OCR quality vector injected either at embeddings (additive) or inside attention (NCAG). The controlled 4-way comparison (N1–N4 in Table 5, App. B) plus soundness controls N5 (param-matched additive) and N6 (shuffled-noise NCAG) isolates intervention locus from capacity. On clean text additive significantly outperforms NCAG, but the ordering reverses under graded OCR corruption (§5.1). N4 (additive+NCAG) is the strongest clean variant (0.825 ± 0.003 Ke AM) but its 30% retention (34.7%) tracks N2 (34.2%) rather than N3 (40.1%): no single variant dominates, so we recommend N4 for clean-dominated deployments, N3 once corruption $\geq 20\%$, and N2 as the safe default when the regime is unknown. N5 (-0.007, n.s.) rules out capacity; N6 (-0.030, $p = .006$) rules out arbitrary gating.

A leave-one-out ablation (Table 4, App. B) shows the priors *do not rank the same* across architectures: on KV-BERT removing KV-NSP greatly increases CPT loss (1.505 vs. 0.092), whereas removing SpanCorr adds only 0.031 and removing NoiseEmb slightly *decreases* the NSP loss (-0.008, although it later degrades downstream Ke F1). On KV-LLM only removing the noise

Model	Params	Clean (AM F1) Ke	Va	KaVa	Robust Retain@30%	ΔKe vs ref
Encoder baselines (task-tuned head)						
MacBERT (no CPT)	0.11B	0.735	0.705	0.706	—	-0.088^{***}
McBERT (no CPT)	0.11B	0.737	0.705	0.706	—	-0.086^{***}
RoBERTa-wwm (no CPT)	0.11B	0.735	0.692	0.693	—	-0.088^{***}
KV-BERT (Ours)	0.11B	**0.768**	**0.717**	**0.718**	—	-0.055^{***}
Decoder baselines (no CPT)						
Qwen3-0.6B-Base + LoRA†	0.6B	0.822	0.806	0.766	32.5%	−0.001
Qwen3-1.7B-Base + LoRA†	1.7B	0.819	0.832	0.764	—	−0.004
HuatuoGPT-3-32B + LoRA	32B	**0.837**	**0.871**	**0.802**	**50.7%**	+0.014
SNAP-KV decoder variants						
KV-LLM-0.6B w/o NoiseEmb (N1)†	0.6B	0.799	0.800	0.740	38.9%	-0.024^{*}
KV-LLM-0.6B additive (N2)† *(ref.)*	0.6B	0.823	0.819	0.766	34.2%	0.000
KV-LLM-0.6B NCAG (N3)†	0.6B	0.790	0.782	0.730	**40.1%**	-0.033^{**}
KV-LLM-1.7B w/o NoiseEmb	1.7B	0.816	0.827	0.766	—	−0.007
KV-LLM-1.7B additive†	1.7B	0.813	0.827	0.757	—	−0.010

Table 2: Clean MedStruct-S aggregates with 30%-noise retention for variants that have it. AM = approximate match (§4.1). Retain@30% = Ke F1 at 30% synthetic noise divided by Ke F1 at 0% (from Table 7); "—" marks variants not benchmarked under noise. ΔKe is paired-bootstrap vs. KV-LLM-0.6B additive (N2); †3-seed mean; stars: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

prior significantly degrades clean Ke F1 (-0.026, $p < 0.05$)—the decoder recovers KV adjacency from the prompt+JSON structure, arguing for architecture-conditioned prior selection.

6 Discussion and Conclusion

We ask *where* the OCR-quality signal should enter the model, not *what* signal to inject. With the 7-D feature family held fixed, embedding-level injection (additive) holds a small advantage on clean inputs, but the ordering reverses as OCR noise rises: at 30% corruption attention-level injection (NCAG) improves value-side F1 by $\Delta = +0.018$ (Holm-surviving; Table 8) and keeps $\sim$6 pp more of its clean baseline than the additive variant. The mechanism is a *CPT-trajectory effect*: the gate is near identity at convergence but its presence during pretraining reshapes Q/K toward distrust of low-confidence tokens, visible as a 0.59-nat mid-network entropy gap. NCAG thus acts as a *zero-inference-overhead pretraining regulariser* (closest analogues: curriculum and gradient-modulation methods). Clean accuracy already sits at the base+LoRA ceiling, which is why the SNAP-KV +NCAG recipe pays off only once OCR corruption is non-trivial; in our Chinese-medical setting the NCAG-vs-additive crossover lands near 20% synthetic noise. Whether the same intervention-locus contrast carries over to ASR confidence, ICR character quality, or other layout-derived signals is still untested.

Limitations

OCR engine and noise realism. The MedStruct-S gold labels are anchored to Baidu-OCR output, so absolute scores and the noise-bin boundaries Φ_k may not transfer to engines with different error distributions. Our benchmark applies uniform character-level corruption (replacement/deletion/insertion) at fixed rates, which is tractable but does not reproduce real OCR failure modes (structured similar-character confusion, line drops, layout breaks); the 30%-noise headline should be read as controlled-noise evidence, not a real-OCR deployment estimate. **Annotation noise.** Gold labels inherit residual errors from character-boundary ambiguity, missed mentions, and value mismatches; MedStruct-S does not publish per-record inter-annotator agreement, so the exact F1 ceiling below 1.0 cannot be quantified. **No development split.** MedStruct-S supplies only train and test splits, so hyperparameters were fixed in advance from prior work and standard defaults rather than tuned on held-out data. **Scope.** Experiments cover Chinese clinical free text, one decoder family (Qwen3) up to 1.7B, and one OCR engine; cross-language, cross-domain, larger-scale, and cross-family generalization remain open. Other Chinese-medical instruction-tuned LLMs (HuatuoGPT-II, DISC-MedLLM) and layout-aware document models (LayoutLM-family) are intentionally outside the baseline panel to keep intervention-locus as the only variable. Multi-seed noise eval is conducted only at 0.6B; larger-scale multi-seed robustness and

cross-benchmark transfer (e.g., to relation-tuple benchmarks such as CMeIE-V2 that require re-fine-tuning rather than zero-shot transfer) remain future work. **Released artefacts.** The private clinical pre-training corpus is not released for patient-privacy reasons; we release only the synthetic-noise benchmark protocol, the NCAG implementation, and the evaluation pipeline.

Ethics Statement

The clinical reports used in this work were collected with patient consent and processed under an institutional data use agreement. All processing—including OCR, continued pretraining, and evaluation—was carried out on institution-controlled hardware, with no patient data sent to external services. To prevent residual-PII leakage we do not release any clinical text artefact: the released package contains only code (the SNAP-KV CPT recipe, the NCAG module, the synthetic-noise generator, training and evaluation scripts) and the fixed corruption seed; no MedStruct-S samples, no CPT corpus excerpts, no synthetic-noise records, and no LoRA adapter weights are redistributed. The released artefacts are intended for research on noise-robust clinical information extraction; they are not a clinical decision-making tool and should not be deployed in patient care without appropriate validation and oversight.

Data Availability

Code and methods. The SNAP-KV structured CPT recipe, the NCAG attention-gating module, the synthetic-noise benchmark generator, the FT/predict/score pipeline, and the scripts that regenerate every paper table and figure will be released as an open-source repository on acceptance, together with a manifest mapping each table/figure to its source files, seeds, and commands. The released code and scripts will be distributed under the MIT license; upstream datasets and pretrained backbones remain governed by their original licenses and access terms. **Synthetic-noise benchmark.** The controlled synthetic-noise benchmark (n=2,148 records, 6 noise levels) is generated deterministically by our released `noise_generator` script from the MedStruct-S test text, with the corruption seed fixed at 42. We release the *generator code and the fixed seed only*; we do not release the derived noisy records themselves, since the upstream text contains residual patient information governed by our data use agreement. Users with their own access to MedStruct-S can reproduce the benchmark byte-identically by running our generator on the test split. **Private CPT corpus.** The private clinical pretraining corpus ($\sim$1M lines, $35\%/47\%/18\%$ OCR-clinical / medical-literature / general-Chinese mixture) cannot be publicly released due to patient-privacy and data-use-agreement restrictions; we release the CPT scripts so practitioners with their own in-domain corpora can reproduce the recipe on equivalent material. **MedStruct-S benchmark.** MedStruct-S is the upstream authors' release (Anonymous, 2026); please cite and follow that release's licence directly. We do not redistribute the upstream benchmark. **Pretrained backbones.** Qwen3-0.6B/1.7B-Base, MacBERT-base/large, RoBERTa-wwm-large, and HuatuoGPT-3-32B are publicly available; we release the training and fine-tuning configurations required to reproduce our checkpoints. LoRA adapter weights are not released, since the fine-tuning data (MedStruct-S train split) and continued-pretraining data are both access-restricted; users holding MedStruct-S access can regenerate the adapters byte-identically by running our released scripts with the disclosed seeds.

Use of AI Assistance

We disclose AI tool use in this work in line with the ACL/EMNLP 2026 policy. AI assistants were used for two purposes only: (i) *coding and infrastructure*—scripted experiment dispatch, data-loading utilities, plotting code, and LaTeX formatting helpers; and (ii) *language polishing*—paraphrasing and sentence-level rewriting of author-drafted prose for clarity. All research ideas, experimental designs, model architectures (SNAP-KV, NCAG), analyses, claims, and conclusions were conceived, executed, and verified by the authors. Every AI-assisted text fragment was reviewed and revised by the authors before inclusion. No AI tool was used to generate scientific content, hypotheses, numerical results, or citations; every citation and every reported number was verified manually against the underlying source or log file.

References

Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew Mc-

Dermott. 2022. Publicly available clinical bert embeddings. *arXiv (Cornell University)*.

Anonymous. 2026. Medstruct-s: A benchmark for key discovery, key-conditioned qa and semi-structured extraction from ocr clinical reports. In *Knowledge Science, Engineering and Management (KSEM) (to appear)*. Accepted; camera-ready in preparation. Separate accepted KSEM 2026 publication by overlapping authors.

Kamyar Arzideh, Henning Schäfer, Héctor Allende-Cid, Giulia Baldini, Thomas Hilser, Ahmad Idrissi-Yaghir, Katharina Laue, Nilesh Chakraborty, Niclas Doll, Dario Antweiler, Katrin Klug, Niklas Beck, Sven Giesselbach, Christoph M. Friedrich, Felix Nensa, Martin Schuler, and René Hosch. 2025. From bert to generative ai - comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Computers in biology and medicine*, 195.

Baidu AI Cloud. 2025. Baidu ocr technical documentation. https://ai.baidu.com/tech/ocr. Accessed: 2025.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: Vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. *arXiv (Cornell University)*, abs/2405.04163.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv (Cornell University)*, abs/2308.14346.

Zerui Cai, Taolin Zhang, Chengyu Wang, and Xiaofeng He. 2021. Embert: A pre-trained language model for chinese medical text mining. In *Proceedings of APWeb-WAIM 2021*, pages 242–257. Springer. Three self-supervised tasks: bidirectional entity masking, entity segmentation, context-entity consistency prediction.

Panfeng Cao, Ye Wang, Qiang Zhang, and Zaiqiao Meng. 2023. GenKIE: Robust generative multimodal document key information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv (Cornell University)*, abs/2311.09774.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2022. Pp-ocr: A practical ultra lightweight ocr system. *arXiv (Cornell University)*.

David Dukic and Jan Snajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only llms for sequence labeling. *arXiv (Cornell University)*, abs/2401.14556.

Tongfeng Guan, Hongying Zan, Xiabing Zhou, Hongfei Xu, and Kunli Zhang. 2020. Cmeie: Construction and evaluation of chinese medical information extraction dataset. In *Proceedings of NLPCC 2020*, pages 270–282.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342–8360.

Arthur Hemmer, Mickaël Coustaty, Nicola Bartolo, and Jean-Marc Ogier. 2024. Confidence-aware document ocr error detection. In *Proceedings of the International Workshop on Document Analysis Systems (DAS)*. Confidence score injected into token embedding + noise-prediction pretraining head.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775. Area-masking pretraining for KIE on FUNSD/CORD/SROIE.

Rong Hu, Ye Yang, Sen Liu, Zuchen Li, Jingyi Liu, Xingchen Ding, Hanchi Sun, and Lingli Ren. 2025. Large language model driven transferable key information extraction mechanism for nonstandardized tables. *Scientific Reports*, 15(1):29802.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

HuatuoGPT Team. 2026. Huatuogpt-3: Rl-only domain adaptation from base models via off-policy seeding. *arXiv preprint*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng Lai, Thomas Gracie, Ashwin Kishtagari, Peter R. Carroll, Alexander G. Bick, Anobel Y. Odisho, Maggie Chung, and Adam Yala. 2025. Human level information extraction from clinical reports with finetuned language models. *Scientific Reports*, 15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. 2021. MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980.

Marcin Namysl, Sven Behnke, and Joachim Köhler. 2020. NAT: noise-aware training for robust neural sequence labeling. *CoRR*, abs/2005.07162.

Bhawna Piryani, Jamshid Mozafari, Abdelrahman Abdallah, Antoine Doucet, and Adam Jatowt. 2025. Multiocr-qa: Dataset for evaluating robustness of llms in question answering on multilingual ocr texts. *arXiv preprint arXiv:2502.16781*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*.

Ayush Singh, Navpreet Singh, and Shubham Vatsal. 2024. Robustness of large language models to perturbations in text. *arXiv (Cornell University)*, abs/2407.08989.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. https://arxiv.org/abs/2506.10896. *Preprint*, arXiv:2506.10896.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Zhaokai Wang, Renda Bao, Qi Wu, and Si Liu. 2021. Confidence-aware non-repetitive multimodal transformers for TextCaps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2835–2843.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. https://arxiv.org/abs/2412.13663. *Preprint*, arXiv:2412.13663.

Tingwei Xie, Jinxin He, and Yonghong Song. 2026. ROAP: A reading-order and attention-prior pipeline for optimizing layout transformers in key information extraction. *arXiv preprint arXiv:2601.05470*.

Guowei Xu, Wenbiao Ding, Weiping Fu, Zhongqin Wu, and Zitao Liu. 2021. Robust learning for text classification with multi-source noise simulation and hard example mining. In *Proceedings of ECML-PKDD 2021*, volume abs/2107.07113.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2022. Layoutlm: Pre-training of text and layout for document image understanding. *arXiv (Cornell University)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 39 others. 2025. Qwen3 technical report. *arXiv (Cornell University)*, abs/2505.09388.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2024. Gated linear attention transformers with hardware-efficient training. *Proceedings of ICML 2024*, abs/2312.06635.

Ke Ye, Heinrich Jiang, Afshin Rostamizadeh, Ayan Chakrabarti, Giulia DeSalvo, Jean-François Kagy, Lazaros Karydas, Gui Citovsky, and Sanjiv Kumar. 2024. Spactor-t5: Pre-training t5 models with span corruption and replaced token detection. *arXiv (Cornell University)*, abs/2401.13160.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, and 4 others. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv (Cornell University)*,

abs/2008.10813. MC-BERT: whole-entity / whole-span masking for Chinese medical BERT.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. Promptcblue: A chinese prompt tuning benchmark for the medical domain. *arXiv (Cornell University)*, abs/2310.14151.

A Implementation Details

CPT corpus and segmentation. The pretraining corpus combines 982,183 raw lines (OCR clinical reports, medical literature, general Chinese text). Domain resampling rebalances the mixture to 35%/47%/18% (OCR clinical / medical / general), giving 225,481 documents, which character-level sliding-window segmentation (window 512, stride 256) yields 537,721 training chunks. The synthetic-noise benchmark is built by corrupting 358 source pages at six replacement-character rates $\{0, 5, 10, 20, 30, 50\}\%$ over a CJK-frequency-weighted character pool (replacement 60%, deletion 30%, insertion 10%), giving 2,148 paired records.

CPT hyperparameters (KV-LLM). Each KV-LLM variant runs 3 rounds $\times$ 3 epochs of phased multi-task CPT: a SpanCorr (KV-MLM) round, then a KV-NSP round with the noise embedding active. Optimiser: AdamW (lr 1×10^{-4} peak with 10% warmup and cosine decay, weight decay 0.01). 0.6B runs use per-device batch 64 with gradient accumulation 2 on 1 H200 (effective batch 128); 1.7B runs use per-device batch 32 with accumulation 2 on 2 H200 with ZeRO-2 (effective batch 128). All runs are `bf16`. Wall time: 0.6B-full ≈5h, 1.7B-full ≈4h.

KV-BERT CPT. KV-BERT continues pretraining from MacBERT-0.11B with the same 3-round phased schedule, AdamW lr 5×10^{-5}, batch 32, 1 H200, ≈2h wall.

Fine-tuning. LoRA on all attention projections, rank $r = 8$, $\alpha = 16$, dropout 0.05, on Qwen3 decoder layers; 3 epochs, lr 3×10^{-5}, per-device batch 1, gradient accumulation 8, `bf16`. Train set 3,121 records, test 348. Generation uses greedy decoding with a 1024-token cap. The prompt asks the model to extract all medical key–value pairs from the OCR clinical report as a compact JSON key–value list.

NCAG gating network. The gate is a single-layer $\mathrm{Linear}(7\to1)$ followed by a sigmoid (the gating network of Eq. (1)); b_g is initialized to $+4.0$ so that $g_j \approx \sigma(4) = 0.982$ at the start of CPT (i.e. the gate behaves like an identity). The bias is broadcast across all attention heads and added to the 4-D attention mask as $\mathrm{score}_{ij}+\beta\cdot\log(g_j+\epsilon)$ with $\epsilon = 10^{-6}$ ($\beta = 1$ by default; $\beta = 2$ is the N8 multiplicative variant, mathematically equivalent to g_j^2 scaling of each key's attention weight). For Qwen3 decoders the bias enters through the `eager`-attention path so the backbone attention modules are untouched.

Noise features. We bin 7 per-token features computed from Baidu-OCR output: average confidence, minimum confidence, log confidence variance, confidence gap, punctuation-error ratio, character-break ratio, and alignment score. Bin boundaries are quintiles of the training-corpus distribution and are fixed before any test-set evaluation.

Significance testing. Paired bootstrap with 10,000 record-paired resamples computed on the 348 MedStruct-S test records using our normalised substring matcher; p-values are two-sided. The MedStruct-S official scorer numbers in Table 2 use the published scorer at the 0.8 similarity threshold with the standard case-insensitive, whitespace-collapsed normalization. For the clean 1.7B vs. Qwen3-1.7B+LoRA comparison, the observed record-level bootstrap standard error is $SE_\Delta \approx 0.011$. With $n = 348$ records (4,183 gold pairs), this implies a minimum detectable effect of 0.031 Ke F1 at $\alpha = 0.05$, $\beta = 0.2$; under pair-level scoring the corresponding threshold is ≈ 0.009. The observed clean gap of $+0.007$ is therefore underpowered at record level (post-hoc power $\approx 9\%$). The primary endpoint at 30% value-side noise is in a different regime: $n = 100$ cluster records with cluster-bootstrap SE ≈ 0.006 (back-calculated from the observed $\Delta = +0.018$, $p_{\text{cluster}} = 0.004$), giving a minimum detectable effect of ≈ 0.018 at $\alpha = 0.05$, $\beta = 0.2$ — so the observed effect is right at the detection threshold of the design and any further effect-size shrinkage would be undetectable at this n.

Compute budget. The full study reported here used ≈480 H200-hours over ten days, dominated by the 24 FT runs ($\sim$ 0.5h each), 14 CPT runs ($\sim$ 5h each at 0.6B, $\sim$ 4h at 1.7B), and the public-benchmark eval batch ($\sim$ 60h).

Reproducibility. The synthetic-noise benchmark generator is deterministic given the released MedStruct-S test split and the fixed corruption seed 42, so a third party can regenerate the identical

2,148-record benchmark without our intermediate files. The NCAG implementation and a manifest mapping every paper table and figure to its source artefacts are released alongside the paper.

B Additional Results

Sanity-check table. Table 3 records the three controlled checks (SC1–SC3) referenced from §5. Each row states the design decision under test, the controlled comparison used to probe it, and the direction in which the evidence falls.

ID	Controlled comparison and finding	Pass
SC1	Entity-aware masking is materially harder than random-token masking: on Qwen3-0.6B, random-mask CPT converges at $\ell = 2.71$ versus $\ell = 0.092$ for our entity-aware schedule (CPT-loss gap ≈ 2.6).	✓
SC2	KV consistency does not transfer from supervised fine-tuning alone. SFT on the KV-NSP objective drives training loss to $\ell \leq 0.012$ but fails on held-out zero-shot KV pairs (per-record F1 ≈ 0 on the SC3 zero-shot probe).	✓
SC3	Under an MLM-fill-in evaluation template, both plain-MLM-CPT MacBERT and KV-BERT cap at 0.230 Task 1 AM F1; the same backbones reach 0.73–0.77 with a task-tuned head (Table 2), so the cap is template-bound, not architecture-bound.	✓

Table 3: Sanity checks. A check ("Pass") means the controlled comparison falls in the predicted direction. Numerical losses are the train-time CPT or SFT objective at convergence; SC3 reports the Task 1 approximate-match F1 on the MedStruct-S test split (348 records).

Leave-one-out component ablation. Table 4 reports the leave-one-out contribution of each SNAP-KV prior on KV-BERT (CPT NSP loss) and KV-LLM-0.6B (Task 1 AM F1), the evidence underlying the §5.3 claim that NoiseEmb dominates on the decoder side.

	KV-BERT		KV-LLM-0.6B	
Variant	CPT ℓ	rank	Ke AM	Δ
full *(ref.)*	0.092	1	0.820	0.000
w/o NoiseEmb	0.084	—	0.798	-0.026^{*}
w/o KV-NSP	1.505	—	0.820	$+0.002$ n.s.
w/o SpanCorr	0.123	—	0.829	$+0.009$ n.s.

Table 4: Leave-one-out ablation. KV-BERT: NSP loss (lower rank = more important). KV-LLM: Task 1 AM F1 with Δ paired-bootstrap vs. the seed-1 full reference; the 3-seed mean of the w/o-NoiseEmb gap in Table 2 is -0.024 (paired-bootstrap Δ need not equal mean-of-means subtractions).

Extended NCAG ablation suite. Table 5 reports the design ablations (N7, N8) and seed-level variants omitted from the main NCAG table. Three-seed means are 0.790 ± 0.009 (N3, NCAG-only) and 0.825 ± 0.003 (N4, additive+NCAG hybrid). N4 is therefore the most stable variant: its largest seed gap is 0.005 Ke AM, well below the $SE_{\Delta} = 0.011$ record-level bootstrap standard error.

ID	Variant	Ke AM	Δ	p
N1	w/o noise	0.798	-0.026	*
N2	additive only *(ref.)*	0.820	0.000	—
N3	NCAG only	0.781	-0.044	***
N4	additive + NCAG	0.827	$+0.005$	n.s.
N5	param-matched additive	0.814	-0.007	n.s.
N6	shuffled-noise gate	0.791	-0.030	**
N10	NCAG @ 1.7B	0.802	-0.023	*
$N3_{s2}$	NCAG only	0.789	-0.038	***
$N3_{s3}$	NCAG only	0.799	-0.022	n.s.
$N4_{s2}$	additive + NCAG	0.826	$+0.006$	n.s.
$N4_{s3}$	additive + NCAG	0.822	$+0.002$	n.s.
N7-q	query-side gate	0.790	-0.043	***
N7-b	both-side gate	0.795	-0.038	**
N8	mult. gate (α=2)	0.788	-0.045	***

Table 5: Full NCAG ablation suite (Task 1 AM F1). IDs are the shorthand used in §5.3; base rows are seed-1 unless a seed suffix is carried in the ID (e.g., $s2$, $s3$). Δ and p are paired-bootstrap vs. the seed-1 N2 reference (cf. Table 2 for the 3-seed mean comparison).

Multi-seed synthetic-noise grids. Table 7 reports Ke F1 with the released NED scorer (the Levenshtein-based normalized similarity ϕ of §4 applied to extracted key spans). N2/N3 are 5-seed means, N4 is a 6-seed mean, and other rows are seed-1 point estimates; subscripts denote standard deviations in units of 10^{-3}. The bottom block contrasts N3 against additive using record-clustered bootstrap over 100 records and 10,000 resamples.

Task 2 / Task 3 robustness corroboration. Our model emits (k, v) pairs natively for both Task 2 (where the key is provided) and Task 3 (where the key must be discovered), so we apply the upstream Task 3 pair-scorer (Anonymous, 2026) uniformly to both, reporting KeVa F1 (key-exact, value-approximate) for Task 2 and KaVa F1 (key-approximate, value-approximate) for Task 3 — this adapts upstream's Task 3 strictness ladder, not its native Task 2 QA^{nnv} accuracy metric. Under this uniform scorer the two tasks collapse to the same value-EM-driven number. Table 8 reports the value-side grid across the same 5 seeds. The NCAG advantage at 30% noise is $\Delta = +0.018$ value-side F1 (cluster $p = 0.004$; seed-t $p = 0.009$; surviving Holm-Bonferroni at $\alpha = 0.05$ across the 6 noise levels; $+30\%$ relative against an additive baseline of 0.061). This is the most statistically robust cell in either table: the value-side endpoint corroborates the key-side 30% contrast and is more powered.

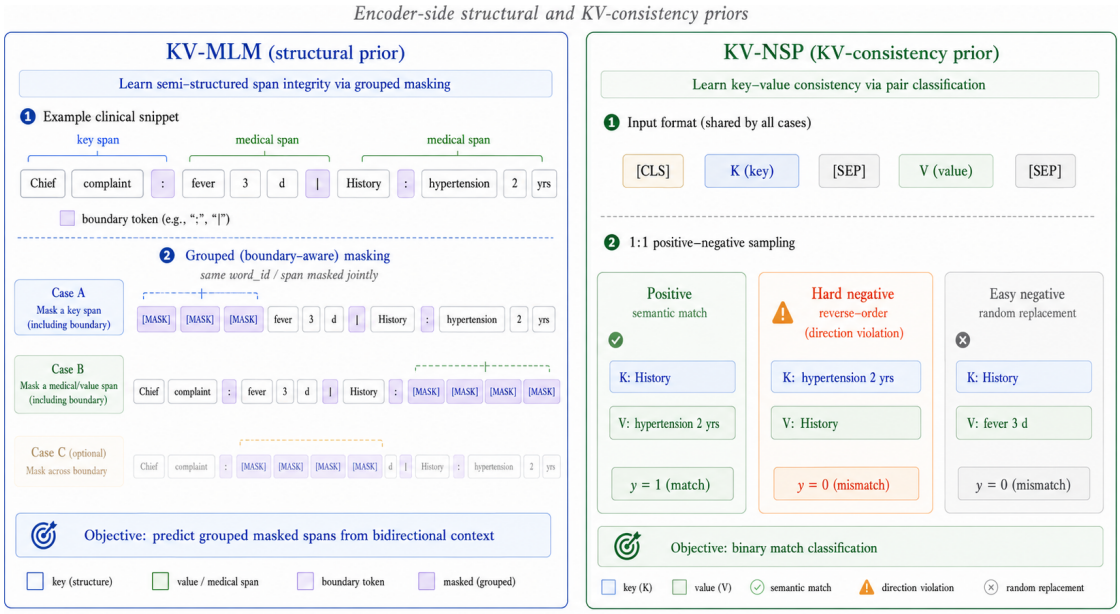

Figure 4: Encoder-side priors in KV-BERT: boundary-aware KV-MLM masking and directional KV-NSP key–value matching.

Comparison	Δ_A	Δ_B	Δ_{all}	Sign kept?
NoiseEmb: full − w/o NoiseEmb	+0.024	+0.015	+0.019	✓
Hybrid: N4 − full	+0.019	+0.001	+0.009	✓
NCAG clean cost: full − N3	+0.028	+0.044	+0.037	✓
Scale: 1.7B full − Qwen3-1.7B+LoRA (seed-1)	+0.011	+0.006	+0.008	✓
SNAP-KV vs. LoRA @0.6B	−0.006	−0.003	−0.005	✓

Table 6: Subset sensitivity (A/B halves, $n = 159/189$). Sign is preserved on both halves for all five headline contrasts, ruling out subset-specific cherry-picking.

	Synthetic OCR noise level						
Variant	0%	5%	10%	20%	30%	50%	**R@30**
Qwen3-0.6B + LoRA	0.717	0.604	0.515	0.362	0.233	0.140	32.5%
HuatuoGPT-3-32B + LoRA	0.730	0.625	0.547	0.423	0.370	0.173	**50.7%**
KV-LLM-0.6B w/o NoiseEmb (N1)	0.631	0.580	0.514	0.365	0.245	0.152	38.9%
KV-LLM-0.6B additive (N2, *ref.*)†	$0.725_{\pm 12}$	$0.604_{\pm 14}$	$0.525_{\pm 5}$	$0.373_{\pm 9}$	$0.248_{\pm 9}$	$0.143_{\pm 11}$	34.2%
KV-LLM-0.6B NCAG-only (N3)†	$0.674_{\pm 13}$	$0.567_{\pm 23}$	$0.485_{\pm 18}$	$0.358_{\pm 4}$	$\mathbf{0.270}_{\pm 4}$	$0.144_{\pm 25}$	**40.1%**
KV-LLM-0.6B additive+NCAG (N4)‡	$0.701_{\pm 23}$	$0.606_{\pm 21}$	$0.501_{\pm 18}$	$0.364_{\pm 16}$	$0.243_{\pm 13}$	$0.139_{\pm 13}$	34.7%
Record-clustered bootstrap: N3 − additive, seed-averaged within-record diffs over 5 seeds							
ΔKe F1 (N3−N2)	−0.051	−0.036	−0.040	−0.015	**+0.022**	+0.001	—
p_{cluster} (two-sided)	0.001^{***}	0.014^{*}	0.009^{**}	0.227	$\mathbf{0.055}^{\dagger}$	0.894	—
$p_{\text{seed-}t}$ (t-test, df=4)	0.009^{**}	0.051	0.013^{*}	0.063	$\mathbf{0.021}^{*}$	0.907	—
p_{Holm} (6 cells)	$\mathbf{0.005}^{**}$	0.057	$\mathbf{0.045}^{*}$	0.454	0.275	0.894	—

Table 7: Key-side synthetic-noise grid (Ke F1). R@30 = Ke@30%/Ke@0%.

	Synthetic OCR noise level					
Variant	0%	5%	10%	20%	30%	50%
KV-LLM-0.6B additive (N2)†	$0.647_{\pm 11}$	$0.437_{\pm 13}$	$0.321_{\pm 4}$	$0.135_{\pm 3}$	$0.061_{\pm 1}$	$0.025_{\pm 1}$
KV-LLM-0.6B NCAG (N3)†	$0.604_{\pm 7}$	$0.416_{\pm 15}$	$0.299_{\pm 16}$	$0.142_{\pm 5}$	$\mathbf{0.078}_{\pm 6}$	$0.027_{\pm 9}$
KV-LLM-0.6B N4‡	$0.624_{\pm 13}$	$0.438_{\pm 12}$	$0.309_{\pm 13}$	$0.137_{\pm 5}$	$0.058_{\pm 5}$	$0.023_{\pm 6}$
Record-clustered bootstrap: N3 − additive, KeVa/KaVa F1, n=100 records, 5 seeds						
Δ Value F1	−0.043	−0.021	−0.022	+0.007	**+0.018**	+0.002
p_{cluster}	0.001^{***}	0.102	0.035^{*}	0.391	$\mathbf{0.004}^{**}$	0.605
$p_{\text{seed-}t}$ (df= 4)	0.004^{**}	0.060	0.051	0.085	$\mathbf{0.009}^{**}$	0.726
p_{Holm}	$\mathbf{0.005}^{**}$	0.408	0.176	0.808	$\mathbf{0.026}^{*}$	0.808

Table 8: Value-side synthetic-noise grid (Task 2/3 F1, with the uniform KeVa/KaVa scorer of §5.1). The 30% cell is the headline endpoint (only Holm-surviving cell in NCAG's favour); subscripts again denote standard deviations (10^{-3}).

1D confidence-only additive control. To test whether the robustness gain comes merely from injecting *some* OCR confidence signal, rather than from the full 7-D feature family, we trained a 1-D additive variant that uses only the bucketed per-token mean OCR confidence. The 1-D, 7-D additive, and NCAG variants share the same CPT recipe and differ only in the noise prior. The 1-D control reproduces the same clean cost / robustness gain trade-off as NCAG: relative to 7-D additive, it is lower on clean text (0.669 ± 26 vs. 0.728 ± 5 at 0% noise), slightly higher at severe noise (0.169 ± 19 vs. 0.147 ± 8 at 50% noise), and still below NCAG at the main 30% endpoint (0.249 ± 8 vs. $\mathbf{0.273} \pm \mathbf{3}$). This supports the claim that the trade-off is intrinsic to noise-conditioning during CPT rather than specific to the 7-D feature family, while also showing that NCAG is the better operating point at moderate corruption.