

Session-Guided Pre-training for Query Understanding: Constraining Mask Prediction and Improving Query Fine-tunability

Sosuke Nishikawa

The University of Tokyo

LY Corporation

sonishik@iis.u-tokyo.ac.jp

sonishik@lycorp.co.jp

Naoki Yoshinaga

Institute of Industrial Science,

The University of Tokyo

ynaga@iis.u-tokyo.ac.jp

Abstract

Large-scale search systems must process vast numbers of diverse, short queries for query understanding (QU) tasks such as named entity recognition (NER) and query classification, under strict latency constraints. Yet general-purpose language models (LMs), trained largely on context-rich open-domain text, are poorly matched to context-poor queries; large generative LMs also incur high serving costs. To achieve accurate and efficient QU, we introduce SEG-QU, a session-guided continual pre-training framework that tailors standard encoder pre-training objectives to search queries. It jointly optimizes token- and query-level objectives over session-derived query pairs: Session-Guided Masked Language Modeling (SG-MLM) reduces uncertainty in single-query MLM, while Same-Session Query Prediction (SSQP) improves downstream fine-tunability. On commercial-search NER and classification tasks, SEG-QU outperforms encoder baselines and fine-tuned generative LMs up to $43\times$ larger, while requiring substantially lower latency and energy. SEG-QU also transfers effectively to a public e-commerce benchmark. Our code is publicly available at <https://github.com/lycorp-jp/segqu>.

1 Introduction

Accurately interpreting short, context-poor queries (Query Understanding; QU) is a core challenge in search engines. Since even marginal increases in response time can degrade user experience (Brutlag, 2009; Arapakis et al., 2014), efficient encoder-only transformer models are still widely adopted as deployed QU models: bidirectional pre-training allows direct fine-tuning (FT) for discriminative language understanding tasks (Suganthan et al., 2025; Gisserot-Boukhlef et al., 2026), and single-pass inference achieves low latency at modest serving cost.

Pre-training encoder-only models for QU remains challenging. The standard pre-training objec-

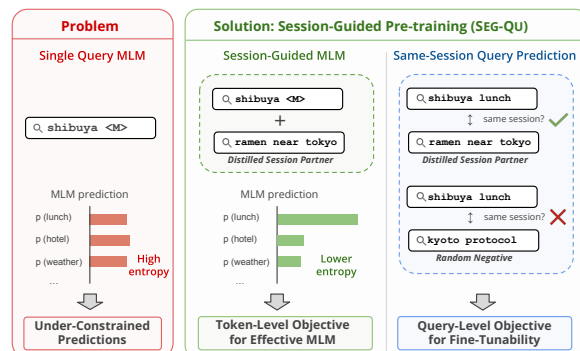


Figure 1: Overview of SEG-QU. Standard MLM on a short query is under-constrained; SEG-QU pairs it with a session partner and jointly trains SG-MLM and SSQP.

tive, Masked Language Modeling (MLM) (Devlin et al., 2019), relies on the distributional hypothesis that a word is characterized by the company it keeps (Harris, 1954). However, unlike formal texts such as Wikipedia articles, search queries are short and often lack sufficient surrounding tokens. This context starvation makes it hard for MLM to learn meaningful representations from queries alone (Xie et al., 2023; Jiang et al., 2022). Prior work mitigates this by supplying additional context, either by synthetic extension (Jiang et al., 2022) or by exploiting click-through logs such as clicked items and co-clicked queries (Li et al., 2023; Xie et al., 2023; Peng et al., 2023). Yet such context only indirectly reflects a query’s linguistic company, and is unavailable for click-less queries because they surface no satisfactory result.

To address these limitations, we propose SEG-QU (Figure 1), a continual pre-training framework for QU encoders that leverages search sessions as a direct and click-independent source of distributional context. Search sessions, *i.e.*, sequences of queries issued within a bounded time window, record the reformulations and related terms a user produces during search, offering a more direct source of a query’s linguistic company. Because

sessions accumulate regardless of click outcomes, they remain available even for click-less queries.¹ In this study, we distill reliable query pairs from raw sessions and jointly train two complementary objectives: Session-Guided MLM (SG-MLM), which adds a session partner as context for mask prediction, and Same-Session Query Prediction (SSQP), which predicts whether two queries share a session. These objectives are designed to provide complementary supervision for token- and query-level QU, respectively.

Experiments on session logs from a commercial web search engine show that our approach, applied as continual pre-training on Japanese ModernBERT (Warner et al., 2025; Tsukagoshi et al., 2025), consistently outperforms the base model and all single-query baselines in query named entity recognition (NER) and classification (§ 5). Analysis reveals why: (1) SG-MLM restores an effective learning signal by shifting mask prediction entropy toward general-domain levels (§ 6.1), and (2) BERT-style joint training with SSQP provides an initialization that adapts well under fine-tuning (§ 6.2). Furthermore, the resulting encoders outperform generative, instruction-following LLMs up to $43\times$ their size with $72\times$ lower p95 latency at 1.1% of the energy cost (§ 6.3), and session and click-through signals combine for further gains (§ 6.4). Finally, SEG-QU transfers to a public Japanese e-commerce relevance benchmark (§ 6.5).

2 Related Work

We review pre-training strategies for QU encoders and prior uses of search sessions in information retrieval (IR).

Query Understanding QU encompasses a range of language processing tasks applied to search queries, such as query classification and NER (Shen et al., 2006; Guo et al., 2009). To counter the context starvation of search queries, existing approaches supply additional context at two levels. At the token level, Jiang et al. (2022) introduce extended token classification (ETC), which inserts tokens instead of masking and detects them, while other work concatenates clicked items as MLM context (Xie et al., 2023; Peng et al., 2023). At the query level, co-click query prediction aligns representations of queries that led to clicks on the same item (Li et al., 2023; Xie et al., 2023). These

methods largely rely on click-through logs, which are unavailable for click-less queries; search sessions remain underexplored as click-independent distributional context for pre-training QU encoders.

Exploiting Search Sessions Search sessions have predominantly been used as dynamic context at inference time in IR: in session search, multi-turn query history is incorporated to disambiguate the current query and improve retrieval (Cheng et al., 2021; Feng et al., 2022; Yu et al., 2021), while in QU tasks such as query rewriting and suggestion, session context similarly helps resolve ambiguity when generating queries (Sordoni et al., 2015; Zuo et al., 2023; Nishikawa et al., 2026). In contrast, SEG-QU uses sessions as offline pre-training signals for QU encoders, rather than as inference-time context (Appendix A).

Related work learns session-informed query embeddings through multi-view learning (Ortiz et al., 2022) or contrastive learning (Nishikawa et al., 2025). In contrast, SEG-QU continually pre-trains the encoder itself at both token and query levels, yielding a fine-tunable initialization for diverse downstream QU tasks.

3 Methodology

We propose a pre-training framework for QU encoders that exploits session structure. First, we distill reliable query associations from raw session logs by extracting co-occurring query pairs (§ 3.1). Then, using these pairs, we jointly optimize complementary objectives that target the token and query levels of downstream QU tasks: *Session-Guided MLM* (SG-MLM; § 3.2), a token-level objective that adds a session partner as context for MLM, and *Same-Session Query Prediction* (SSQP; § 3.3), a query-level objective that predicts whether two queries originate from the same session.

3.1 Distilling Pairs from Sessions

Raw session logs are noisy (Sordoni et al., 2015): intent shifts, navigational queries, and single-query sessions dilute pre-training. To extract reliable query associations, we aggregate transition statistics across all sessions rather than relying on any single user’s trajectory (Huang et al., 2003; Nishikawa et al., 2025). Concretely, for every ordered pair of consecutive distinct queries (q_i, q_j) within the same session, we count each transition at most once per user and compute Normalized

¹In our logs (§ 4.1), 71% of click-less queries co-occur with at least one other distinct query in a session.

Pointwise Mutual Information (NPMI):

$$\begin{aligned} \text{NPMI}(q_i \rightarrow q_j) &= \frac{\text{PMI}(q_i \rightarrow q_j)}{-\log p(q_i \rightarrow q_j)}, \\ \text{PMI}(q_i \rightarrow q_j) &= \log \frac{p(q_i \rightarrow q_j)}{p_{\text{src}}(q_i) p_{\text{tgt}}(q_j)}. \end{aligned} \quad (1)$$

Here, $p(q_i \rightarrow q_j)$ is the fraction of counted directed transitions from q_i to q_j , while $p_{\text{src}}(q_i)$ and $p_{\text{tgt}}(q_j)$ are the corresponding source and target marginals, respectively. We retain pairs whose NPMI exceeds a threshold δ . NPMI suppresses high-frequency navigational queries (e.g., portal URLs), which occur broadly across transitions but carry little discriminative context.

3.2 Session-Guided MLM

Search queries suffer from context starvation under standard MLM: for instance, predicting the mask in a query like “Shibuya [MASK]” is virtually unconstrained, as the answer could be *lunch*, *weather*, *hotel*, or countless other completions.

To address this, we modify the input for MLM to operate on session pairs rather than isolated queries. Given a pair (q_i, q_j) extracted via NPMI (§ 3.1), we concatenate them with a separator token and apply masking to both queries. Formally, SG-MLM minimizes the standard MLM loss computed over the masked positions in both queries:

$$\mathcal{L}_{\text{SG-MLM}} = - \sum_{t \in \mathcal{M}} \log p_{\theta}(x_t | \tilde{\mathbf{x}}_{i \oplus j}), \quad (2)$$

where x_t is the original token at masked position t , $\mathcal{M}$ is the set of masked positions, $\tilde{\mathbf{x}}_{i \oplus j}$ denotes the [SEP]-concatenated input with masks applied, and θ represents the model parameters. Note that the encoder accepts a single query at inference; the session partner serves solely as a guidance signal during pre-training.²

3.3 Same-Session Query Prediction

Following prior QU pre-training that adds query-level objectives to address context starvation (Li et al., 2023; Xie et al., 2023), we predict whether two queries originate from the same session. Given a pair (q_i, q_j) extracted via NPMI (§ 3.1), we use it as a positive example ($y=1$) and construct a negative example ($y=0$) by replacing q_j with a randomly sampled query q_r . Let $\tilde{\mathbf{x}}$ denote $\tilde{\mathbf{x}}_{i \oplus j}$ when $y=1$ and

$\tilde{\mathbf{x}}_{i \oplus r}$ when $y=0$. A binary classifier over the [CLS] representation predicts $y \in \{0, 1\}$:

$$\mathcal{L}_{\text{SSQP}} = -\log p_{\theta}(y | \tilde{\mathbf{x}}). \quad (3)$$

Although prior work adopts contrastive query-level objectives, this formulation revisits BERT-style Next Sentence Prediction (Devlin et al., 2019), which is widely considered ineffective in general-domain pre-training (Liu et al., 2019; Lan et al., 2020). Despite this, the resulting joint objective improves downstream performance (§ 5), which we attribute to its high fine-tunability (§ 6.2).

Following BERT-style joint training, both losses use the same input $\tilde{\mathbf{x}}$. The final objective is:

$$\mathcal{L}_{\text{SEG-QU}} = - \sum_{t \in \mathcal{M}} \log p_{\theta}(x_t | \tilde{\mathbf{x}}) + \mathcal{L}_{\text{SSQP}}. \quad (4)$$

For positive instances, the first term corresponds to SG-MLM in Eq. (2); for negative instances, it is computed using a randomly sampled query as context. We refer to the resulting pre-training framework as SEG-QU (SEssion-Guided pre-training for QU encoders).

4 Experimental Setup

We describe the pre-training setup (§ 4.1), baselines and ablations (§ 4.2), and downstream tasks (§ 4.3). Detailed settings are provided in Appendix B.

4.1 Pre-training Setup

We use session logs from Yahoo! JAPAN Search,³ one of the largest search engines in Japan. The logs span November–December 2025, and we segment sessions using a 15-minute inactivity timeout (Göker and He, 2000). We extract 100M query pairs whose NPMI exceeds threshold $\delta = 0.5$ (selected on dev macro- F_1), covering 68M unique queries. Appendix B examines the session timeout and NPMI threshold choices.

We continually pre-train ModernBERT-ja-130m.⁴ All MLM-based pre-training uses a mask rate of 0.30 (Warner et al., 2025).⁵ For SSQP, positive and negative examples are sampled 1:1, with negatives randomly drawn from the extracted pairs. For SEG-QU, the model is trained for 5 epochs with a batch size of 2,048 and a peak learning rate of 1e-4, using AdamW with linear decay following a warmup over the first 1% of training steps.

³<https://search.yahoo.co.jp/>

⁴<https://huggingface.co/sbintuitions/models>

⁵Reducing the mask rate for single-query MLM to ease prediction difficulty degraded downstream performance.

²Concatenating three or more session queries does not yield further gains, likely because a single session partner provides enough context for MLM (§ 6.1).

4.2 Baselines and Ablations

We compare SEG-QU against ModernBERT-ja-130m (its base encoder without continual pre-training), SG-MLM, and the following continual pre-training baselines:

Single-Query Pre-training Four baselines learn from individual queries without session context. **Session-Derived MLM** flattens the NPMI-extracted session pairs into queries and trains with MLM, matching our method in the set of queries seen during pre-training. **Full-Log MLM** trains with MLM on the most frequent unique queries from the entire search log, matching our method in the number of unique queries. **Full-Log MLM w/ SimCSE** extends Full-Log MLM with an unsupervised SimCSE objective (Gao et al., 2021) to strengthen query-level representations. **ETC** (Jiang et al., 2022) detects tokens inserted into queries, using the same query set as Full-Log MLM.

Session-Aware Pre-training These methods leverage search sessions for pre-training. **UBIQUE** (Nishikawa et al., 2025) is our re-implementation of its session-based variant on the same raw sessions as ours. **SG-MLM w/ random pairing** concatenates randomly paired queries drawn from the NPMI-extracted query set, testing whether the gains require session-based pairing or merely concatenating two queries. **SG-MLM w/o NPMI pairing** concatenates all queries within a raw session without NPMI selection, testing whether statistical pair selection is necessary. **SG-MLM w/ SSQP_{cl}** jointly optimizes SG-MLM with an InfoNCE (van den Oord et al., 2019) objective over session pairs, providing a contrastive joint-training variant. **SEG-QU w/o SSQP** uses the same balanced positive and negative inputs as SEG-QU but omits the SSQP loss, isolating the contribution of the binary query-level objective.

Except UBIQUE, all continual pre-training models start from ModernBERT-ja-130m and use *the same expected number of masked-token prediction targets* as SEG-QU, with training steps adjusted accordingly. We exclude click-based approaches, as limited click coverage and the distribution shift prevent a fair comparison. We instead examine click signals as a complementary source in § 6.4.

4.3 Downstream Tasks

Our main evaluation covers two internal QU tasks at complementary representation levels: query

Setting	NER	Class.	Avg.
<i>General-Domain Baseline</i>			
ModernBERT	41.99	65.04	53.52
<i>Single-Query Pre-training</i>			
Session-Derived MLM	44.89	74.17	59.53
Full-Log MLM	48.28	74.72	61.50
w/ SimCSE	49.47	74.73	62.10
ETC (Jiang et al., 2022)	48.27	74.76	61.52
<i>Session-Aware Pre-training</i>			
UBIQUE (Nishikawa et al., 2025)	41.69	63.93	52.81
SG-MLM	54.78	75.82	65.30
w/ random pairing	51.10	73.23	62.17
w/o NPMI pairing	52.38	73.58	62.98
w/ SSQP _{cl}	53.60	77.01	65.31
SEG-QU	56.99	78.48	67.74
w/o SSQP	55.08	77.60	66.34

Table 1: Main results on QU tasks (macro-F₁).

NER (token-level) and query classification (query-level). These tasks can support core search applications, including relevance ranking, targeted advertising, entity linking to knowledge bases, and routing to verticals such as local or shopping search.

Query NER identifies and classifies entity spans in queries (e.g., *ramen near Shibuya station* → [(Shibuya station, STATION), (ramen, GENRE)]). We formulate it as BIO sequence labeling with a linear layer over token representations followed by a softmax, trained with cross-entropy loss. The dataset contains 7,298 queries from 2023, manually annotated with 13 entity types, split into 3,651/1,457/2,190 for train/dev/test.

Query Classification assigns one or more topic categories to a given query (e.g., *ramen near Shibuya station* → [GOURMET]). We frame it as multi-label classification with a linear layer over the [CLS] representation followed by a sigmoid activation, trained with binary cross-entropy loss. The dataset contains 15,623 queries from 2014 to 2024, manually annotated with 53 categories, split into 7,813/3,112/4,698 for train/dev/test.

We fine-tune all encoder models for 10 epochs with a learning rate of 5e-5 and a batch size of 64, using three random seeds. We report the average macro-F₁ to account for class imbalance.

5 Results

Table 1 shows the main results on the two QU tasks. SEG-QU achieves the best performance on both, outperforming ModernBERT and the strongest single-query baseline by 14.22 and 5.64 average macro-F₁ points, respectively. SG-MLM alone

already surpasses every single-query baseline by >5.3 on NER ($p < 0.001$)⁶ and >1.0 on classification ($p < 0.05$), which we attribute to the session partner restoring mask prediction entropy to general-domain levels (§ 6.1). The additional SEG-QU gain is particularly notable on classification (+2.66 over SG-MLM; $p < 0.001$), where SG-MLM alone improved little. UBIQUE underperforms the other encoder baselines, consistent with its goal of producing frozen retrieval embeddings rather than a fine-tunable initialization.

Component analyses examine each design choice. Replacing session partners with random queries (SG-MLM w/ random pairing) reduces average macro- F_1 by 3.13, showing that the gains require session-based pairs rather than arbitrary concatenation. Concatenating queries from raw sessions (SG-MLM w/o NPMI pairing) reduces it by 2.32, indicating that statistical pair selection is needed to suppress noisy transitions. The contrastive joint-training variant (SG-MLM w/ SSQP_{cl}) performs on par with SG-MLM alone, indicating that the BERT-style joint formulation fits better here; we analyze its fine-tunability in § 6.2.

Removing the SSQP loss from SEG-QU (SEG-QU w/o SSQP) lowers average macro- F_1 by 1.40, suggesting that the loss contributes beyond balanced-pair MLM. Interestingly, this variant outperforms SG-MLM by 1.04, indicating that replacing half the session partners with random queries is itself beneficial. Random-pair MLM may regularize the encoder by discouraging over-reliance on session context and reducing mismatch with single-query downstream inputs. Beyond this ablation, click-stratified analysis confirms consistent SEG-QU gains on both click-rich and click-poor test subsets (Appendix C).

6 Further Evaluation and Analysis

We analyze why SEG-QU works (§ 6.1, § 6.2), compare it against larger LLMs (§ 6.3), test its complementarity with click-through signals (§ 6.4), and evaluate transfer to a public benchmark (§ 6.5).

6.1 SG-MLM Constrains Mask Prediction

SG-MLM outperforms single-query baselines under the same token budget (§ 5). We attribute this to the session partner supplying context that constrains mask prediction well, and test this by mea-

⁶All p-values are calculated via a paired permutation test with $n = 100,000$ permutations (Smucker et al., 2007).

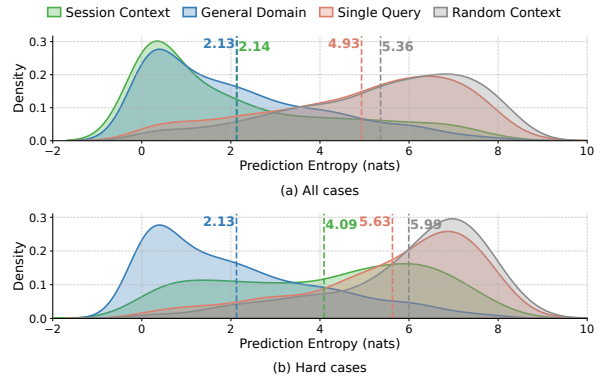


Figure 2: Entropy distribution of masked token predictions under different input conditions. Density below zero reflects finite kernel density estimate bandwidth.

suring the entropy of the predicted distribution at masked positions as a proxy for prediction uncertainty. We sample 1,000 queries from the NPMI-distilled pairs (§ 3.1) and mask a single token at a random position in each query.⁷ We then compute the entropy ($H = -\sum_{i=1}^{|V|} p_i \log p_i$) of the predicted distribution at the masked position using the base ModernBERT under three input conditions: **Single Query**, the masked query alone; **Session Context**, the query concatenated with its session partner; and **Random Context**, the query concatenated with a randomly sampled query. As a reference, **General Domain** masks one random token in Wikipedia sentences, a difficulty regime under which MLM is known to be effective (Devlin et al., 2019; Liu et al., 2019).

Figure 2 shows that Single Query yields high entropy (mean 4.93 nats), reflecting high-uncertainty predictions under context starvation. Adding the session partner moves the distribution onto the general-domain reference (mean difference 0.01, 95% CI $[-0.17, 0.20]$, Welch’s t -test), while Random Context pushes it further away (mean 5.36),⁸ confirming that semantically related context is required. This effect is not the model copying the answer from the partner: restricting to hard cases where the ground-truth token is absent from the session partner ($\sim 47\%$ of instances), session context still cuts mean entropy from 5.63 to 4.09 (Figure 2b). The effect therefore rests on semantic association rather than surface overlap.

⁷Entropy is measured per masked position, so we mask one token per input to keep it comparable across lengths.

⁸We attribute this to the unrelated query introducing a conflicting intent that disperses the predicted distribution.

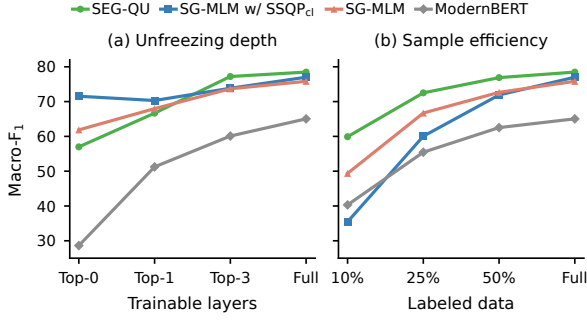


Figure 3: Classification macro-F₁ across (a) FT depth (top-0 = linear probe) and (b) labeled-data fraction.

6.2 BERT-Style Joint Training with SSQP Improves Fine-Tunability

Query-level objectives in QU pre-training are typically contrastive (Li et al., 2023; Xie et al., 2023), yet our BERT-style joint objective improves query classification performance over both SG-MLM alone and the contrastive variant, SG-MLM w/ SSQP_{cl} (§ 5). We hypothesize that this gain stems from fine-tunability rather than frozen embedding quality. We assess fine-tunability along two dimensions: (1) the performance attainable through full-parameter fine-tuning and (2) label efficiency across FT set sizes.

Unfreezing-Depth Sweep We fine-tune only the top- k layers, varying k from 0 (a linear probe over the frozen [CLS] representation) to full fine-tuning. Figure 3 reveals a crossover: under the linear probe, SG-MLM w/ SSQP_{cl} leads (71.54) while SEG-QU trails (56.99), below SG-MLM alone (61.87); this ordering reverses with depth, and SEG-QU becomes best under full fine-tuning. From top-0 to full it gains 21.49 points, against 5.47 for SG-MLM w/ SSQP_{cl}. Frozen embedding quality and fine-tunability thus dissociate: the contrastive variant yields the strongest frozen embeddings but adapts little, whereas SEG-QU is weakest among the session-based variants yet adapts the most.

Sample-Size Sweep A natural concern is whether this adaptation requires more FT data. We sub-sample the classification training set at 10%, 25%, 50%, and 100%. Figure 3(b) shows that SEG-QU dominates at every fraction, and the gap widens as data shrinks: +1.47 points over SG-MLM w/ SSQP_{cl} at full data, +4.98 at 50%, +12.46 at 25%, and +24.46 at 10%, where SG-MLM w/ SSQP_{cl} degrades sharply (35.45).

SEG-QU thus yields strong FT adaptation (even

Setting	Avg. $\uparrow$	Latency (ms) $\downarrow$		Energy $\downarrow$
	(macro-F ₁)	mean	p95	(J/q)
LLMs (generative)				
Sarashina2.2 _{0.5B}	56.50	25.78	35.04	9.68
Sarashina2.2 _{1B}	59.56	36.68	50.13	18.71
Sarashina2.2 _{3B}	61.82	63.65	91.38	27.75
Encoder Models				
ModernBERT _{130M} [†]	53.52	1.42	1.65	0.37
SEG-QU _{70M}	63.27	1.10	1.26	0.30
SEG-QU _{130M} [†]	67.74	1.42	1.65	0.37
SEG-QU _{310M}	70.07	1.91	2.06	0.57

Table 2: Task performance and serving efficiency. [†]SEG-QU_{130M} shares ModernBERT_{130M}’s architecture, so their inference costs are identical.

on limited data) despite weaker frozen embeddings. This aligns with Gao et al. (2021), who find that sentence-level contrastive learning does not necessarily improve downstream transfer. One possible explanation for why BERT-style SSQP is effective despite NSP’s limited utility in general-domain pre-training is that NPMI-distilled session pairs provide more discriminative supervision (Appendix D).

6.3 Comparison to LLMs

To quantify the practical advantage of encoders over LLMs (§ 1), we train SEG-QU at three sizes (70M, 130M, and 310M)⁴ and compare them against the instruction-tuned Sarashina2.2 models (0.5B, 1B, and 3B),⁴ Japanese LLMs further fine-tuned on the same downstream tasks (§ 4.3) and evaluated generatively.⁹ We measure mean and p95 latency as well as energy cost per query on the classification test set (§4.3) using CodeCarbon.¹⁰ Detailed per-task results are reported in Appendix E (Table 9). All measurements use batch size 1 on a single NVIDIA A100 SXM4 GPU to simulate online serving, where queries arrive individually.

Table 2 shows that all three SEG-QU variants outperform all LLM baselines in task performance despite being up to 43 $\times$ smaller. Even SEG-QU_{70M} surpasses Sarashina2.2-3B in average F₁ (63.27 vs. 61.82) while achieving 72 $\times$ lower p95 latency at 1.1% of the energy cost. Scaling to SEG-QU_{310M} raises performance to 70.07 F₁ (+6.80 over SEG-QU_{70M}) with only small absolute increases in p95 latency (+0.80 ms) and energy (+0.27 J/q). In summary, SEG-QU encoders outperform generative

⁹We did not continue causal-LM pre-training on search queries because doing so yielded no downstream improvement, possibly due to degraded instruction-following ability.

¹⁰<https://github.com/mlco2/codecarbon>

Setting	NER	Class.	Avg.
ModernBERT	41.99	65.04	53.52
Click Only	50.54	72.58	61.56
SEG-QU	56.99	78.48	67.74
SEG-QU $\rightarrow$ Session (control)	57.17	78.08	67.63
SEG-QU $\rightarrow$ Click	58.43	78.47	68.45

Table 3: Complementarity of session and click-through signals. All metrics are macro-F₁.

LLMs up to $43\times$ their size while remaining practical to deploy in production.

6.4 Combination with Click-Through Signals

Click-through logs offer supervision complementary to sessions: clicked items add content absent from the search box, and co-clicked queries link different users’ expressions of the same intent. We test whether sessions and clicks combine additively.

We adopt two click-based objectives: clicked title MLM, which applies MLM to a query concatenated with its clicked page title (Xie et al., 2023; Peng et al., 2023), and co-click query prediction, which aligns representations of co-clicked queries via contrastive learning (Li et al., 2023; Xie et al., 2023). **Click Only** trains both from ModernBERT on 10M click-derived instances. Following stage-wise continual pre-training over heterogeneous sources (Gururangan et al., 2020), **SEG-QU $\rightarrow$ Click** applies them as a second stage on the SEG-QU checkpoint; the control **SEG-QU $\rightarrow$ Session** instead continues on session data at a matched token budget and configuration (Appendix F).

Table 3 shows that Click Only improves over ModernBERT (+8.04), and adding it to SEG-QU reaches the highest average (+0.71 over SEG-QU), while the control SEG-QU $\rightarrow$ Session does not; the gain therefore stems from click signals, not added steps. It concentrates on NER (+1.44), with classification flat, likely because SEG-QU already improves it (§ 5). These results suggest that session and click-through signals are complementary.

6.5 Transfer to a Public E-Commerce Benchmark

We evaluate whether the gains from SEG-QU transfer beyond our internal, in-domain NER and query classification tasks. We use the Japanese ESCI benchmark (Reddy et al., 2022), whose relevance task requires relating a short query to a concise product title. We formulate it as four-way classification over EXACT, SUBSTITUTE, COMPLE-

Setting	Macro-F ₁
ModernBERT	53.20
Full-Log MLM w/ SimCSE	51.81
UBIQUE	51.15
SEG-QU	55.36

Table 4: Results on Japanese ESCI relevance prediction. All models are based on ModernBERT-ja-130m.

MENT, and IRRELEVANT, concatenating the two texts with a [SEP] token for joint encoding. We use the official train and test splits and divide the 7,284 unique training queries 9:1 into 6,556 training and 728 development queries. The test set contains 3,123 queries. We fine-tune all models under the same configuration (Appendix G) with three random seeds and report mean macro-F₁.

Table 4 shows that SEG-QU outperforms ModernBERT by 2.16 macro-F₁ points ($p < 0.001$). It also outperforms the strongest single-query baseline, Full-Log MLM w/ SimCSE, and the session-based UBIQUE baseline. Full-Log MLM w/ SimCSE performs below ModernBERT, suggesting that query-only continual pre-training does not necessarily improve transfer (Xie et al., 2023). SEG-QU therefore transfers beyond our internal, in-domain evaluation to a public e-commerce task.

7 Conclusion

We presented SEG-QU, a simple session-guided pre-training framework for QU encoders. Using session partners distilled from raw sessions, it combines SG-MLM with SSQP through BERT-style joint training. Analysis shows that session partners restore an effective learning signal for MLM, while this joint formulation yields a highly fine-tunable initialization.

SEG-QU is practical to deploy: sessions are a natural byproduct of any search system; the framework requires no architectural changes; SEG-QU encoders outperform generative LLMs up to $43\times$ their size at far lower latency and energy cost; and it yields further gains when combined with click-through signals. Its benefits also transfer to a public Japanese e-commerce relevance benchmark.

Future work includes extending session-guided pre-training beyond within-session transitions and dynamically leveraging session context during inference for discriminative QU tasks.

Limitations

Our pre-training experiments use session logs from a single Japanese commercial web search engine. Although SEG-QU transfers to a public Japanese e-commerce relevance task (§ 6.5), cross-lingual and cross-engine generalization remains unverified.

Our evaluation covers query NER, query classification, and query-product relevance, all discriminative tasks well suited to encoder-only models; whether SEG-QU benefits generative QU tasks such as query rewriting and suggestion, which typically rely on decoder-based architectures, is an open question. Relatedly, our LLM comparison (§ 6.3) covers a single model family (Sarashina2.2) at up to 3B parameters, evaluated only in the generative instruction-following formulation, the common way LLMs are applied to QU. Discriminatively adapted decoders (Suganthan et al., 2025) and larger LLMs are left to future work.

Ethics Statement

Throughout the construction of pre-training data from search session logs (§ 4.1) and evaluation datasets (§ 4.3), user identifiers were pseudonymized before use. Specifically, session user IDs were replaced with one-way hashed identifiers before dataset construction, and researchers had no access to the original identifiers. All annotation was performed by in-house reviewers without access to user identifiers. All example queries in the running text are constructed for illustration and do not reproduce specific user queries from the logs.

References

- Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. 2014. [Impact of response latency on user behavior in web search](#). In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 103–112, New York, NY, USA. Association for Computing Machinery.
- Jake Brutlag. 2009. [Speed matters for Google web search](#). Technical report, Google Inc. Accessed: 2025-09-13.
- Qiannan Cheng, Zhaochun Ren, Yujie Lin, Pengjie Ren, Zhumin Chen, Xiangyuan Liu, and Maarten de Rijke. 2021. [Long short-term session search: Joint personalized reranking and next query prediction](#). In *Proceedings of the Web Conference 2021*, WWW '21, page 239–248, New York, NY, USA. Association for Computing Machinery.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [LLM.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rui Feng, Chen Luo, Qingyu Yin, Bing Yin, Tuo Zhao, and Chao Zhang. 2022. [CERES: Pretraining of graph-conditioned transformer for semi-structured session data](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 219–230, Seattle, United States. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hippolyte Gisserot-Boukhlef, Nicolas Boizard, Manuel Faysse, Duarte Miguel Alves, Emmanuel Malherbe, André F. T. Martins, Céline Hudelot, and Pierre Colombo. 2026. [Should we still pretrain encoders with masked language modeling?](#) In *International Conference on Learning Representations*.
- Ayse Göker and Daqing He. 2000. [Analysing web search logs to determine session boundaries for user-oriented learning](#). In *Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000)*, volume 1892 of *Lecture Notes in Computer Science*, pages 319–322, Trento, Italy. Springer.
- Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. [Named entity recognition in query](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.
- 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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

- Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. 2003. [Relevant term suggestion in interactive web search based on contextual information in query session logs](#). *J. Am. Soc. Inf. Sci. Technol.*, 54(7):638–649.
- Haoming Jiang, Tianyu Cao, Zheng Li, Chen Luo, Xi-anfeng Tang, Qingyu Yin, Danqing Zhang, Rahul Goutam, and Bing Yin. 2022. [Short text pre-training with extended token classification for e-commerce query understanding](#). *Preprint*, arXiv:2210.03915.
- 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*.
- Juanhui Li, Wei Zeng, Suqi Cheng, Yao Ma, Jiliang Tang, Shuaiqiang Wang, and Dawei Yin. 2023. [Graph enhanced BERT for query understanding](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’23, page 3315–3319, New York, NY, USA. Association for Computing Machinery.
- 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](#). *Preprint*, arXiv:1907.11692.
- Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. [Identifying task-based sessions in search engine query logs](#). In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, page 277–286, New York, NY, USA. Association for Computing Machinery.
- Sosuke Nishikawa, Jun Hirako, Nobuhiro Kaji, Koki Watanabe, Hiroki Asano, Souta Yamashiro, and Shumpei Sano. 2025. [Search query embeddings via user-behavior-driven contrastive learning](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 138–147, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sosuke Nishikawa, Naoki Yoshinaga, and Nobuhiro Kaji. 2026. [Recasting web-scale query suggestion as dense retrieval: Efficient, up-to-date, and context-aware suggestions](#). In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’26, page 1416–1428, New York, NY, USA. Association for Computing Machinery.
- Diego Ortiz, Jose Moreno, Gilles Hubert, Karen Pinel-Sauvagnat, and Lynda Tamine. 2022. [Exploring the value of multi-view learning for session-aware query representation](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 304–315, Seattle, United States. Association for Computational Linguistics.
- Zhiyuan Peng, Vachik Dave, Nicole McNabb, Rahul Sharnagat, Alessandro Magnani, Ciya Liao, Yi Fang, and Sravanthi Rajanala. 2023. [Entity-aware multi-task learning for query understanding at Walmart](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’23, page 4733–4742, New York, NY, USA. Association for Computing Machinery.
- Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. [Shopping queries dataset: A large-scale ESCI benchmark for improving product search](#). *Preprint*, arXiv:2206.06588.
- Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2006. [Building bridges for web query classification](#). In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’06, page 131–138, New York, NY, USA. Association for Computing Machinery.
- Mark D. Smucker, James Allan, and Ben Carterette. 2007. [A comparison of statistical significance tests for information retrieval evaluation](#). In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM ’07, page 623–632, New York, NY, USA. Association for Computing Machinery.
- Alessandro Sordani, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. [A hierarchical recurrent encoder-decoder for generative context-aware query suggestion](#). In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM ’15, page 553–562, New York, NY, USA. Association for Computing Machinery.
- Paul Suganthan, Fedor Moiseev, Le Yan, Junru Wu, Jianmo Ni, Jay Han, Imed Zitouni, Enrique Alfonseca, Xuanhui Wang, and Zhe Dong. 2025. [Adapting decoder-based language models for diverse encoder downstream tasks](#). *Preprint*, arXiv:2503.02656.
- Jaime Teevan, Daniel Ramage, and Meredith Ringel Morris. 2011. [#TwitterSearch: A comparison of microblog search and web search](#). In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, page 35–44, New York, NY, USA. Association for Computing Machinery.
- Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. 2025. [ModernBERT-Ja](#). <https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a>.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jian Xie, Yidan Liang, Jingping Liu, Yanghua Xiao, Baohua Wu, and Shenghua Ni. 2023. [QUERT: Continual pre-training of language model for query understanding in travel domain search](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’23*, page 5282–5291, New York, NY, USA. Association for Computing Machinery.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. [Few-shot conversational dense retrieval](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 829–838, New York, NY, USA. Association for Computing Machinery.

Simiao Zuo, Qingyu Yin, Haoming Jiang, Shaohui Xi, Bing Yin, Chao Zhang, and Tuo Zhao. 2023. [Context-aware query rewriting for improving users’ search experience on E-commerce websites](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 616–628, Toronto, Canada. Association for Computational Linguistics.

A Relation to Session-Aware Models

CERES (Feng et al., 2022) pre-trains a graph-conditioned Transformer over semi-structured e-commerce sessions, jointly modeling query/product text and customer-action relations for session-level product and query retrieval. **SaQuEViL** (Ortiz et al., 2022) constructs and aligns user- and system-view embeddings from query reformulations and retrieved-document rankings, and directly evaluates them through centroid-based query classification and session-based retrieval.

In contrast, SEG-QU continually pre-trains the encoder itself on click-independent query pairs distilled from sessions using token-level and query-level objectives. It uses the session partner only during pre-training to produce a fine-tunable initialization. SEG-QU leaves the base encoder architec-

Timeout (min)	NER	Class.
5	54.80	78.19
15	56.99	78.48
30	57.39	78.60

Table 5: Post-hoc robustness to the inactivity timeout. Values are test-set macro-F₁.

ture unchanged and requires neither session history nor retrieved documents at inference.

B Experimental Details

This appendix details baseline configurations and session-pair construction settings.

Baseline Configurations. For Full-Log MLM w/ SimCSE, we add the unsupervised SimCSE InfoNCE loss to the MLM loss, with a dropout rate of 0.1 and a temperature of 0.05. For ETC, we initialize both the generator and the discriminator from the Full-Log MLM checkpoint and continue training with the ETC objective, using the same MLM-to-ETC step ratio (88.6% to 11.4%) and all other hyperparameters as in Jiang et al. (2022). For UBIQUE (Nishikawa et al., 2025), we follow its filtering procedure to extract 100M query pairs from the session logs in §4.1. We continually pre-train ModernBERT-ja-130m on these pairs, following the hyperparameters of the original paper. For a fair comparison, we full-parameter fine-tune the resulting encoder as for the other models; linear classification over its frozen embeddings performed worse.

Session-Pair Construction Settings. We fixed the inactivity timeout at 15 minutes a priori based on prior evidence. Göker and He (2000) found that an 11–15 minute interval balanced splitting related activities against merging unrelated ones, and subsequent studies have also used or evaluated a 15-minute threshold (Teewan et al., 2011; Lucchese et al., 2011). After fixing this value, we conducted a post-hoc robustness analysis using 5- and 30-minute variants with the same NPMI threshold ($\delta = 0.5$) and 100M-pair budget. Table 5 shows that SEG-QU remains effective across the evaluated timeouts, although the 5-minute setting is somewhat weaker on NER.

We selected the NPMI threshold (§ 3.1) based on dev-set performance. Table 6 reports dev macro-F₁ and retained-pair proportions across candidate

δ	NER	Class.	Avg.	Retained (%)
0.1	50.97	77.62	64.30	97
0.3	51.38	77.11	64.25	88
0.5	51.15	77.77	64.46	72
0.7	–	–	–	39

Table 6: Dev-set macro- F_1 and retained-pair proportions across NPMI thresholds. All reported performance results use a fixed 100M-pair training corpus.

Setting	NER		Class.	
	Rich	Poor	Rich	Poor
ModernBERT	48.59	37.85	65.34	64.51
Click Only	57.59	46.90	74.31	68.57
Full-Log MLM w/ SimCSE	55.14	43.02	78.09	69.84
SEG-QU	64.28	51.21	80.59	74.95

Table 7: Click-stratified results on the two QU tasks.

thresholds. For comparability, all reported performance results use equally sized 100M-pair training corpora. We selected $\delta = 0.5$ based on average dev macro- F_1 . At $\delta = 0.7$, too few pairs were retained to construct an equally sized corpus.

C Click-Stratified Evaluation

Our claim concerns the availability of pre-training supervision, rather than larger gains for click-poor queries at inference. We nevertheless examine whether the gains persist across click-frequency strata.

We divide each NER and classification test set into equal-sized click-rich and click-poor halves based on historical click counts from the same log period used for pre-training. We do not use exact zero-click subsets because they are too small and have highly skewed label distributions. We retain the original label sets and report macro- F_1 .

Table 7 compares SEG-QU with representative baselines. Its gains remain consistent for click-poor queries, although they are not uniformly larger than those for click-rich queries. Absolute performance is lower on the click-poor subsets for all models, suggesting that these queries form a more challenging subset, potentially because they are rarer or noisier.

D Why Are Session Pairs Effective for BERT-Style SSQP?

Although NSP has shown limited utility in general-domain pre-training (Liu et al., 2019; Lan et al.,

Source	Pos. med.	Neg. med.	Med. Δ	Mean Δ
Session	0.9382	0.7740	0.1614	0.1544
Wikipedia	0.8506	0.7537	0.0969	0.0952

Table 8: Cosine similarities and paired positive-negative margins by pair source.

2020), our BERT-style SSQP objective improves downstream QU performance. We hypothesize that this difference stems from the stronger semantic association of NPMI-distilled session pairs.

To examine this hypothesis, we sampled 2,000 NPMI-distilled session pairs and 2,000 adjacent sentence pairs from the June 2026 Japanese Wikipedia dump. For each positive pair, we constructed a random negative with the same anchor. We encoded the texts using the Japanese text embedding model `cl-nagoya/ruri-v3-310m`¹¹ and computed pairwise cosine similarities. For each anchor, we define the margin Δ as the positive similarity minus the negative similarity.

As shown in Table 8, the session pairs have substantially larger positive-negative margins than adjacent Wikipedia sentences. In contrast, adjacent sentences are less clearly separated from random pairs, potentially making the discrimination learned through general-domain NSP less reliable. This contrast supports our hypothesis that the effectiveness of SSQP stems from the more discriminative supervision provided by NPMI-distilled session pairs.

E Configuration for § 6.3

This appendix details the fine-tuning and inference settings used in the comparison of § 6.3. The corresponding per-task results are reported in Table 9.

LLMs We fine-tune each instruction-tuned Sarashina2.2 model (0.5B, 1B, and 3B) on the same training splits as the encoders (§4.3). We use LoRA with rank 16 and $\alpha = 32$, a learning rate of $2e-4$, a batch size of 64, and 10 epochs. For inference, each model is cast to `bfloat16`, with LoRA adapters merged, and served with `vLLM`¹²; 8-bit quantization increased latency in our setup, consistent with Dettmers et al. (2022). We use greedy decoding with a maximum of 128 new tokens. We frame both tasks as instruction-following generation, with the prompts shown in Figure 4. For NER,

¹¹<https://huggingface.co/cl-nagoya/ruri-v3-310m>

¹²<https://github.com/vllm-project/vllm>

Setting	Macro-F ₁ ↑			p95 Latency (ms) ↓		Energy (J/q) ↓		Parse Fail (%) ↓	
	NER	Class.	Avg.	NER	Class.	NER	Class.	NER	Class.
<i>LLMs (generative)</i>									
Sarashina2.2 _{0.5B}	43.29	69.70	56.50	51.98	35.04	15.45	9.68	0.096	0.000
Sarashina2.2 _{1B}	48.12	71.00	59.56	69.79	50.13	19.44	18.71	0.048	0.000
Sarashina2.2 _{3B}	46.90	76.73	61.82	133.14	91.38	38.76	27.75	0.000	0.000
<i>Encoder Models</i>									
ModernBERT _{130M}	41.99	65.04	53.52	1.71	1.65	0.43	0.37	–	–
SEG-QU _{70M}	50.54	76.00	63.27	1.39	1.26	0.34	0.30	–	–
SEG-QU _{130M}	56.99	78.48	67.74	1.71	1.65	0.43	0.37	–	–
SEG-QU _{310M}	60.32	79.82	70.07	2.20	2.06	0.65	0.57	–	–

Table 9: Per-task performance, serving cost, and parse failure rate for the LLM comparison (§ 6.3). Parse failure applies only to the generative LLMs. Encoder latency is nearly constant across tasks, while the LLM latency gap widens on NER, where longer generation dominates.

Query NER Prompt
Extract named entities from the following text. You may use only the labels in the following list: [{labels}]. Output only the full original text as a string, replacing each named entity span with the form [text LABEL]. Do not alter any characters, whitespace, or symbols outside the brackets. If no entity is found, output the original text unchanged. No explanation or supplementary text is needed. Text: {text}
Query Classification Prompt
Read the following text and select all applicable labels from the following list: [{labels}]. Output only a JSON array listing the label name strings. If no label applies, output []. No explanation or supplementary text is needed. Text: {text}

Figure 4: Prompts used for the LLM baselines, shown as English translations of the original Japanese prompts.

we parse the bracketed entities from each output; for classification, we parse the JSON array of labels. We compute macro-F₁ from the parsed predictions, counting unparseable outputs as no prediction.

Encoder Models At inference, each model is cast to bfloat16, truncated to a maximum length of 32 tokens, and served with HuggingFace Transformers.¹³

F Click-Based Pre-training Details

We use click-through logs from the same period as the session logs (§4.1) and extract (query, title) pairs in which the query led to a click on the page. To suppress navigational queries, we retain only

titles linked to between 2 and 5 distinct queries, and from each retained title, sample two queries q_1 and q_2 to construct (q_1, title) for clicked title MLM and (q_1, q_2) for co-click query prediction, yielding 10M (q_1, q_2, title) tuples in total. Training runs for 1 epoch with learning rate 5e-5 and batch size 2,048. For the SEG-QU → Session control, we continue training on session data under the same configuration, adjusting the number of steps to match the expected number of MLM training tokens in SEG-QU → Click.

G ESCI Details

We use the Japanese subset of the reduced ESCI dataset. We split the official training set 9:1 by query ID into 6,556 training and 728 development queries. All query-product pairs associated with the same query are assigned to the same split. The resulting training, development, and test sets contain 188,541, 20,553, and 88,789 query-product pairs, respectively. We fine-tune each model for three epochs with a batch size of 64 and a learning rate of 5e-5.

¹³<https://github.com/huggingface/transformers>