


# AGNUS: Robust Entity Disambiguation using decoder-only LMs

Anonymous ACL submission

## Abstract

Entity disambiguation (ED) links ambiguous mentions in text to entries in a knowledge base and is a core task in entity linking systems. While pretrained decoder-only language models (DLMs) offer strong generalization capabilities, their effective use in ED has been restricted due to sensitivity to candidate order, susceptibility to hallucinated outputs, and potential dataset leakage. We introduce AGNUS  a zero-shot ED framework that addresses these challenges through three core innovations: (1) **order-invariant candidate encoding** via shared positional embeddings and modified autoregressive attention masking, which eliminates bias on input ordering; (2) **constrained decoding** that ensures outputs are restricted to valid candidates, effectively preventing hallucinations; and (3) **synthetic dataset** creation approach as a diagnostic tool for data contamination detection and counteraction. AGNUS eliminates up to 15.2% of F1 variability caused by candidate permutations, delivering consistent and order-robust predictions previously unattainable with autoregressive architectures. In our experiments, AGNUS achieves state-of-the-art performance on four standard ED benchmarks, surpassing prior zero-shot approaches by an average 3.7%. We release code, data including candidate sets, and a synthetic benchmark to support reproducibility and controlled evaluation<sup>1</sup>.


## 1 Introduction

Entity Disambiguation (ED) represents the task of linking ambiguous mentions in text to the correct entity from a provided candidate set and is a core component of knowledge-intensive NLP applications such as question answering, semantic search, and entity linking. While Large Language Models (LLMs) have demonstrated remarkable generalization across diverse tasks, their robust application to

ED remains challenging, particularly in zero-shot settings.

In this work, we identify and address three fundamental limitations of Decoder-Only Language Models (DLMs) when applied to ED:

First, these models are highly sensitive to the order in which candidate entities are presented. Autoregressive generation induces positional bias, leading to substantial prediction variability across permutations of candidate inputs - up to 15.2% (F1) in our experiments. Second, DLMs may produce entities that are not part of the candidate set, undermining system reliability in constrained ED settings. Third, meaningfully evaluating LLMs on existing benchmarks is complicated by potential training data contamination, given the opacity of pretraining corpora.

To overcome these challenges, we propose AGNUS , a robust, zero-shot ED framework for DLMs that can be applied to open-weight models out-of-the-box, requiring no fine-tuning or re-training. AGNUS incorporates two key components to achieve order-robustness and hallucination resistance: (1) **Masked Attention Candidate Set (MACS)** to enforce candidate order-invariant encoding; and (2) **Agnus Contextual Decoding (ACDC)** to restrict decoding to valid candidate entities.


Our approach ensures that candidate entities are represented indistinguishably in terms of position (positional embeddings) and interdependencies (attention) to the underlying model, eliminating the influence of input order on token predictions. At the same time, constrained decoding removes hallucinated outputs without sacrificing the model's contextual reasoning ability.


To audit dataset leakage, we design a synthetic dataset construction methodology for ED. The resulting synthetic dataset serves as a diagnostic tool to detect contamination, evaluate with a potentially lesser degree of contamination and therewith al-


<sup>1</sup><https://anonymous.4open.science/r/Agnus/README.md>

lows us to test model generalization capabilities in a controlled setting.

Our contributions are as follows:

- We propose AGNUS , a zero-shot ED framework combining:
  - **MACS** for order-invariant candidate encoding.
  - **ACDC** for constrained autoregressive tree-based decoding.
- We introduce a synthetic dataset construction method to gauge benchmark contamination and apply it to the AIDA (Yosef et al., 2011) benchmark.
- We release all code, data including entity candidates and evaluations to support reproducibility<sup>2</sup> and future comparability.

Across four standard ED benchmarks, AGNUS  achieves state-of-the-art performance in zero-shot settings, while delivering stable predictions under candidate permutations and eliminating hallucinated outputs.

The remainder of this paper is structured as follows. Section 2 reviews related work in entity disambiguation and recent advances in large language models, with particular attention to dataset contamination, hallucination, and order sensitivity. Section 3 introduces our proposed framework, AGNUS , detailing the disambiguation setup, our order-invariant encoding method, and constrained decoding strategy. Section 4 presents our experimental setup, results across standard benchmarks, and a comprehensive ablation analysis, followed by a study on contamination detection. Finally, Section 5 concludes the paper and outlines directions for future work.

## 2 Related Work

### 2.1 Entity Disambiguation

Entity disambiguation (ED) is a critical task in natural language processing and understanding, where the goal is to map ambiguous entity mentions in text to their correct entries in a knowledge base. Current state-of-the-art ED and entity linking models (van Hulst et al., 2020; Barba et al., 2022; Ayoola et al., 2022; Shavarani and Sarkar,

2023; Xiao et al., 2023b; Ding et al., 2024a; Orlando et al., 2024) make use of various deep learning architectures to outperform more traditional works. In recent years, transformer-based systems, such as BLINK (Wu et al., 2020), REL (van Hulst et al., 2020), SpEL (Shavarani and Sarkar, 2023), DeepType (Raiman and Raiman, 2018) and GENRE (Cao et al., 2021) have taken over the stage with many basing themselves on BERT (Devlin et al., 2019) embeddings. In recent years, LLM-based systems have entered the space with (Sun et al., 2023), (Wang et al., 2023a), (Xiao et al., 2023a), EntGPT (Ding et al., 2024a), ChatEL (Ding et al., 2024b), LLMAEL (Xin et al., 2024) and (Tasawong et al., 2024). Particularly, in (Ding et al., 2024a; Xin et al., 2024; Liu et al., 2024; Vollmers et al., 2025) authors improve LLM-based entity disambiguation by tuning inputs and otherwise providing LLM backbones with context-relevant data.

### 2.2 Large Language Models

Applying LLMs to ED is accompanied by a multitude of considerations when contrasted with more traditional ED. Among these, there exist benchmark contamination (Section 2.2.1), hallucinations (Section 2.2.2), decoding mechanisms (Section 2.2.3) and order-specific biases (Section 2.2.4) that endanger robust disambiguation. In the following, we address these areas of prior work.

#### 2.2.1 Dataset Contamination

Benchmark contamination in LLMs (Xu et al., 2024) has become a critical issue as models trained on vast amounts of publicly available data may inadvertently ‘memorize’ aspects of popular benchmark datasets, potentially leading to inflated estimates of their true capabilities.

To address these challenges, researchers have started developing various countermeasures (Chen et al., 2025), including dynamic evaluation benchmarks (Wang et al., 2025; Zhu et al., 2024a,b) to effectively prevent pre-benchmarking disclosure. Another measure is to provide a means of evaluation for the degree of contamination (Xu et al., 2024) by computing perplexity (Li, 2023) – by applying the exponential function to the average negative log likelihood over a particular sequence of text to measure a model’s ‘surprise’ (or inverse confidence) for a particular output.

<sup>2</sup><https://anonymous.4open.science/r/Agнус/README.md>

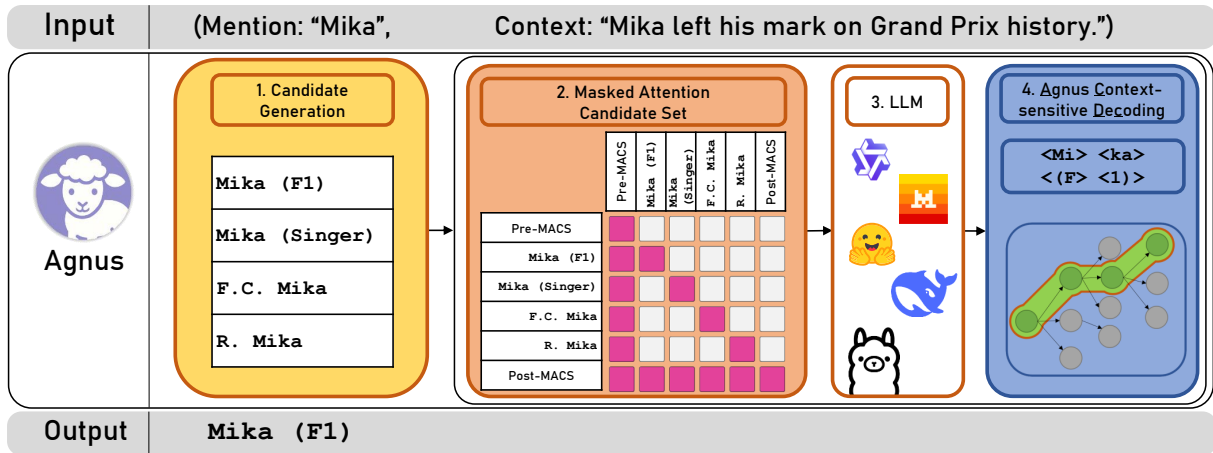


Figure 1: AGNUS Overview – Takes an input document, (1.) generates candidate entities for mentions (e.g. MIKA) using a pre-existing candidate generation method, (2.) applies masked attention and altered positional embeddings to the candidate entity collection (MACS, Section 3.2) and (3.) passes representation to a specified LLM, followed by (4.) constrained decoding (ACDC, Section 3.3) for context-sensitive disambiguation and returns the disambiguated entity (e.g. Mika (F1)).

### 2.2.2 Hallucinations

Despite remarkable capabilities in generating human-like text, LLMs may produce factual inaccuracies or nonsensical sequences, a phenomenon referred to as *hallucination* (Huang et al., 2025). The underlying causes of hallucinations are an active area of research. Some potential contributing factors include the vast scale of the training data, containing potentially noisy data (Petroni et al., 2021; Ji et al., 2023) and the autoregressive nature of text generation based on prior tokens (Holtzman et al., 2020; Maynez et al., 2020). The presence of hallucinations poses a significant challenge for the reliable application of LLMs on downstream Natural Language Processing (NLP) tasks, posing issue for robust and trustworthy ED. Recent research efforts have started counteracting hallucinations through retrieval augmentation, fact verification and the incorporation of knowledge graphs (Lewis et al., 2020; Pusch and Conrad, 2024).

In this paper, we eliminate the possibility for entity candidate hallucinations by defining a specialised constrained decoding strategy for ED.

### 2.2.3 Constrained Decoding

Early work on LLMs (Brown et al., 2020; Radford et al., 2019) demonstrated that decoder-only language models process natural language prompts effectively without an enforced schema, meaning that input-output pairs are structurally not bound by predefined templates or grammars. This flexibility allows for broad applicability but introduces challenges in reliability, consistency, and controlla-

bility (Bender et al., 2021).


To mitigate challenges of unstructured interaction, researchers have developed various prompt engineering methods (Sahoo et al., 2024; Ouyang et al., 2022a; Madaan et al., 2023; Wei et al., 2022) to implicitly guide, but not force DLMs towards more structured outputs. Therefore, constrained decoding (Beurer-Kellner et al., 2024) approaches to enforce strict restrictions on LLM text generation have been developed and even started being applied to the domain of entity linking (Vollmers et al., 2025).

### 2.2.4 Order Bias

Prior work has established that modern generative large language models demonstrate inherent tendencies toward positional preferences when processing ordered lists of candidate answers (Pezeshkpour and Hruschka, 2023; Wei et al., 2024; Zheng et al., 2023; Anonymous, 2025) and that these are also sensitive towards the arrangement order of otherwise identical answer collections (Dominguez-Olmedo et al., 2023; Li et al., 2023; Li and Gao, 2024; Wang et al., 2023b, 2024a; Xue et al., 2024). Approaches to mitigation include compensation for positional preferences (Wei et al., 2024; Zhao et al., 2021), systematic permutation averaging and applying multiple forward passes with varied option sequences (Pezeshkpour and Hruschka, 2023; Wang et al., 2023b), as well as reasoning-enhanced strategies (Wang et al., 2024a,b) to attenuate sequence dependence. AGNUS employs a method to mitigate candidate or-

der bias without requiring additional training by adapting the approach from (Anonymous, 2025) – which investigated the effects of altered positional embeddings on generated output – to entity disambiguation.

### 3 AGNUS

In this section we introduce AGNUS , our proposed approach for LLM-based robust entity disambiguation. In Figure 1, we present AGNUS: from (Step 1.) generating entity candidates for each mention using the DBpedia Lookup<sup>3</sup> service – chosen for improved reproducibility; (Step 2.) applying combined masked attention and position-specific shared positional embeddings (*Masked Attention Candidate Set*, Section 3.2) based on (Anonymous, 2025); (Step 3.) passing the encoded inputs to a chosen model; (Step 4.) constrained decoding (ACDC) to final disambiguation for the input document "*Mika left his mark on Grand Prix history.*" and entity mention *Mika*, yielding contextually disambiguated entity *Mika* (F1)<sup>4</sup>.

#### 3.1 Disambiguation Setup

AGNUS represents an approach leveraging DLMs for the task of disambiguating entities based on entity candidate information while mitigating DLM-specific challenges. For disambiguation, AGNUS takes as input a document providing context, a mention and a collection of candidate entities generated via pre-existing candidate generation approaches.

Due to leveraging the contextual disambiguation capabilities of DLMs, AGNUS does not require candidate entities to solely be a knowledge base-backed IRI<sup>5</sup>. Instead, candidate entity representation may additionally take any identifying or meaningful form, such as a description, label, type or combination thereof. For each mention contained within an input document, we generate a fixed candidate set, employing candidates generated with DBpedia Lookup<sup>6</sup>. Each candidate collection is encoded using MACS (Section 3.2). Subsequently, the resulting encoded prompt is transmitted as a whole to the underlying DLM for contextual parsing and decoded via ACDC (Section 3.3).

<sup>3</sup><https://lookup.dbpedia.org/>

<sup>4</sup>[https://en.wikipedia.org/wiki/Mika\\_Häkkinen](https://en.wikipedia.org/wiki/Mika_Häkkinen)

<sup>5</sup><https://wikipedia.org/wiki/>

Internationalized\_Resource\_Identifier

<sup>6</sup><https://github.com/dbpedia/dbpedia-lookup>

#### 3.2 Masked Attention Candidate Set

Text sequences encoded on modern generative language models rely on underlying position-influenced attention mechanisms and positional embeddings to signal the order of token appearance within a sequence (Anonymous, 2025). This affects desired order-invariant sequences, such as candidate collections – an undesirable property for entity disambiguation. To render an LLM order-agnostic for parts of a sequence, we tackle both aspects: modify positional embeddings (Section 3.2.1) for candidate entities to simulate similar positions and mask the attention mechanism between entity candidates (Section 3.2.2) to the underlying language model.

Each candidate collection is encoded using MACS, embedded into its original textual encoding with text preceding (Pre-MACS) and succeeding (Post-MACS) the collection being encoded in standard LLM-specific fashion.

##### 3.2.1 Positional Embedding

Every sequence of tokens is attributed a certain range of positional embedding values within its LLM-encoded representation. Within a MACS-encoded collection, every token making up an entity candidate is modified to appear as sharing a similar range of positions (see visualization Fig. 2) as other candidates to the underlying LLM.

To do so, we define relative position  $i \in [0, \dots, n_{c_j} - 1]$  of each token  $t_{c_j,i}$  for entity candidate representation  $c_j \in C$  s.t.  $n_{c_j}$  is the number of tokens for entity  $c_j$  and collection of all candidate entities  $C$  for a given mention and  $T_{c_j}$  the set of all tokens for  $c_j$ :  $\forall t_{c_j,i} : i \in [0, \dots, \max_{c \in C} (|n_c - 1|)]$ .

Therefore as visualised in Figure 2, the shared range of possible positional embeddings is defined by the token-wise longest candidate within a MACS collection and starts for each candidate at the end of prior sequence’s token (PRE-MACS) and afterwards continues the candidate encoding with the succeeding sequence’s (POST-MACS) first positional embedding.

##### 3.2.2 Causal Mask

To encode a collection of entity candidates in an order-invariant fashion to the underlying LLM, we apply an adapted version of the commonly-employed triangular attention matrix as causal mask (see Figure 3). Entities within a collection cannot attend to one another (grey entries), but do attend (pink entries) – and are attended to – in oth-



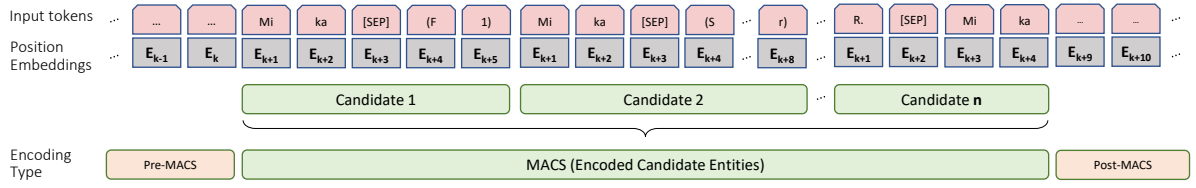


Figure 2: MACS – Positional Embeddings: Each candidate entity entry is encoded as being on the same range of positions for the length of their contents. Candidate entity entries’ positional identifiers are **shared** across common lengths and encoded analogously. Post-MACS – any tokens after a MACS block – starting positional embedding is computed as being subsequent to the longest option contained within MACS (candidate) entries.

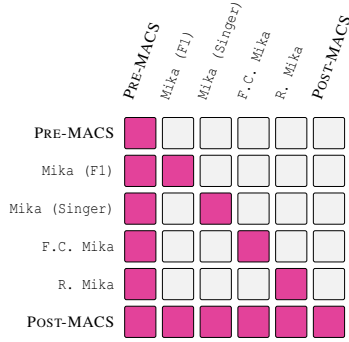


Figure 3: MACS – Causal mask: Example from Fig. 1 for entity candidate representations for entities "Mika (F1)", "Mika (Singer)", "F.C. Mika", "R. Mika". Grey cells signify blocked attention whereas pink signifies enabled attention. Intra entity attention and attention from tokens preceding (Pre-MACS) and succeeding (Post-MACS) MACS is preserved normally s.t. subsequent tokens attend to prior ones.

erwise usual LLM fashion to their own prior tokens (diagonal entries) and rest of the token sequence (to PRE-MACS and by POST-MACS). This means that tokens within each candidate’s representation continue attending to each other, such as <ka> to <Mi> within our example.

### 3.3 Agnus Contextual Decoding


Generative LLMs may hallucinate information in unexpected fashions. This ranges from a corrupt expected result format to non-existing options. Due to the nature of entity disambiguation, only given options may be produced. As such, we define an input-flexible grammar based on entity candidates. We implement this grammar in a logits processor<sup>7</sup> that filters forbidden tokens at each generation step until a single disambiguated entity remains.

Formally: Let the set of candidate sequences be  $O = \{o_1, \dots, o_n\}$  where each candidate option  $o_i \in \Sigma^{l_i}$  is a sequence of length  $l_i$ . The vocabulary is de-

<sup>7</sup><https://anonymous.4open.science/r/Agnus/src/agnus/pipeline/llm.py#L356>

defined as  $\Sigma = \{t_k^i \mid i \in \{1, \dots, n\}, k \in \{1, \dots, l_i\}\} \cup \{\text{EOS}\}$ . We then define the set of nonterminals as  $V = \{X_i^k \mid i \in \{1, \dots, n\}, k \in \{0, \dots, l_i\}\}$  where  $X_i^k$  denotes the state after generating the first  $k$  tokens of candidate  $c_i$ . The start symbol transitions to the initial state of each candidate:  $S \rightarrow X_1^0 \mid X_2^0 \mid \dots \mid X_n^0$ . For each  $o_i$ , we define the following transitions:  $X_i^k \rightarrow t_{k+1}^i X_i^{k+1}, \forall k \in \{0, \dots, l_i - 1\}, X_i^{l_i} \rightarrow \text{EOS}$ .

## 4 Experiments and Results

AGNUS  combines techniques to create an LLM-enabled approach to robust entity candidate disambiguation. In this section, we conduct experiments to evaluate AGNUS with different configurations regarding representations of entity candidates, LLMs, our candidate encoding (MACS) and our constrained decoding (ACDC). We report entity disambiguation results in comparison to prior work in Table 1.

### 4.1 Technical Details

All our experiments were run on a server with NVIDIA RTX 4090 (24GB vRAM), 1TB RAM, 128 CPU cores, Debian (Bookworm), CUDA 12.5 and Python 3.11. As for LLMs, we decided on instruct models for our experiments such that they would run on our hardware and be comparable in size, leading to the following selection: Mistral (7B-Instruct) (Jiang et al., 2023), Llama2 (7B) (Touvron et al., 2023), Llama3 (8B-Instruct) (Dubey et al., 2024) and Qwen (2.5-7B-Instruct) (Yang et al., 2024) – for the rest of the paper we omit detailed version specifications.

### 4.2 Evaluation

In our experiments, we outperform related work on 4 out of 5 common datasets (AIDA (Yosef et al., 2011), KORE 50 (Hoffart et al., 2012), MSNBC (Cucerzan, 2007), ACE04 (Ratinov et al., 2011), AQUAINT (Milne and Witten, 2008)) in

zero-shot settings despite our underlying LLMs’ relatively modest parameter count<sup>8</sup>.

We report our ED F1 results in Table 1 – the *top block* lists scores for trained or finetuned approaches, the *bottom block* compares zero-shot methods. Our model performs strongly across most datasets and even surpasses finetuned or trained prior work in certain cases. Despite being a zero-shot approach, AGNUS attains overall new state-of-the-art results for KORE 50 (82.3%) and ACE04 (95.5%). Unsurprisingly when evaluating on AIDA, approaches trained on AIDA outperform ours, but AGNUS (86.7%) exceeds second-ranked zero-shot approach EntGPT-P (Ding et al., 2024a) (82.1%) F1 measure by 4.6%. Evaluating against KORE 50, AGNUS reaches 82.3% in comparison to ChatEL’s 78.7%, surpassing it by 3.6%. As for ACE and AQUAINT, our results (95.5% and 87.5%) improve upon EntGPT-P’s (91.8% and 79.1%) respectively by 3.7% and 8.4%. For MSNBC, we do not beat the state-of-the-art for zero-shot entity disambiguation and instead reach 82.4%, underperforming ChatEL (Ding et al., 2024b) (88.1%) by 5.7% and finetuned state-of-the-art CoherentED (96.3%) by 13.9%. While AGNUS yields improvements across some benchmarks, we consider our primary benefit lying in enhancing disambiguation robustness via order invariance for candidates and by preventing structurally invalid outputs.

### 4.3 Ablation Study

AGNUS employs multiple techniques to mitigate issues relating to LLM-based ED. Particularly, AGNUS relies on LLMs for disambiguation, MACS for order-invariant candidate encoding, ACDC for entity decoding and particularly candidates’ representation. In our ablation study, we therefore design experiments to verify the impact of these aspects on model results by investigating candidate representation (Section 4.3.1), LLM selection (Section 4.3.2), MACS (Section 4.3.3) and ACDC (Section 4.3.4).

<sup>8</sup>We note that our models are at least an order of magnitude smaller and that Ding et al. (2024b) argue model parameter count having a significant influence on the entity disambiguation task. EntGPT (Ding et al., 2024a) and ChatEL (Ding et al., 2024b) employ Llama2 70B (Touvron et al., 2023) and GPT-3.5 (Ouyang et al., 2022b); ChatEL (Ding et al., 2024b) additionally makes use of PaLM 540B (Chowdhery et al., 2023) and GPT-4 (OpenAI, 2023). OpenAI has not disclosed parameter counts for GPT-3.5 and GPT-4, but each of them is assumed to have at least 175B parameters, with rumors claiming GPT-4 having 1.76 trillion parameters according to <https://en.wikipedia.org/wiki/GPT-4>.

#### 4.3.1 Candidate Representation

To validate LLM disambiguation capabilities based on contextual candidate entity information, we apply AGNUS to candidate representations of different entity information types. We selected DBpedia (& Wikipedia) entity IRIs, entity types, textual entity descriptions and labels as meaningful entity information characterising entity candidates for our experiments (Tables 2 and 3). We note that in Table 2 across all datasets, IRI-based representations perform best with an average F1 performance of 86.9%, outperforming labels by 10.2% – with a tie of 87.9% for AQUAINT. For all datasets beside KORE 50 and AQUAINT, descriptions reach the second-highest score (avg.: 73.5%), but are still surpassed by labels (76.7%) on average by 3.2%. We note that the shorter and more unique a representation is, the better AGNUS seems to perform. In our experiments, we find effects of representation depend on benchmarked dataset with representation-based score differences ranging from 5.6% (ACE04) to 34.9% (AQUAINT) with a mean of 21.12% across our 6 datasets.

#### 4.3.2 Large Language Model

To verify our approach’s generalizability across LLMs, we run AGNUS on 4 LLMs: Llama2 (Touvron et al., 2023), Llama3 (Dubey et al., 2024), Qwen (Yang et al., 2024) and Mistral (Jiang et al., 2023). In Table 3, we notice similar trends across most LLMs for the AIDA dataset with Llama2 representing a slight outlier: All other LLMs attain respective top results using IRIs (Qwen: 84.6%, Mistral: 86.7%, Llama3: 84.0%) as candidate information, whereas our outlier LLM manages to slightly improve on its 80.9% F1 measure, reaching 81.3% by employing labels as candidate representation. Typically, Llama3, Mistral and Qwen reach similar results to each other using IRIs (84.0% – 86.7%) and descriptions (70.2% – 76.3%) as candidate representations. Using labels, Qwen plummets down to 64.6%, whereas Llama3 (74.5%) and Mistral (74.3%) attain F1 scores close to each other. For type candidate information, Mistral (70.5%) noticeably outperforms Qwen (42.2%) and Llama3 (39.2%); Llama2 manages to outperform its successor Llama3 (56.0%). Llama3 (70.2%) and Qwen (72.1%) handle descriptions as meaningful entity information comparably well with Mistral (76.3%) performing slightly better and Llama2 (56.5%) displaying worst results.

Trained (or finetuned) for ED						
Model	AIDA	KORE 50	MSNBC	ACE04	AQUAINT	Mean
End2End (Kolitsas et al., 2018)	0.891	0.569	0.933	0.892	0.894	0.836
GENRE (Cao et al., 2021)	0.933	0.542	0.943	0.901	0.899	0.844
REL (van Hulst et al., 2020)	0.928	<u>0.618</u>	0.935	0.897	0.873	0.850
ReFinED (Ayoola et al., 2022)	<u>0.939</u>	0.567	0.941	0.908	0.918	0.855
LLMAEL × ReFinED <sub>FT</sub> (Xin et al., 2024)	0.923	-	0.888	0.881	0.891	0.900
EntGPT-I (GPT3.5) (Ding et al., 2024a)	0.920	<b>0.753</b>	0.922	<b>0.937</b>	0.906	0.888
EOEDbMSL (Tasawong et al., 2024)	<b>0.941</b>	-	0.935	0.917	0.894	0.922
ExtEnD (Barba et al., 2022)	0.926	-	<u>0.947</u>	0.918	<u>0.916</u>	<u>0.927</u>
CoherentED (Xiao et al., 2023b)	0.894	-	<b>0.963</b>	<u>0.934</u>	<b>0.946</b>	<b>0.934</b>

LLM 0-shot ED						
Model	AIDA	KORE 50	MSNBC	ACE04	AQUAINT	Mean
ChatEL (Ding et al., 2024b)	-	<u>0.787</u>	<b>0.881</b>	0.893	0.767	<u>0.832</u>
EntGPT-P (GPT3.5) (Ding et al., 2024a)	<u>0.821</u>	0.716	<u>0.867</u>	<u>0.918</u>	<u>0.791</u>	0.823
EntGPT-P (Llama2 70B) (Ding et al., 2024a)	0.708	0.647	0.741	0.746	0.635	0.695
Ours – AGNUS (Llama2 8B)	0.809	0.529	0.562	0.897	0.576	0.675
Ours – AGNUS (Mistral)	<b>0.867</b>	<b>0.823</b>	0.824	<b>0.955</b>	<b>0.875</b>	<b>0.869</b>
Baseline: Mistral (hidden candidates)	0.791	0.794	0.739	0.953	0.720	0.799
Ablation: w.o. MACS (best)	0.865	0.811	0.814	0.962	0.907	(0.872)
Ablation: w.o. MACS (worst)	0.833	0.779	0.766	0.950	0.847	(0.835)

Table 1: ED evaluation table – **Upper category**: ED systems trained or finetuned for ED (mainly with AIDA). **Lower category**: 0-shot ED systems. Top scores per column and category **bolded**, second highest underlined. Scores obtained from respective papers. Note that baseline with *hidden candidates* also uses matching to candidates (else naive results would tend to 0) and MACS ablations are run over multiple iterations, showing score variability.

Entity Representation	AIDA	AIDA-Syn	KORE 50	MSNBC	ACE04	AQUAINT	Mean
AGNUS (w. IRI)	<b>0.867</b>	<b>0.863</b>	<b>0.823</b>	<b>0.824</b>	<b>0.955</b>	<b>0.879</b>	0.869
AGNUS (w. Label)	0.743	0.706	<u>0.785</u>	0.589	0.899	<b>0.879</b>	0.767
AGNUS (w. Type)	0.705	0.719	0.595	0.591	0.934	0.530	0.679
AGNUS (w. Description)	<u>0.763</u>	<u>0.790</u>	0.515	<u>0.679</u>	<u>0.954</u>	<u>0.706</u>	0.735
Mean	0.769	0.770	0.679	0.671	0.936	0.748	0.762

Table 2: Ablation Study (Candidate Representation over datasets): AGNUS (Mistral) F1 measures on AIDA, AIDA-Syn, KORE 50, MSNBC, ACE04 and AQUAINT with different candidate entity representations (IRI, label, entity type, entity description), along with per representation and per dataset averages. Top entry by dataset in **bold**, second underlined.

Model	AIDA				Mean
	IRI	Label	Type	Description	
AGNUS (Qwen)	0.846	0.646	0.422	0.721	0.659
AGNUS (Mistral)	<b>0.867</b>	0.743	0.705	0.763	0.770
AGNUS (Llama2)	0.809	0.813	0.560	0.565	0.687
AGNUS (Llama3)	0.840	0.745	0.392	0.702	0.670
AGNUS (Llama3) w.o. ACDC	0.765	0.698	0.331	0.677	0.618

Table 3: Ablation Study (LLM, Candidate Representation, ACDC): AGNUS F1 measures for different types of candidate representations for Qwen, Mistral, Llama2, Llama3 and without constrained decoding via ACDC. AGNUS without ACDC utilises fuzzy search, ranking reply and candidate, matching to candidate with highest similarity.

### 4.3.3 Masked Attention Candidate Set

We investigate how MACS affects qualitative results and whether it actually renders disambiguation order-invariant. To this end, we run experiments shuffling candidates over 10 iterations and

display results in Figure 4. Our experiments over 3 different LLMs (Llama3, Mistral, Qwen) display how disambiguation varies without the use of MACS and remains unchanged when applying MACS. Order invariance persists across all 10 iterations of shuffled candidates when MACS is employed whereas not applying the causal mask to candidate entities yields result variations. Without MACS, Llama3 averages at 66.53% (MACS: 66.40%) and varies between 59.56% – 73.02%, a difference of 13.46%. Mistral on the other hand varies in the range of 32.47% – 43.69%, averaging at 38.07% without MACS across iterations of candidate shuffles (with MACS: 38.20%). In Table 1, we also display MACS ablations over 2 iterations, from one non-MACS execution to another exhibiting 3.7% F1 difference on average and beating AGNUS in ACE04 (96.2% vs. 95.5%). Finally, Qwen

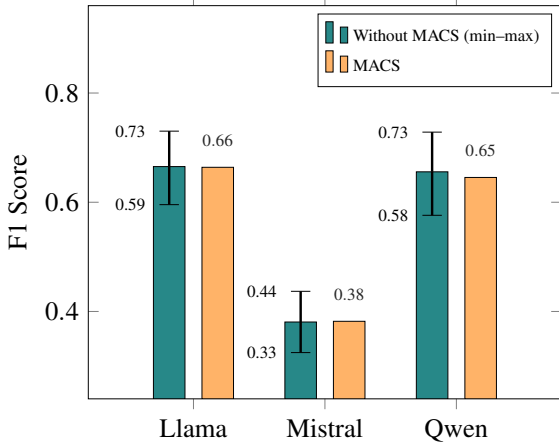


Figure 4: Ablation Study (MACS) - F1 Score Variability: Error Bar plot on disambiguation without (*left*) and with MACS (*right*) with randomised candidate shuffles over 10 iterations with Llama3, Mistral and Qwen on perplexity decoding – disambiguates to highest confidence – for AIDA. MACS and non-MACS results are similar on average. Without MACS, performance varies (Llama3: 13.5%, Mistral: 11.2%, Qwen: 15.2%).

also exhibits changes resulting from candidate order changes: with an average of 65.57% (MACS: 64.54%) its candidate order-dependant results vary within the range 57.61% – 72.84%. Based on our experiments, we conclude that MACS effectively removes order-based bias from candidates with an overall minor average reduction in F1 score.

#### 4.3.4 Agnus Contextual Decoding

In Table 3, additionally to checking out the impact of candidate representations across language models, we also evaluate AGNUS without our constrained decoding method (ACDC). We find that AGNUS hallucinates across the board, decreasing F1 for all types of candidate representation. In non-ACDC experiments, we apply fuzzy matching to improve the likelihood of finding at least one entity. Exact disambiguation to candidate matches in our zero-shot experiments yield extremely subpar results (close to 0) and would otherwise be misrepresenting the added value of our robustness-oriented approach. On average, F1 performance without ACDC is lowered by 5.2%, the largest drops appearing with IRI (-7.5%) and type (-6.1%) candidate representations, followed by label (-4.7%) and descriptions (-2.5%).

#### 4.4 Contamination Detection

To diagnose potential contamination, we employ perplexity (Li, 2023) to quantify a model’s uncer-

tainty for a given token sequence prediction. Perplexity reflects the inverse likelihood assigned to a particular token sequence by a model: lower perplexity indicates higher predictive confidence and a higher likelihood of contamination. To detect contamination and evaluate the generalizability of DLMs, we propose synthetically generating a novel dataset derived from an existing one by replacing each entity mention with a distinct, contextually similar mention and corresponding entity. We apply our method with the DeepSeek-R1 (DeepSeek-AI et al., 2024) model<sup>9</sup> to AIDA (Yosef et al., 2011) and release AIDA-Syn<sup>10</sup>. For each sequence, we produced five mention-entity options, but for AIDA-Syn only one was retained to reduce the risk of future pretraining exposure. All alternatives, along with a generation script, are made available<sup>11</sup>. To assess contamination levels across different LLMs, we introduce a modified decoding strategy, illustrated in Figure 4 with disambiguation performed by selecting a candidate entity with highest confidence. The model that performs worst with this strategy is presumed to be least contaminated. Our findings show that Mistral (Jiang et al., 2023) yields the lowest performance with a perplexity-based decoding method on AIDA, suggesting being least affected by benchmark contamination. Applying the same decoding strategy with AIDA-Syn, F1 score decreases from 38.20% to 22.82%, a substantial relative drop of 15.38%. This reduction supports the hypothesis that AIDA-Syn exhibits reduced contamination and that our underlying DLMs may suffer from contamination.

## 5 Conclusion

We propose a set of techniques to enable robust LLM-based entity disambiguation by addressing the issues of unwanted order bias and hallucinations. Our experimental results show that our zero-shot approach outperforms prior work on average by 3.5%. Further, we introduce a methodology to detect data contamination and publish a novel diagnostic dataset AIDA-Syn. While yields modest improvements across benchmarks, our primary benefit lies in enhancing output robustness and controlling generation behavior, particularly in cases where unconstrained to semantically or structurally invalid outputs.

<sup>9</sup>Version from May 2025: <https://www.deepseek.com/>

<sup>10</sup>Made up of 888 documents. More details in Appendix.

<sup>11</sup><https://anonymous.4open.science/r/Agnus/README.md>



## 6 Limitations

Due to our introduction of order-invariance by application of a causal mask, modifying positional IDs and introducing a custom logits processor, we are limited to open-weight DLMs, making evaluation with DeepSeek (DeepSeek-AI et al., 2024), GPT-3.5 (Ouyang et al., 2022b), GPT-4 (OpenAI, 2023) impossible, unfortunately. Also, due to hardware restrictions, we are limited to evaluating on significantly smaller models than related work – with us running experiments on models with around 10 billion parameters. For instance, in (Ding et al., 2024a), authors employ GPT3.5 and Llama, claiming that results improve with increased model size.

Fundamentally, we design a causal mask due to being interested in disambiguating entities with causal decoder-only language models. The general idea could likely be transferred to other types of language models by analogously adapting the causal mask to fit another paradigm’s attention masking strategy.

Alike other deep learning approaches to entity disambiguation, AGNUS is limited by its generated candidate sets and by only working with candidate entities that have some form of textual label, description, types or otherwise meaningful information for a LLM to predict.

Our prompt design (for further details, see Figure 7) does not take into account character offsets within input documents due to DLMs running into issues when handling numerical values. Consequently, AGNUS may run into issues when the same mention at different positions in a document refers to different entities. (Example: Tim, CEO of **Apple**, likes to eat an **apple** a day.)

While ACDC does mitigate hallucinations, a given LLM’s next token prediction may be to continue with non-entity tokens, such as a greeting, acknowledgment of task or similar, therewith potentially negatively affecting entity disambiguation depending on decoding algorithm. Designing a specific decoding strategy to include such behaviour could prove to be a benefit in future endeavours.

In this paper, our underlying models are not fine-tuned for the entity disambiguation task nor given particular domain-specific information that could boost context and potential results. Therefore, we concede that going for a few-shot approach could yield improved results.

Further, despite having the out-of-the-box structural capabilities for it, we could not evaluate our

approach on knowledge bases other than Wikipedia and DBpedia due to not being aware of comparable and valid evaluation benchmarks for it.

DLMs are language-dependant and have mainly been trained with English in mind. Therefore, results may vary greatly when our approach is applied to other languages.

Regarding evaluation contamination and the creation of AIDA-Syn, we did not go as in-depth explaining our procedure, safeguards against LLM hallucinations, inherent surrounding bias as we would have liked, nor provide in-depth statistics or analyses. We introduce it mainly as a diagnostic tool to evaluate our approach and show that despite there being novel entities and candidates, AGNUS is capable of attaining similar results as for the non-synthetic version with the suggested least contaminated LLM.

## References

- Anonymous. 2025. Removed to honour ACL anonymity policies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Accepted for publication.
- Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. **Re-fined: An efficient zero-shot-capable approach to end-to-end entity linking**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2022, Hybrid: Seattle, Washington, USA + Online, July 10-15, 2022*, pages 209–220. Association for Computational Linguistics.
- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. ExtEnD: Extractive entity disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online and Dublin, Ireland. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. **On the dangers of stochastic parrots: Can language models be too big?** In *FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.
- Luca Beurer-Kellner, Marc Fischer, and Martin T. Vechev. 2024. **Guiding llms the right way: Fast, non-invasive constrained generation**. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

680	Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie	Damai Dai, Daya Guo, Dejian Yang, Deli Chen,	740
681	Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind	Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai,	741
682	Neelakantan, Pranav Shyam, Girish Sastry, Amanda	Fuli Luo, Guangbo Hao, Guanting Chen, Guowei	742
683	Askell, Sandhini Agarwal, Ariel Herbert-Voss,	Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng	743
684	Gretchen Krueger, Tom Henighan, Rewon Child,	Wang, Haowei Zhang, Honghui Ding, Huajian Xin,	744
685	Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,	Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang,	745
686	Clemens Winter, Christopher Hesse, Mark Chen, Eric	Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang,	746
687	Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,	Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie	747
688	Jack Clark, Christopher Berner, Sam McCandlish,	Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu,	748
689	Alec Radford, Ilya Sutskever, and Dario Amodei.	Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean	749
690	2020. <a href="#">Language models are few-shot learners</a> . In <i>Ad-</i>	Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao,	750
691	<i>Advances in Neural Information Processing Systems 33:</i>	Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang,	751
692	<i>Annual Conference on Neural Information Process-</i>	Mingchuan Zhang, Minghua Zhang, Minghui Tang,	752
693	<i>ing Systems 2020, NeurIPS 2020, December 6-12,</i>	Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang,	753
694	<i>2020, virtual</i> .	Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu	754
695	Nicola De Cao, Gautier Izacard, Sebastian Riedel, and	Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge,	755
696	Fabio Petroni. 2021. <a href="#">Autoregressive entity retrieval</a> .	Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin	756
697	In <i>9th International Conference on Learning Repre-</i>	Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao	757
698	<i>sentations, ICLR 2021, Virtual Event, Austria, May</i>	Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu,	758
699	<i>3-7, 2021</i> . OpenReview.net.	Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu	759
700	Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang,	Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou,	760
701	Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu,	Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun,	761
702	Haizhou Li, Tao Xie, and Baishakhi Ray. 2025. <a href="#">Re-</a>	W. L. Xiao, and Wangding Zeng. 2024. <a href="#">Deepseek-v3</a>	762
703	<a href="#">cent advances in large language model benchmarks</a>	<a href="#">technical report</a> . <i>CoRR</i> , abs/2412.19437.	763
704	<a href="#">against data contamination: From static to dynamic</a>		
705	<a href="#">evaluation</a> . <i>CoRR</i> , abs/2502.17521.		
706	Aakanksha Chowdhery, Sharan Narang, Jacob Devlin,	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and	764
707	Maarten Bosma, Gaurav Mishra, Adam Roberts,	Kristina Toutanova. 2019. <a href="#">BERT: pre-training of</a>	765
708	Paul Barham, Hyung Won Chung, Charles Sutton,	<a href="#">deep bidirectional transformers for language under-</a>	766
709	Sebastian Gehrmann, Parker Schuh, Kensen Shi,	<a href="#">standing</a> . In <i>Proceedings of the 2019 Conference of</i>	767
710	Sasha Tsvyashchenko, Joshua Maynez, Abhishek	<i>the North American Chapter of the Association for</i>	768
711	Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vin-	<i>Computational Linguistics: Human Language Tech-</i>	769
712	odkumar Prabhakaran, Emily Reif, Nan Du, Ben	<i>nologies, NAACL-HLT 2019, Minneapolis, MN, USA,</i>	770
713	Hutchinson, Reiner Pope, James Bradbury, Jacob	<i>June 2-7, 2019, Volume 1 (Long and Short Papers),</i>	771
714	Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin,	pages 4171–4186. Association for Computational	772
715	Toju Duke, Anselm Levskaya, Sanjay Ghemawat,	Linguistics.	773
716	Sunipa Dev, Henryk Michalewski, Xavier Garcia,	Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim	774
717	Vedant Misra, Kevin Robinson, Liam Fedus, Denny	Weninger, Balaji Veeramani, and Sanmitra Bhat-	775
718	Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim,	tacharya. 2024a. <a href="#">Entgpt: Linking generative large</a>	776
719	Barret Zoph, Alexander Spiridonov, Ryan Sepassi,	<a href="#">language models with knowledge bases</a> . <i>CoRR</i> ,	777
720	David Dohan, Shivani Agrawal, Mark Omernick, An-	abs/2402.06738.	778
721	drew M. Dai, Thanumalayan Sankaranarayanan Pil-	Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024b.	779
722	lai, Marie Pellat, Aitor Lewkowycz, Erica Moreira,	<a href="#">Chatel: Entity linking with chatbots</a> . In <i>Proceeed-</i>	780
723	Rewon Child, Oleksandr Polozov, Katherine Lee,	<i>ings of the 2024 Joint International Conference on</i>	781
724	Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark	<i>Computational Linguistics, Language Resources and</i>	782
725	Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy	<i>Evaluation, LREC/COLING 2024, 20-25 May, 2024,</i>	783
726	Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov,	<i>Torino, Italy</i> , pages 3086–3097. ELRA and ICCL.	784
727	and Noah Fiedel. 2023. <a href="#">Palm: Scaling language mod-</a>	Ricardo Dominguez-Olmedo, Moritz Hardt, and Cele-	785
728	<a href="#">eling with pathways</a> . <i>J. Mach. Learn. Res.</i> , 24:240:1–	stine Mender-Dünner. 2023. Questioning the survey	786
729	240:113.	responses of large language models. <i>arXiv preprint</i>	787
730	Silviu Cucerzan. 2007. <a href="#">Large-scale named entity dis-</a>	<i>arXiv:2306.07951</i> .	788
731	<a href="#">ambiguation based on wikipedia data</a> . In <i>EMNLP-</i>	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,	789
732	<i>CoNLL 2007, Proceedings of the 2007 Joint Con-</i>	Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,	790
733	<i>ference on Empirical Methods in Natural Language</i>	Akhil Mathur, Alan Schelten, Amy Yang, Angela	791
734	<i>Processing and Computational Natural Language</i>	Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang,	792
735	<i>Learning, June 28-30, 2007, Prague, Czech Republic,</i>	Archi Mitra, Archie Sravankumar, Artem Korenev,	793
736	pages 708–716. ACL.	Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien	794
737	DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingx-	Rodriguez, Austen Gregerson, Ava Spataru, Bap-	795
738	uan Wang, Bochao Wu, Chengda Lu, Chenggang	tiste Rozière, Bethany Biron, Binh Tang, Bobbie	796
739	Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,	Chern, Charlotte Caucheteux, Chaya Nayak, Chloe	797
		Bi, Chris Marra, Chris McConnell, Christian Keller,	798
		Christophe Touret, Chunyang Wu, Corinne Wong,	799

800	Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-	pages 519–529. Association for Computational Lin-	859
801	lonsius, Daniel Song, Danielle Pintz, Danny Livshits,	guistics.	860
802	David Esiobu, Dhruv Choudhary, Dhruv Mahajan,		
803	Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes,	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio	861
804	Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,	Petroni, Vladimir Karpukhin, Naman Goyal, Hein-	862
805	Emily Dinan, Eric Michael Smith, Filip Radenovic,	rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-	863
806	Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Geor-	täschel, Sebastian Riedel, and Douwe Kiela. 2020.	864
807	gia Lewis Anderson, Graeme Nail, Grégoire Mialon,	<a href="#">Retrieval-augmented generation for knowledge-</a>	865
808	Guan Pang, Guillem Cucurell, Hailey Nguyen, Han-	<a href="#">intensive NLP tasks</a> . In <i>Advances in Neural In-</i>	866
809	nah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov,	<i>formation Processing Systems 33: Annual Confer-</i>	867
810	Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan	<i>ence on Neural Information Processing Systems 2020,</i>	868
811	Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan	<i>NeurIPS 2020, December 6-12, 2020, virtual</i> .	869
812	Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,		
813	Jeet Shah, Jelmer van der Linde, Jennifer Billock,	Ruizhe Li and Yanjun Gao. 2024. Anchored answers:	870
814	Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi,	Unravelling positional bias in gpt-2’s multiple-choice	871
815	Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu,	questions. <i>arXiv preprint arXiv:2405.03205</i> .	872
816	Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph		
817	Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia,	Yucheng Li. 2023. <a href="#">Estimating contamination via</a>	873
818	Kalyan Vasuden Alwala, Kartikeya Upasani, Kate	<a href="#">perplexity: Quantifying memorisation in language</a>	874
819	Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and	<a href="#">model evaluation</a> . <i>CoRR</i> , abs/2309.10677.	875
820	et al. 2024. <a href="#">The llama 3 herd of models</a> . <i>CoRR</i> ,		
821	abs/2407.21783.	Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan	876
822	Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen,	Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023.	877
823	Martin Theobald, and Gerhard Weikum. 2012.	Split and merge: Aligning position biases in large	878
824	<a href="#">KORE: keyphrase overlap relatedness for entity dis-</a>	language model based evaluators. <i>arXiv preprint</i>	879
825	<a href="#">ambiguation</a> . In <i>21st ACM International Confer-</i>	<i>arXiv:2310.01432</i> .	880
826	<i>ence on Information and Knowledge Management,</i>		
827	<i>CIKM’12, Maui, HI, USA, October 29 - November</i>	Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu,	881
828	<i>02, 2012, pages 545–554. ACM</i> .	and Enhong Chen. 2024. <a href="#">Onenet: A fine-tuning free</a>	882
829	Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and	<a href="#">framework for few-shot entity linking via large lan-</a>	883
830	Yejin Choi. 2020. <a href="#">The curious case of neural text</a>	<a href="#">guage model prompting</a> . In <i>Proceedings of the 2024</i>	884
831	<a href="#">degeneration</a> . In <i>8th International Conference on</i>	<i>Conference on Empirical Methods in Natural Lan-</i>	885
832	<i>Learning Representations, ICLR 2020, Addis Ababa,</i>	<i>guage Processing, EMNLP 2024, Miami, FL, USA,</i>	886
833	<i>Ethiopia, April 26-30, 2020. OpenReview.net</i> .	<i>November 12-16, 2024, pages 13634–13651. Associ-</i>	887
834		ation for Computational Linguistics.	888
835	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong,		
836	Zhangyin Feng, Haotian Wang, Qianglong Chen,	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler	889
837	Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting	Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,	890
838	Liu. 2025. <a href="#">A survey on hallucination in large lan-</a>	Nouha Dziri, Shrimai Prabhumoye, Yiming Yang,	891
839	<a href="#">guage models: Principles, taxonomy, challenges, and</a>	Shashank Gupta, Bodhisattwa Prasad Majumder,	892
840	<a href="#">open questions</a> . <i>ACM Transactions on Information</i>	Katherine Hermann, Sean Welleck, Amir Yazdan-	893
841	<i>Systems</i> , 43(2):1–55.	bakhsh, and Peter Clark. 2023. <a href="#">Self-refine: Itera-</a>	894
842	Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu,	<a href="#">tive refinement with self-feedback</a> . In <i>Advances in</i>	895
843	Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea	<i>Neural Information Processing Systems 36: Annual</i>	896
844	Madotto, and Pascale Fung. 2023. <a href="#">Survey of halluci-</a>	<i>Conference on Neural Information Processing Sys-</i>	897
845	<a href="#">nation in natural language generation</a> . <i>ACM Comput.</i>	<i>tems 2023, NeurIPS 2023, New Orleans, LA, USA,</i>	898
846	<i>Surv.</i> , 55(12):248:1–248:38.	<i>December 10 - 16, 2023</i> .	899
847	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-		
848	sch, Chris Bamford, Devendra Singh Chaplot, Diego	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and	900
849	de Las Casas, Florian Bressand, Gianna Lengyel,	Ryan T. McDonald. 2020. <a href="#">On faithfulness and fac-</a>	901
850	Guillaume Lample, Lucile Saulnier, L��lio Ren-	<a href="#">tuality in abstractive summarization</a> . In <i>Proceedings</i>	902
851	nard Lavaud, Marie-Anne Lachaux, Pierre Stock,	<i>of the 58th Annual Meeting of the Association for</i>	903
852	Teven Le Scao, Thibaut Lavril, Thomas Wang, Timo-	<i>Computational Linguistics, ACL 2020, Online, July</i>	904
853	th��e Lacroix, and William El Sayed. 2023. <a href="#">Mistral</a>	<i>5-10, 2020, pages 1906–1919. Association for Com-</i>	905
854	<a href="#">7b</a> . <i>CoRR</i> , abs/2310.06825.	putational Linguistics.	906
855	Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas		
856	Hofmann. 2018. <a href="#">End-to-end neural entity linking</a> .	David N. Milne and Ian H. Witten. 2008. <a href="#">Learning to</a>	907
857	In <i>Proceedings of the 22nd Conference on Compu-</i>	<a href="#">link with wikipedia</a> . In <i>Proceedings of the 17th ACM</i>	908
858	<i>tational Natural Language Learning, CoNLL 2018,</i>	<i>Conference on Information and Knowledge Manage-</i>	909
	<i>Brussels, Belgium, October 31 - November 1, 2018,</i>	<i>ment, CIKM 2008, Napa Valley, California, USA,</i>	910
		<i>October 26-30, 2008, pages 509–518. ACM</i> .	911
		Andrea Moro, Alessandro Raganato, and Roberto Nav-	912
		igli. 2014. Entity Linking meets Word Sense Disam-	913
		biguation: a Unified Approach. <i>Transactions of the</i>	914
		<i>Association for Computational Linguistics (TACL),</i>	915
		<i>2:231–244</i> .	916








1032	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	multiple choice selectors than you think. <i>arXiv preprint arXiv:2404.08382</i> .	1090
1033	Isabel Kloumann, Artem Korenev, Punit Singh Koura,		1091
1034	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-		
1035	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-	1092
1036	tiniet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy,	1093
1037	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	and Barbara Plank. 2024b. "my answer is c": First-	1094
1038	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,	token probabilities do not match text answers in	1095
1039	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	instruction-tuned language models. <i>arXiv preprint</i>	1096
1040	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	<i>arXiv:2402.14499</i> .	1097
1041	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,		
1042	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	1098
1043	Melanie Kambadur, Sharan Narang, Aurélien Ro-	Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,	1099
1044	driguez, Robert Stojnic, Sergey Edunov, and Thomas	and Denny Zhou. 2022. <a href="#">Chain-of-thought prompting</a>	1100
1045	Scialom. 2023. <a href="#">Llama 2: Open foundation and fine-</a>	<a href="#">elicits reasoning in large language models</a> . In <i>Ad-</i>	1101
1046	<a href="#">tuned chat models</a> . <i>CoRR</i> , abs/2307.09288.	<i>Advances in Neural Information Processing Systems 35:</i>	1102
		<i>Annual Conference on Neural Information Process-</i>	1103
1047	Johannes M. van Hulst, Faegheh Hasibi, Koen Derck-	<i>ing Systems 2022, NeurIPS 2022, New Orleans, LA,</i>	1104
1048	sen, Krisztian Balog, and Arjen P. de Vries. 2020.	<i>USA, November 28 - December 9, 2022</i> .	1105
1049	<a href="#">Rel: An entity linker standing on the shoulders of gi-</a>		
1050	<a href="#">ants</a> . In <i>Proceedings of the 43rd International ACM</i>	Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang,	1106
1051	<i>SIGIR Conference on Research and Development in</i>	and Hsin-Hsi Chen. 2024. Unveiling selection bi-	1107
1052	<i>Information Retrieval, SIGIR '20</i> , page 2197–2200,	ases: Exploring order and token sensitivity in large	1108
1053	New York, NY, USA. Association for Computing	language models. <i>arXiv preprint arXiv:2406.03009</i> .	1109
1054	Machinery.		
		Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian	1110
1055	Ruben Verborgh, Michael Röder, Ricardo Usbeck, and	Riedel, and Luke Zettlemoyer. 2020. <a href="#">Scalable zero-</a>	1111
1056	Axel-Cyrille Ngonga Ngomo. 2018. <a href="#">Gerbil – bench-</a>	<a href="#">shot entity linking with dense entity retrieval</a> . In	1112
1057	<a href="#">marking named entity recognition and linking consis-</a>	<i>Proceedings of the 2020 Conference on Empirical</i>	1113
1058	<a href="#">tently</a> . <i>Semant. Web</i> , 9(5):605–625.	<i>Methods in Natural Language Processing, EMNLP</i>	1114
		<i>2020, Online, November 16-20, 2020</i> , pages 6397–	1115
1059	Daniel Vollmers, Hamada M. Zahera, Diego Mous-	6407. Association for Computational Linguistics.	1116
1060	salleem, and Axel-Cyrille Ngonga Ngomo. 2025. <a href="#">Con-</a>		
1061	<a href="#">textual augmentation for entity linking using large</a>	Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun	1117
1062	<a href="#">language models</a> . In <i>Proceedings of the 31st Inter-</i>	Shou, and Daxin Jiang. 2023a. <a href="#">Instructed language</a>	1118
1063	<i>national Conference on Computational Linguistics,</i>	<a href="#">models with retrievers are powerful entity linkers</a> . In	1119
1064	<i>COLING 2025, Abu Dhabi, UAE, January 19-24,</i>	<i>Proceedings of the 2023 Conference on Empirical</i>	1120
1065	<i>2025</i> , pages 8535–8545. Association for Computa-	<i>Methods in Natural Language Processing, EMNLP</i>	1121
1066	tional Linguistics.	<i>2023, Singapore, December 6-10, 2023</i> , pages 2267–	1122
		2282. Association for Computational Linguistics.	1123
1067	Kehang Wang, Qi Liu, Kai Zhang, Ye Liu, Hanqing Tao,	Zilin Xiao, Linjun Shou, Xingyao Zhang, Jie Wu, Ming	1124
1068	Zhenya Huang, and Enhong Chen. 2023a. <a href="#">Class-</a>	Gong, and Daxin Jiang. 2023b. <a href="#">Coherent entity dis-</a>	1125
1069	<a href="#">dynamic and hierarchy-constrained network for en-</a>	<a href="#">ambiguation via modeling topic and categorical de-</a>	1126
1070	<a href="#">tity linking</a> . In <i>Database Systems for Advanced Ap-</i>	<a href="#">pendency</a> . In <i>Findings of the Association for Compu-</i>	1127
1071	<i>plications - 28th International Conference, DASFAA</i>	<i>tational Linguistics: EMNLP 2023, Singapore, De-</i>	1128
1072	<i>2023, Tianjin, China, April 17-20, 2023, Proceedings,</i>	<i>cember 6-10, 2023</i> , pages 7480–7492. Association	1129
1073	<i>Part II</i> , volume 13944 of <i>Lecture Notes in Computer</i>	for Computational Linguistics.	1130
1074	<i>Science</i> , pages 622–638. Springer.		
		Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng	1131
1075	Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu,	Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2024. <a href="#">LL-</a>	1132
1076	Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and	<a href="#">MAEL: large language models are good context aug-</a>	1133
1077	Zhifang Sui. 2023b. Large language models are not	<a href="#">menters for entity linking</a> . <i>CoRR</i> , abs/2407.04020.	1134
1078	fair evaluators. <i>arXiv preprint arXiv:2305.17926</i> .		
		Cheng Xu, Shuhao Guan, Derek Greene, and M. Tahar	1135
1079	Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing	Kechadi. 2024. <a href="#">Benchmark data contamination</a>	1136
1080	Huang, and Zhongyu Wei. 2025. <a href="#">Benchmark self-</a>	<a href="#">of large language models: A survey</a> . <i>CoRR</i> ,	1137
1081	<a href="#">evolving: A multi-agent framework for dynamic</a>	abs/2406.04244.	1138
1082	<a href="#">LLM evaluation</a> . In <i>Proceedings of the 31st Inter-</i>		
1083	<i>national Conference on Computational Linguistics,</i>	Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang	1139
1084	<i>COLING 2025, Abu Dhabi, UAE, January 19-24,</i>	Li, Honglin Han, Meng Zhao, and Chengguo Yin.	1140
1085	<i>2025</i> , pages 3310–3328. Association for Computa-	2024. Strengthened symbol binding makes large	1141
1086	tional Linguistics.	language models reliable multiple-choice selectors.	1142
		<i>arXiv preprint arXiv:2406.01026</i> .	1143
1087	Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger,		
1088	and Barbara Plank. 2024a. Look at the text:		
1089	Instruction-tuned language models are more robust		

- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordini, Marc Spaniol, and Gerhard Weikum. 2011. [AIDA: an online tool for accurate disambiguation of named entities in text and tables](#). *Proc. VLDB Endow.*, 4(12):1450–1453.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.
- Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024a. [Dyval: Dynamic evaluation of large language models for reasoning tasks](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024b. [Dyval 2: Dynamic evaluation of large language models by meta probing agents](#). *CoRR*, abs/2402.14865.

## A Appendix

Over the course of researching and developing AG-NUS , we implemented some further aspects that we could not allude to in depth. Here are some supplemental materials about them that might be of interest to fellow researchers.

### A.1 Prompt Setup & Candidate Representation

In Figure 7, we illustrate how our prompts are set up: we provide a system prompt, a user prompt including a task definition, specific mention to disambiguate, input document text, desired output representation type and a list of (by default) 10 permutable candidate entities. We note that since we do not provide offsets (due to DLMs handling them questionably), we potentially run into issues when multiple mentions referring to different entities are to be disambiguated.

#### A.1.1 IRI

In many of our experiments, utilising human-readable IRI-based representations linking to DBpedia provide best results. Our choice is motivated by IRIs representing the most straightforward – even if knowledge base-dependent – representation for entities. We assume for this to likely be the case due to these being succinct, human-readable, similar to labels and that the start of the candidate (<https://dbpedia.org/resource/>) provides a useful bias to contextually reinforce the representation. We provide an example IRI-based candidate representation prompt in Figure 6.

#### A.1.2 Label

A second candidate representation type we investigate is labels. The choice is motivated by prior work and structural similarity to human-readable Wikipedia- and DBpedia-based IRIs. We illustrate a specific prompt for a chosen example (see Figure 1) in Figure 8.

#### A.1.3 Description

While some of our other representations are likely to have occurred in appropriate contexts within chosen models’ training data, representing candidate entities as descriptions follows the rationale of "contextual reasoning". Descriptions describe entities, give deeper context and use more specific language to define what an entity represents at its core. As such, utilising descriptions for our experiments tests whether (1) longer bits of text may

bias models into different directions and (2) models can reason over contexts, reaching desired answers. Unfortunately, our results with descriptions tend to be worse than with IRIs or labels. We theorise that ACDC’s strictly constrained decoding causes issue with descriptions due to looking to reproduce the exact description while descriptive texts may oftentimes begin with generic ambiguous formulas. A possible improvement could be to change the decoding algorithm to beam search rather than the oftentimes defaulted-to greedy decoding. Further, descriptions may greatly vary in length, causing an underlying DLM to run into unexpected "behaviour" regarding attention and positional embeddings when a large "gap" is perceived between candidate descriptions’ number of ingested tokens.

In Figure 9, we provide a description-based version for our chosen example in which we shorten certain texts for visualization purposes.

#### A.1.4 Type

Finally, we introduce 'types' as a source of information for entities. While simple and potentially ambiguous, the idea was that in combination with other representation types, it could help improve disambiguation by providing more context as used in more traditional entity disambiguation approaches (e.g. applying named entity recognition incl. types and disambiguating based on type-filtered candidates). On another hand, types can also be particularly specific, such as defining an entity as a "Formula One racer" which would prove beneficial to identify a mention Mika as the Finnish-born race driver Mika Häkkinen. While we acknowledge the potentially non-negligible effect, we did not investigate in further depth the provided order of types as it would be out of scope for our current research endeavour. We present our types-specific prompt in Figure 11.

#### A.1.5 Pairwise Representation

In Table 4, we apply disambiguation based on multiple entity representations in a pairwise fashion AIDA. These experiments’ prompts are set up analogously to the ones as illustrated in Figures 6, 7, 8, 9 and 10 with the format defined in Figure 11. In these experiments, pairwise representations are ordered in descending fashion by mean representation scores reached in single-representation experiments (see Table 2): IRI > Label > Description > Type. For instance, in a pair of Label and Type, "candidate representa-

Entity Representation	AGNUS w. IRI	AGNUS w. Type	AGNUS w. Label	AGNUS w. Desc.
AGNUS w. IRI	<b>0.867</b>	<b>0.855</b>	<b>0.763</b>	<b>0.854</b>
AGNUS w. Type	<u>0.855</u>	0.705	0.734	<u>0.766</u>
AGNUS w. Label	0.763	0.734	0.743	0.744
AGNUS w. Desc.	0.854	<u>0.766</u>	<u>0.744</u>	0.763

Table 4: Ablation Study (Entity Representation - Single and Pairwise): Disambiguation results (F1-measure) on AIDA for pairwise and singular (diagonal) entity representation information types for candidates on AGNUS (Mistral): entity IRI, entity type(s), entity label and entity description. Per column top-ranked score in **bold**, second-ranked underlined.

# Documents	Mentions	# Type-Consistent Docs.	Type Consist. (Mean)
888	15,314	331	46.60%

Table 5: Some data statistics for AIDA-Syn. Type-Consistency compares pre-transformation types of entities to post-transformation types of entities and checks overlap.

System	AIDA-Syn	AIDA	ASM-10	ASM-50	ASM-100
Babelify (Moro et al., 2014)	0.7503	0.6729	0.7660	0.7111	0.6912
WAT (Piccinno and Ferragina, 2014)	0.8641	0.6986	0.9355	0.8235	0.8332
REL (van Hulst et al., 2020)	-	?	0.9030	0.7942	0.6829

Table 6: F1 measures on datasets AIDA-Syn, AIDA for AGNUS and GERBIL-available systems (all other publicly available systems on GERBIL (Verborgh et al., 2018) timed out or returned "The annotator caused too many single errors" for the platform despite repeated attempts).

tion type 1" would be defined by Label and "candidate representation type 2" by Type.

## A.2 Candidate Representation - Pairwise Effects

We investigated effects of single candidate representation types within our paper. We considered it interesting to have a look at pairwise combinations thereof as well to verify to what extent adding more information could yield better results – as would be an initial human intuition.

In Table 4, we evaluated AGNUS on pairwise combinations of candidate representation types to verify effects as well as the extent of increased information content on results. We note that disambiguating based on meaningful IRIs, such as from Wikipedia (e.g. [https://en.wikipedia.org/wiki/Mika\\_\(singer\)](https://en.wikipedia.org/wiki/Mika_(singer))), yields the best scores regardless of representation it may be combined with. Any further representation type worsens results, seemingly indicating that highly-defining compact representations may yield best results.

Types by themselves return mixed results, slightly improving upon description-based candidates, but deteriorate label-based results slightly. This may be due to the high overlap among candidates for this representation, potentially causing

confusion upon disambiguation and yielding worst results (7.0%) in our experiments. Adding labels (7.3%) or descriptions (7.7%) to types increases candidate information, decreasing ambiguity and leading to improved results. Labels as an entity characteristic by themselves (7.43%) are relatively ambiguous, but benefit slightly from further information in the form of descriptions (7.44%). Overall, top scores are reached with IRI representations regardless of other combined information – actually suffering from any additional representations (by itself: 8.67%, with type(s): 8.55%, with description: 8.54%) –, most notably suffering from labels (7.63%).

## A.3 Masked Attention Candidate Set - Details

MACS hides certain tokens' positions from other tokens without requiring retraining by restricting attention and sharing positional embeddings for successive token predictions. In Figure 12, we illustrate an example of the encoding to an underlying DLM: the positional embedding ID for each candidate is reset to the first candidate's positional embedding and incremented for each token until another candidate entity or the end of the candidate entity set is encountered. The positional embedding of the first token succeeding a masked attention candidate set is set to the longest candidate's final token's positional ID incremented by one.

As such, to the underlying DLM it will appear as though there was a gap in positions. Consequently, prompts including candidates with highly varying lengths may lead to weirdness for the underlying DLM's decoding process.



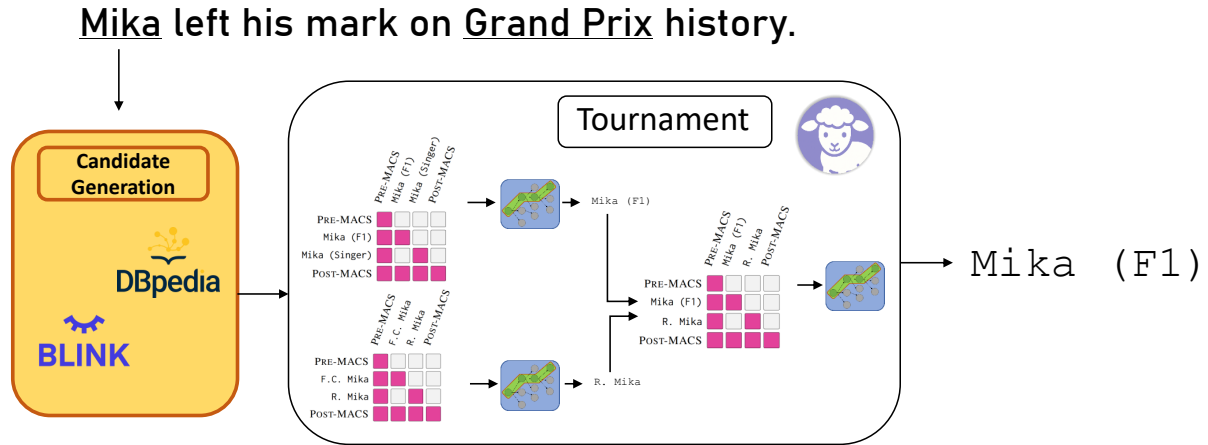


Figure 5: Agnus Hierarchical Elimination Tree: When context window exceeds LLM capabilities, Agnus makes use of a hierarchical elimination tree, splitting the disambiguation task into smaller ones, each of size  $k$  where  $k$  is the number of allowed candidates to not exceed the context window.

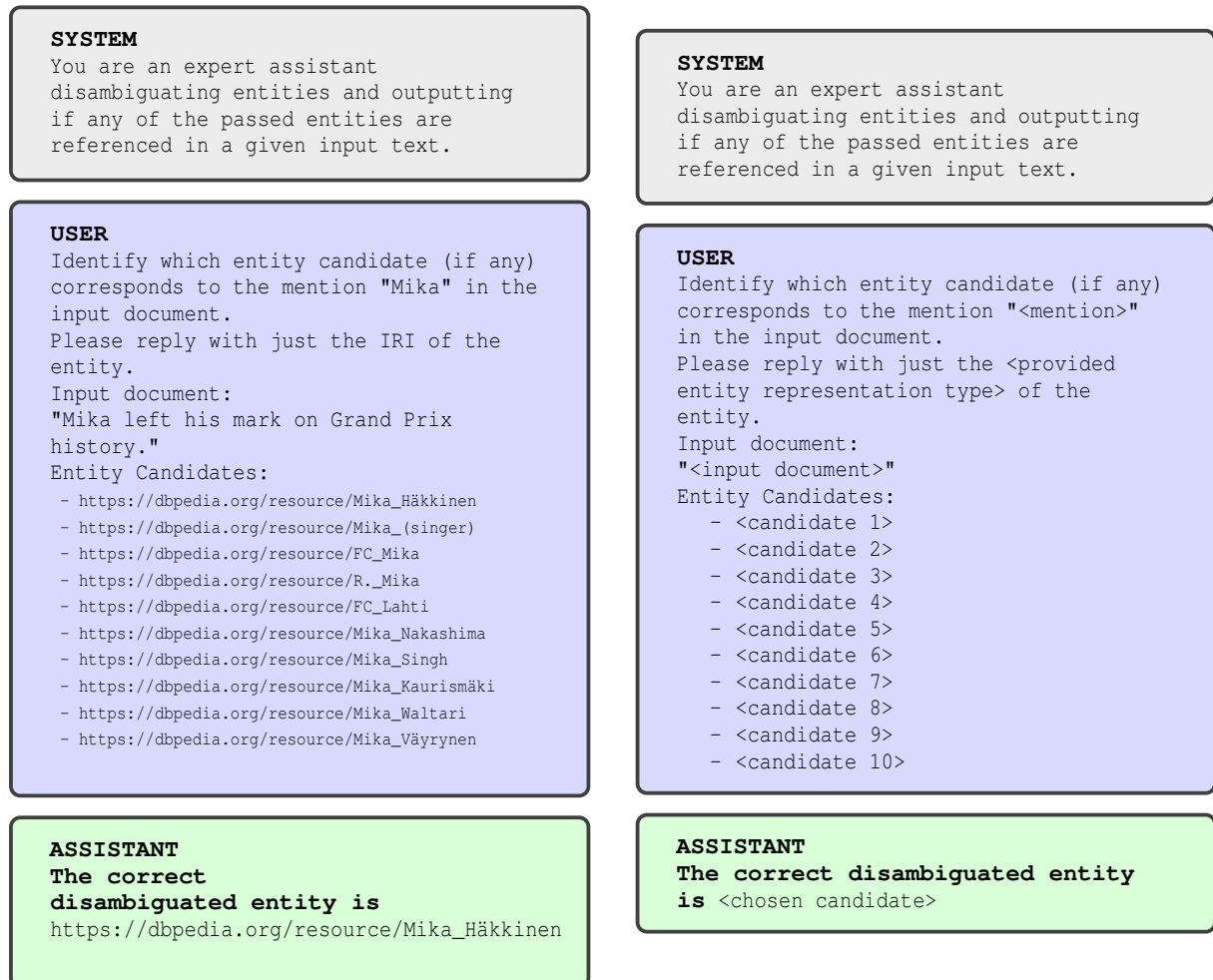


Figure 6: Prompt - IRI: Entity Candidates represented by their DBpedia-grounded entity IRIs.

Figure 7: Prompt - Template: Template variables are surrounded by less than (<) and greater than (>) symbols.

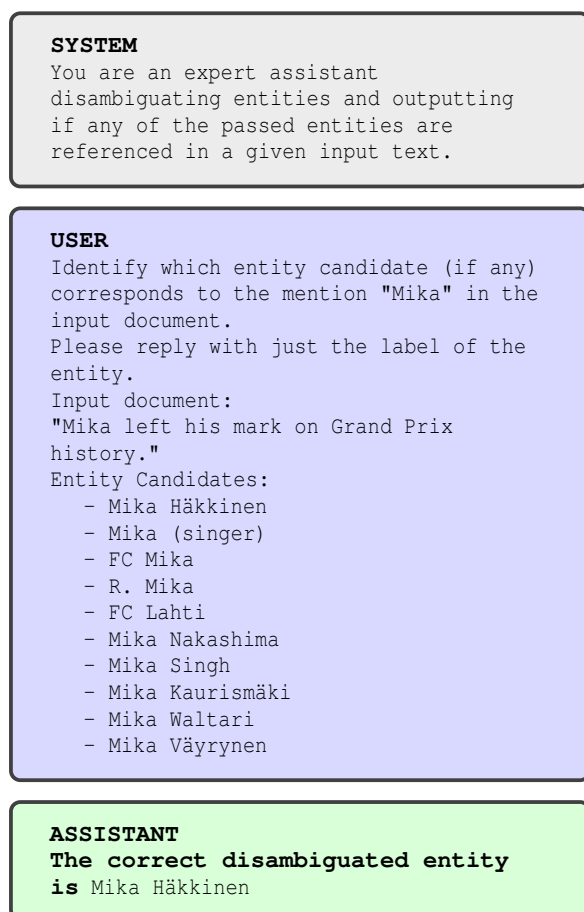


Figure 8: Prompt - Label: Candidates represented by each entity's respective label (rdfs:label).

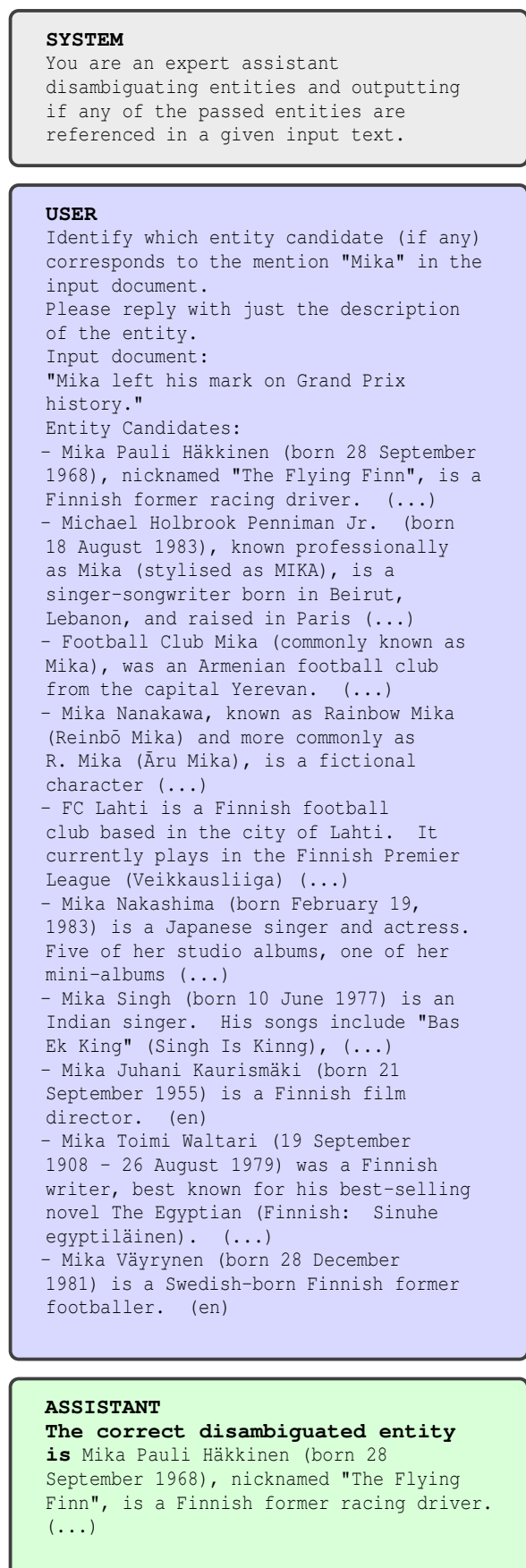
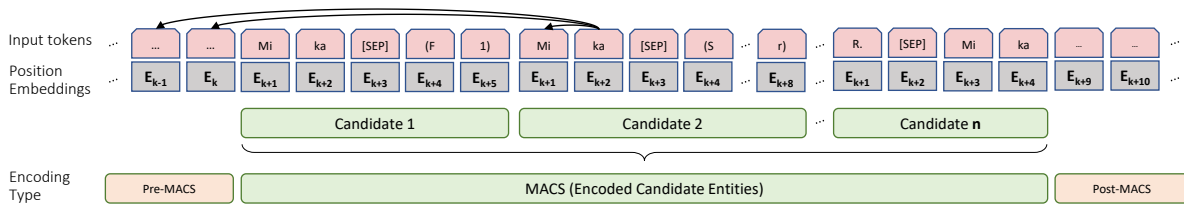


Figure 9: Prompt - Description: Candidates represented by respective descriptions queried from DBpedia (dbo:abstract). "(...)" indicates text was clipped for visualization purposes.



Figure 11: Prompt - Pairwise Template: Prompt Template for pairwise representation type experiments. Template variables are surrounded by less than (<) and greater than (>) symbols.

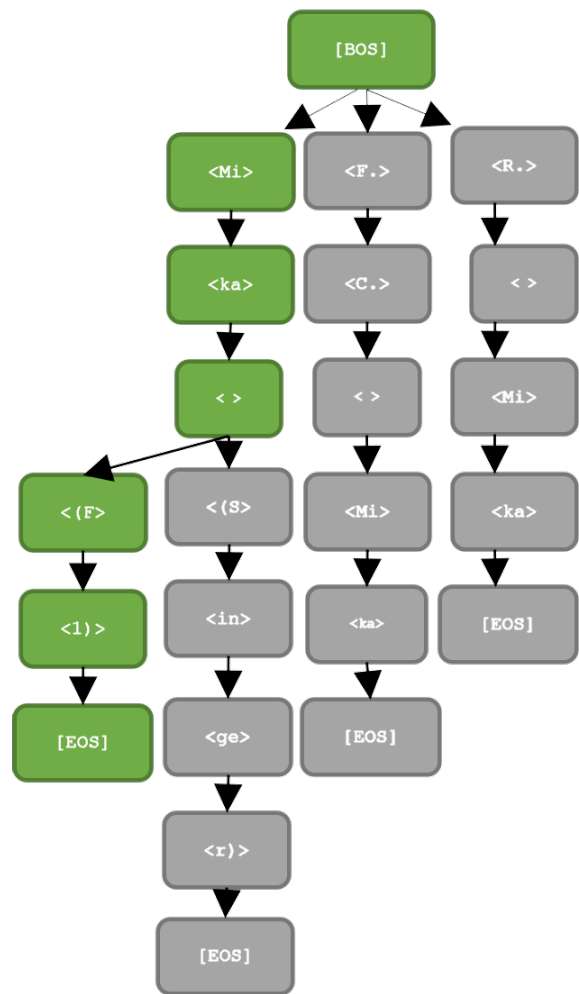
Figure 10: Prompt - Types: Candidates represented by respective types' (rdf:type) in string representation. "(...)" indicates text was clipped for visualization purposes.



#### A.4 Agnus Contextual Decoding - Details

Our constrained decoding mechanism functions in a tree-based fashion in accordance to the grammar defined in Section 3.3 and only allows for specific tokens at each step by setting disallowed tokens to negative infinity ( $-\text{inf}$ ) with a customised logit processor. Effectively, this leads to undesired tokens being impossible to be generated by the language model. In Figure 13, we visualize in further detail a step-by-step decoding correctly disambiguating "Mika (F1)". Each node represents a token: a green node signifies a token being decoded by the model, grey means not decoded despite previously having been a possible choice. Our approach aims to be minimally invasive and maximally generalisable to open-weight models in the sense that we do not define nor modify a particular decoding algorithm. Instead, our models use the default or otherwise defined decoding algorithms for the respective language model (i.e. *greedy decoding*, *top-p sampling* or similar).

Furthermore, we introduce a simple optimization involving an "early stopping" mechanism when decoded entity representations start being no longer ambiguous (see constrained decoding visualization in Figure 13). Halting representation decoding prior to an end of sequence token yields the same results as when generation is continued, but speeds up the disambiguation process significantly – particularly for longer representation types, such as descriptions (see Figure 9).



## A.5 AIDA-Syn

We created AIDA-Syn using DeepSeek-R1 and generated 5 variants of coherent mentions and entities each. We automatically filtered out variants and documents where entities did not correspond to a valid DBpedia entity or where other LLM-related issues may have arisen. Some issues were related to DeepSeek’s maximum number of generated tokens



activating prior to reaching the end. Key criteria for our generation included semantic coherence, lexical diversity, naturalness, plausibility within the surrounding text, and alignment with existing entities – our employed setups, prompts (Figure 14) and results (Figure 15) are publicly available<sup>12</sup>. In the end, we generated a collection of 888 synthetic documents with 5 variants of mentions and entities for each. The reasoning behind the latter being more options for future evaluation endeavours, as well as switching to alternative mention contexts in case of faulty generations. Note that due to hardware limitations, we relied on DeepSeek’s API rather than employing ACDC with DBpedia entities for the synthetic dataset generation.

As a means of verifying that our generated mentions and entities are sensical, we used a two-pronged approach. First, two researchers manually validated a random sample of 222 (25%) documents, verifying contextual coherence for all variants. In 14% (32 documents) of documents, generated mentions created excessive entity ambiguity or were incorrect, leading to using another set of generated mentions and entities for the documents.

Second, we attempted to run the full suite of annotators via GERBIL (Verborgh et al., 2018) to see whether existing approaches could annotate documents effectively – we report the results in Table 6. Unfortunately, many D2KB annotators did not run on our full AIDA-Syn (or ASM-10<sup>13</sup>, ASM-50<sup>14</sup>, ASM-100<sup>15</sup>) and the original AIDA datasets, instead returning timeout errors and similar. In Table 5 we display some details about the synthetic AIDA-Syn dataset including number of documents (888), total number of mentions (15,314) as well as entity type consistency between the original dataset and the transformed documents. Our assumption is that a certain degree of overlap between types should persist, but that it shouldn’t be an absolute overlap the sake of document diversity.

### A.5.1 Synthetic Data Generation Caveats

Generating novel mentions can result in a variety of ways that lessen expressiveness of data. For instance, in Figure 15, "Mika"

<sup>12</sup><https://anonymous.4open.science/r/Agnus/README.md>

<sup>13</sup><http://gerbil.aksw.org/gerbil/experiment?id=202505190000>

<sup>14</sup><http://gerbil.aksw.org/gerbil/experiment?id=202505190001>

<sup>15</sup><http://gerbil.aksw.org/gerbil/experiment?id=202505190002>

is transformed into following alternative mentions: Ayrton Senna, Michael Schumacher, Alain Prost, Lewis Hamilton and Sebastian Vettel. In this case, a simple first name ("Mika") is transformed into a first name followed by a last name, both belonging to famous race drivers. The presence of the last name creates a lower degree of ambiguity than in the initial dataset, reducing complexity of the disambiguation task.

Further, in this setup, the linked wikipedia ID ("wiki") may be hallucinated, making the alternative document effectively unuseable.

Similarly, a DLM may hallucinate a mention and link an unrelated entity to it.

All of these points put into question the validity of utilising synthetic data for evaluation, but still allow for a certain degree of expressiveness regarding contamination diagnostics.

### A.6 Context Length

The number of tokens an LLM may process at once and is therefore limited to is known as *context length*. If tokens surpass the maximal context length an LLM was trained for, it produces gibberish at an increased likelihood – to the point that some LLMs will opt to instead raise an error when a threshold is reached. To counteract the issue of context length for LLM-based entity disambiguation, we introduce *Hierarchical Elimination Tree Disambiguation*, a linearly scalable disambiguation method for iterative pruning of unwanted candidate entities – alike single-elimination tournaments.

We did not run into issues relating to context length in our experiments when comparing with prior work due to the limited number of candidates. Regardless, we developed a relatively simple approach allowing to sidestep the context length issue (see Figure 5).

The approach resolves the problem of context length by transforming the disambiguation task of  $|C|$  candidates into tasks of smaller subsets of at most  $k$  disambiguation candidates instead, aggregating results and repeating the AGNUS disambiguation process (see Fig. 1) with further subsets of candidate entities until disambiguation converges on one entity. Formally, with  $N = |C|$  entity candidates,  $k$  maximum threshold for concurrent candidates and  $j \in (1, \dots, \lceil \log_k N \rceil)$ , Hierarchical Elimination Tree Disambiguation (HET) leads to  $1 + \sum_{j=1}^{\lceil \log_k(n) \rceil} \lceil \frac{n}{k^j} \rceil$  disambiguation tasks of at most size  $k$  each being computed.

### USER

Prompt for Generating Multiple New Mentions Entities Task:  
Given an input text with entity mentions, generate 5 distinct modified versions of the text, each with:

New mentions replacing the original mentions.

Valid Wikipedia IDs for each new mention.

Semantic coherence (new mentions fit the context and work together logically).

Input:

A text snippet with entity mentions.

A list of mentions to replace (with Wikipedia IDs if available).

Output:

A JSON list of 5 alternatives, each adhering to the original structure. Each alternative must:

Replace all original mentions with new ones.

Include unique combinations of mentions/entities across alternatives.

Output Requirements:

Placeholder-ized Text: Replace each mention in input\_text with {0}, {1}, etc., in order of occurrence.

Replacements: For each alternative, provide:

entities: Array of {new: "Replacement", wiki: "Wikipedia\_ID"} (aligned with mentions order).

Constraints:

All new mentions must fit the original context.

All wiki IDs must exist on Wikipedia.

Handle duplicate mentions.

Output Format (JSON): json

```
{ "input_text": "Text with {0}, {1} and so on.", "mentions": ["Mention1", "Mention2", ...],  
  "replacements": [ { "entities": [ { "new": "Replacement1", "wiki": "Wiki_ID1"}, { "new":  
    "Replacement2", "wiki": "Wiki_ID2"} // as many as there are mentions ] } // ... N-1 more  
alternatives ] }
```

Additional Guidelines:

- Diversity: Ensure the 5 alternatives cover meaningfully different entities (e.g., different cities, agencies, brands or similar).
- Naturalness: New mentions should sound natural, plausible and coherent in the text.
- Wikipedia Validation: All wikipedia\_id fields must link to existing pages.
- Consistency: Within each alternative, new mentions should not contradict each other (e.g., "FBI" and "Tokyo" is fine; "FBI" and "Medieval Knights" is not).

Example Input: text

Original Text: "Samsung released a new phone in Seoul."

Mentions: ["Samsung", "Seoul"]

Wikipedia IDs: ["Samsung", "Seoul"]

Example Output: json

```
{ "modified_text": "{0} launched a new smartphone in {1}.", "mentions": ["Samsung",  
  "Seoul"], "replacements": [ { "entities": [ { "new": "Apple", "wiki": "Apple_Inc."},  
    { "new": "Cupertino", "wiki": "Cupertino_California"}, ] }, { "entities": [ { "new":  
    "Nokia", "wiki": "Nokia"}, { "new": "Helsinki", "wiki": "Helsinki"}, ] }, // ... more  
alternatives ] }
```

Actual input: text

Original Text: "Mika left his mark on Grand Prix history."

Mentions: ["Mika", "Grand Prix"]

Wikipedia IDs: ["Mika\_Häkkinen", "Formula\_One"]

Figure 14: Prompt - AIDA-Syn: A prompt with example to generate 5 alternative mentions based on an existing AIDA document which will then be validated using automatic and human-in-the-loop methods.

**ASSISTANT**

```
{
  "input_text": "{0} left his mark on {1} history.",
  "mentions": ["Mika", "Grand Prix"],
  "replacements": [
    {
      "entities": [
        {"new": "Ayrton Senna", "wiki": "Ayrton_Senna"},
        {"new": "Monaco Grand Prix", "wiki": "Monaco_Grand_Prix"}
      ],
    },
    {
      "entities": [
        {"new": "Michael Schumacher", "wiki": "Michael_Schumacher"},
        {"new": "F1 World Championship", "wiki": "Formula_One_World_Championship"}
      ],
    },
    {
      "entities": [
        {"new": "Alain Prost", "wiki": "Alain_Prost"},
        {"new": "French Grand Prix", "wiki": "French_Grand_Prix"}
      ],
    },
    {
      "entities": [
        {"new": "Lewis Hamilton", "wiki": "Lewis_Hamilton"},
        {"new": "British Grand Prix", "wiki": "British_Grand_Prix"}
      ],
    },
    {
      "entities": [
        {"new": "Sebastian Vettel", "wiki": "Sebastian_Vettel"},
        {"new": "Singapore Grand Prix", "wiki": "Singapore_Grand_Prix"}
      ],
    },
  ]
}
```

Figure 15: Synthetic Data Generation Response: JSON data with the passed prior input text, mentions, as well as replacement mentions and entities. Replaced mentions for "Mika" are fellow race drivers and the replacements are some form of racing championship.

AIDA*	100 cand., $k = 2$	10 candidates
AGNUS (Mistral)	0.8329	0.8669

Table 7: ED for AGNUS on 581 AIDA documents with 100 candidates (HET,  $k = 2$ ) and 10 candidates.

In Table 7, we show the result of a computed 'stress test' for HET with 100 candidates and  $k = 2$  to maximise the number of disambiguation runs to see how much performance would deteriorate for AIDA. Our 'worst-case' HET experiment creates 7 elimination rounds and 98 disambiguation tasks for each mention. By doing so, the likelihood of potentially propagating errors increases, but the performance difference between our HET-activated stress test and usual-setting AGNUS is only 3.4% for AIDA<sup>16</sup>. Nevertheless, we do not recommend running HET with  $k = 2$  unless absolutely necessary for a small context length (e.g. when introducing multiple shots) due to the unnecessarily large amount of disambiguations to be performed for large candidate sizes. For reference, with  $k = 10$  and 100 candidates, it would still be 7 elimination rounds, but with a total of 12 disambiguations of at most 10 candidates each.

### A.7 Notes on Baseline Experiments

Do note that in the case of "*w.o. ACDC*" (without constrained decoding), we apply fuzzy matching between candidate representations for both predicted and expected values, ranking similarity for the sake of comparison fairness and picking the highest-overlap-similarity candidate as a match. Just using the results as-is for a "baseline" comparison seemed disingenuous as "exact matching" criteria would put baseline results very close (if not exactly) to 0 in most cases.

Applying hard-prompting based finetuning to our employed suite of large language models would likely alleviate the effects to a certain degree, but would simultaneously render the comparison invalid due to comparing our zero-shot model to a 1-shot baseline, therewith having only limited expressivity over our existing ED evaluation table (Table 1).

Due to similar reasons, our baseline without candidates still uses matching to candidates (it did not see or produce) rather than dryly applying an exact

<sup>16</sup>Please note that we perform our evaluation on 581 documents from the AIDA dataset due to a flat multiplier of 98 costing unnecessary amounts of time and electricity on our limited hardware.

matching scheme, therewith heightening the likelihood of correct results. Hence, we urge readers to not overestimate baseline performance.

### A.8 Evaluating with Related Work Candidates & Details

Unfortunately, to the best of our knowledge, the large majority of prior work does not provide candidate entities for their entity disambiguation methods. We provide our candidates, data and results (see code repository). We have some comments and concerns with prior work's (EntGPT (Ding et al., 2024a)) provided candidates<sup>17</sup>, but we want to validate our approach best as possible and see an added benefit for comparability in doing so.

AGNUS always chooses a single entity from among a set of 10 candidates. EntGPT's candidate sets are variable in size and less than or equal in amount to 10.

Further, EntGPT relies on a prompt structure allowing underlying DLMs to express that "None" of the provided candidates correspond to the desired one. This unfortunately entails a few additional considerations regarding evaluation: Theoretically, (1) if a candidate generation technique were to never contain a desired entity for disambiguation, a DLM could technically always choose "None" and reach a perfect score. (2) Limited comparability to existing ED methods.

Also, according to our analyses (see Table 8), some datasets are incomplete in terms of documents and mentions, therefore making meaningful comparisons with other existing work difficult. Despite concerns regarding generalizability to other methods, we regard comparing AGNUS to the best of our knowledge the only prior work that explicitly provides candidate entities a meaningful endeavour. We note that while some prior work do technically provide code to generate candidates, provided code being impossible to run without possible major changes (e.g. local dependencies (Liu et al., 2024)<sup>18</sup> or paths to inaccessible datasets (Xiao et al., 2023a)<sup>19</sup>) significantly impedes a comparable and clean evaluation process.

We evaluate AGNUS with existing candidate en-

<sup>17</sup>[https://github.com/yifding/In-Context\\_EL/tree/main/RUN\\_FILES/4\\_13\\_2023/rel\\_blink/evaluation\\_new\\_one\\_step](https://github.com/yifding/In-Context_EL/tree/main/RUN_FILES/4_13_2023/rel_blink/evaluation_new_one_step)

<sup>18</sup>[https://github.com/laquabe/OneNet/blob/main/pointwise\\_process/listwise\\_cand.py](https://github.com/laquabe/OneNet/blob/main/pointwise_process/listwise_cand.py)

<sup>19</sup>[https://github.com/MrZilinXiao/InsGenEntityLinking/blob/master/data\\_scripts/create\\_candidates\\_dict.py](https://github.com/MrZilinXiao/InsGenEntityLinking/blob/master/data_scripts/create_candidates_dict.py)



1555 titles for the limited documents where available  
1556 (Table 8) and display our results in Table 9. Unfor-  
1557 tunately, AGNUS (Mistral) with MACS performs  
1558 significantly worse than EntGPT-P (with GPT3.5).  
1559 The difference in results may partially be attributed  
1560 to (1) utilising EntGPT’s prompt setup without fur-  
1561 ther prompt-tuning to our expected representation  
1562 types, potentially causing issue for constrained de-  
1563 coding and (2) utilising its prompt as-is preventing  
1564 us from benefitting from EntGPT’s context self-  
1565 augmentation. We have noticed significant uplifts  
1566 with constrained decoding when expressly adapting  
1567 prompts to only output an expected candidate.

Dataset Dataset	In-Set True Exists	"None" None Correct	Docs. Samples	Docs. (Original) Mentions	Mentions Candidates	Candidates Min	Min Max	Max Avg Cand	Avg Cand
KORE50	113	35	50	50	148	1365	1	10	9.223
ACE2004	242	15	35	106	257	1953	1	10	7.599
AIDA-B	4250	125	230	231	4375	31651	1	10	7.235
AQUAINT	700	27	50	50	727	4935	1	10	6.788
CLUEWEB	9961	1193	320	-	11154	83285	1	10	7.467
MSNBC	617	39	20	20	656	4525	1	10	6.898
OKE2015	441	95	101	101	536	3625	1	10	6.763
OKE2016	240	48	55	-	288	2179	1	10	7.566
Reuters-128	544	106	113	128	650	4686	1	10	7.209
RSS-500	447	77	357	500	524	3199	1	10	6.105
WIKI	6076	717	319	-	6793	42296	1	10	6.226

Table 8: Prior Work (Ding et al., 2024a) Dataset Statistics: Number of mentions for which correct entity is within candidate set (In-Set), is not in candidate set ("None" being correct), number of documents provided and number of documents within dataset originally (as far as could be determined reasonably). Dash (-) means varying values have been found from different sources.

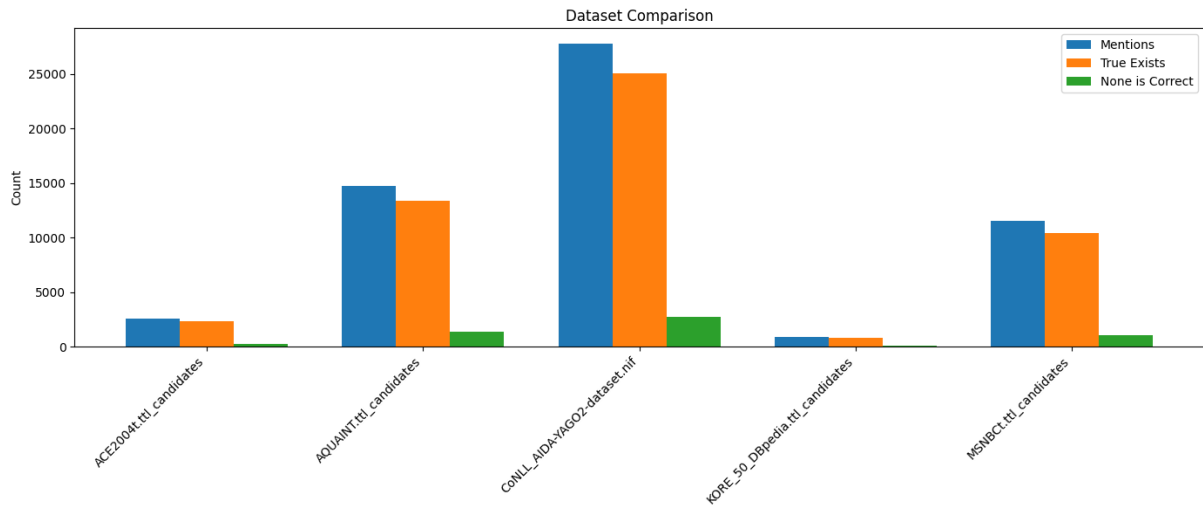


Figure 16: AGNUS Candidates using DBpedia Lookup Finder: Number of mentions (blue); True candidate exists within candidates (Orange); True candidate is not within candidates (and "NIL" / "None" should be chosen) (Green).



Model	ACE04*	AIDA*	AQU*	CWEB*	KORE50*	MSN*	OKE15*	OKE16*	REU*	RSS*	WIKI*
AGNUS  (Mistral)	0.751	0.676	0.634	0.538	0.607	0.767	0.633	0.598	0.752	0.709	0.641
EntGPT-P (GPT3.5)	0.884	0.821	0.791	0.709	0.716	0.867	0.767	0.770	0.785	0.808	0.771

Table 9: Evaluation (Micro F1): AGNUS  and EntGPT (Ding et al., 2024a) entity disambiguation performed with EntGPT’s candidates and limited documents. Some provided datasets (see Table 8) fewer documents and mentions than in the originals. Therefore, we recommend to not compare these values to other prior work with similar-named benchmark names. EntGPT scores taken from their repository.