Appendix of "Selecting the Right Experts: Generalizing Information Extraction for Unseen Scenarios via Task-Aware Expert Weighting"

Lubingzhi Guo[a,*], Javier Sanz-Cruzado[a] and Richard McCreadie[a]

[a]University of Glasgow, Glasgow, United Kingdom
ORCID (Lubingzhi Guo): https://orcid.org/0009-0001-9283-5747, ORCID (Javier Sanz-Cruzado): https://orcid.org/0000-0002-7829-5174, ORCID (Richard McCreadie): https://orcid.org/0000-0002-2751-2087

A Dataset Details

Tables 1 and 2 present the detailed statistics for the training and zero-shot datasets, respectively. To provide a more comprehensive evaluation of generalization ability of our model on unseen data, we extend the original zero-shot datasets. This includes a set of relation extraction datasets and the PHEE dataset for EE and EAE. For newly added datasets, we follow the guidelines provided by the authors of each dataset where available. If such guidelines are not publicly available, we manually create them based on the underlying concept. These are then used as docstring annotations to align with the format and style of the training dataset.

Table 1. Supervised Dataset Statistics

Task	Dataset	Domain	Train	Dev	Test
NER	$ACE05_{NER}$	News	19217	901	676
	$ACE05_{VER}$	News	19217	-	-
	CoNLL 2003	News	14041	3250	3453
	Ontonotes 5	News	30000	15680	12217
	WNUT 2017	Social Media	3394	1009	1287
	BC5CDR	Biomedical	4561	4582	4798
	NCBIDisease	Biomedical	5433	924	941
	DIANN	Biomedical	3976	793	1309
RE	$ACE05_{RE}$	News	19217	901	676
	$ACE05_{RC}$	News	5691	-	-
EE&EAE	$ACE05_{EE}$	News	19217	901	676
	$ACE05_{EAE}$	News	3843	397	368
	RAMS	News	7329	924	871
SF	TACRED	News	10027	3896	2311

B Task Encoder Details

In our ablation experiments, we evaluate two categories of task encoders. The specific model names used for each category are listed below.

- **Code Encoders:**
  - codet5: Salesforce/codet5-base
  - codebert: microsoft/codebert-base
  - graphcodebert: microsoft/graphcodebert-base
  - codet5p-110m: Salesforce/codet5p-110m-embedding
  - jina-embeddings: jinaai/jina-embeddings-v2-base-code
- **Text Encoders:**
  - bge: BAAI/bge-large-en
  - gte: Alibaba-NLP/gte-large-en-v1.5

* Corresponding Author. Email: l.guo.1@research.gla.ac.uk

Table 2. Zero-shot Dataset Details

Task	Dataset	Domain	Test
NER	BroadTwitter	Social Media	2002
	HarveyNER	Social Media	1303
	CrossNER_AI	AI	431
	CrossNER_Literature	Literature	416
	CrossNER_Music	Music	465
	CrossNER_Politics	Politics	651
	CrossNER_Science	Scientific	543
	FabNER	Scientific	2064
	MIT Movie	Social Media	2443
	MIT Restaurants	Social Media	1521
	E3C	Biomedical	851
	MultiNERD	Wikipedia	16454
	WikiEvents	Wikipedia	573
RE	SciERC	Scientific	397
	SemEval	Scientific	2714
	ADECorpus	Biomedical	428
	CoNLL 2004	News	287
	NYT11-HRL	News	362
	KBP37	News	3405
	GIDS	News	4307
EE	WikiEvents	Wikipedia	573
	CASIE	Cybersecurity	2000
	PHEE	Biomedical	968
EAE	WikiEvents	Wikipedia	321
	CASIE	Cybersecurity	2000
	PHEE	Biomedical	968

C Affects of Ontonotes Dataset Sampling

As mentioned in the paper, the training dataset still presents some variations from the original GoLLIE framework. Specifically, the OntoNotes 5 dataset, which is automatically annotated and contains over 100,000 examples, was randomly sampled to 30,000 examples per training epoch to mitigate potential bias. This process was performed independently for each epoch without a fixed seed, making it challenging to replicate the exact sampling behavior. This led to

a performance drop, particularly on the OntoNotes 5 dataset, affecting both supervised and zero-shot performance. To investigate this performance variation and minimize the influence of outliers, we conducted five sampling iterations using different random seeds, training rank-32 LoRA models. The results, shown in Table 3, consistently demonstrate poorer performance across these runs. Note that the zero-shot results are derived from a smaller set of datasets, as reported by GoLLIE. Additionally, the original framework used QLoRA with a LoRA rank of 8, an alpha of 16, and a larger per-device batch size. However, due to resource constraints and implementation differences, we could not replicate this configuration in our experiments.

Table 3. Ablation Results Across OntoNotes Reproduction Variants

	Ontonotes 5	Supervised Avg. F1	Zero-shot Avg. F1
GoLLIE	83.4±0.2	73.0±0.3	55.3±0.2
Reproduced LoRA	70.5±2.8	71.2±0.4	53.4±0.4

D Seen and Unseen Labels

Table 5 lists the labels, categorized into two groups: those with concepts overlapping with the training corpus and those without overlap. For the newly added datasets, we follow a strict strategy: if a concept is similar to one in the training corpus, even when associated with a different task, we classify it as a seen label. Table 4 reports the performance of the compared models on both the seen and unseen label sets.

Table 4. Performance for Seen and Unseen Labels in Zero-Shot Setting. The best averaged performance is highlighted in bold. Arrows (↑ and ↓) indicate comparisons with the best-performing LoRA variant.

PEFT Method	LoRA				MoELoRA					
Variant	rank 32		rank 36		Token-Only		Schema-Aware		Task-Aware	
Label	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
NER										
BroadTwitter	47.2	-	46.7	-	48.7	-	49.2	-	50.2	-
HarveyNER	-	38.5	-	35.7	-	36.8	-	41.3	-	36.7
AI	65.8	54	61.6	48.4	62.2	60.5	67	60.6	63.7	59.1
Literature	58.4	66	56.4	67.7	60.8	68.6	63.7	73.1	67.3	71.5
Music	52.9	72.2	52.3	70.2	66.9	77.7	68.2	79.7	68.9	77.2
Politics	66.9	52.8	63.5	32.8	68.3	52.4	70.7	62.3	66.9	39.5
Science	62.3	54	57.7	49.2	61.2	60.1	63.8	58.2	58.1	56.5
FabNER	21.7	25.6	13.2	23	22.6	24.8	17.9	24.8	21.6	25.4
MIT Movie	61.1	60.1	57.6	62.1	65.4	63.6	64.5	62.3	70	63.7
MIT Restaurants	34.6	46.4	32.	47.9	41	51.9	39.9	53.8	47.9	55.7
E3C	57.8	-	56.4	-	61.6	-	60.6	-	62.7	-
MultiNERD	80.1	36.4	80.3	43.1	81	45.8	80.9	46.8	81.8	50.8
$\text{WikiEvents}_{\text{NER}}$	82.1	0	81.5	3.4	78.4	0	75	3.5	76.1	0
Average	57.6	46.0	54.9	44.0	$59.8^{\uparrow 2.2\%}$	$49.3^{\uparrow 3.3\%}$	$60.1^{\uparrow 2.5\%}$	$\mathbf{51.5}^{\uparrow 5.5\%}$	$\mathbf{61.3}^{\uparrow 3.7\%}$	$48.7^{\uparrow 2.7\%}$
RE										
SciERC	3	0	11.9	0	7.6	5.2	11.9	1.7	25	12.9
SemEval	19.8	4.7	29.9	10.3	18	14.3	27.9	16.3	36	25.9
ADECorpus	-	24	-	24.6	-	38.3	-	37	-	53.4
CoNLL 2004	20.2	-	21.2	-	38.7	-	53.3	-	59	-
NYT11-HRL	39.2	2.3	47.5	0	44.2	2.8	39.1	1.2	43.9	3.3
KBP37	9.5	-	16.1	-	22.6	-	22.8	-	25.4	-
GIDS	57.6	40.1	67	42.9	68.9	67	76.3	79.7	78.8	83.4
Average	24.9	14.2	32.3	15.6	$33.3^{\uparrow 1.0\%}$	$25.5^{\uparrow 9.9\%}$	$38.6^{\uparrow 6.3\%}$	$27.2^{\uparrow 11.6\%}$	$\mathbf{44.7}^{\uparrow 12.4\%}$	$\mathbf{35.8}^{\uparrow 20.2\%}$
EE&EAE										
$\text{WikiEvents}_{\text{EE}}$	44.2	48.7	45.4	48.1	43.9	45	43.8	49.7	41.5	42.5
CASIE_{EE}	-	48.8	-	50.7	-	56.5	-	61.6	-	68
PHEE_{EE}	-	55.7	-	57.2	-	63.3	-	67.1	-	70.3
$\text{WikiEvents}_{\text{EAE}}$	45.7	32.2	46.4	29.6	44.5	32.6	43.7	42.4	42	37.9
$\text{CASIE}_{\text{EAE}}$	37.5	21.2	36.9	20.9	40.3	23.4	41.1	24.1	40.4	22.4
PHEE_{EAE}	23.3	37.6	23.8	35.8	24.4	42.6	21.2	45.7	24.8	47.1
Average	37.7	40.7	38.1	40.4	$\mathbf{38.3}^{\uparrow 0.2\%}$	$43.9^{\uparrow 3.2\%}$	$37.5^{\downarrow 0.6\%}$	$\mathbf{48.4}^{\uparrow 7.7\%}$	$37.2^{\downarrow 0.9\%}$	$48.0^{\uparrow 7.3\%}$
Average All	45.0	37.3	45.7	36.5	$48.7^{\uparrow 3\%}$	$42.4^{\uparrow 5.1\%}$	$50.1^{\uparrow 4.4\%}$	$45.1^{\uparrow 7.8\%}$	$\mathbf{52.4}^{\uparrow 6.7\%}$	$\mathbf{45.6}^{\uparrow 8.3\%}$

Table 5. Labels in zero-shot datasets: Concepts overlapping with the training corpus (seen) and concepts not overlapping (unseen).

Dataset	Seen Labels	Unseen Labels
BroadTwitter	Location, Organization, Person	-
HarveyNER	-	Point, Area, Road, River
AI	Product, Country, Person, Organization, Location, Miscellaneous	Field, Task, Algorithm, Researcher, Metric, University, ProgrammingLanguage, Conference
Literature	Event, Person, Location, Organization, Country, Miscellaneous	Book, Writer, Award, Poem, Magazine, LiteraryGenre
Music	Event, Country, Location, Organization, Person, Miscellaneous	MusicGenre, Song, Band, Album, MusicalArtist, MusicalInstrument, Award
Politics	Person, Organization, Location, Election, Event, Country, Miscellaneous	Politician, PoliticalParty
Science	Person, Organization, Country, Location, ChemicalElement, ChemicalCompound, Event, Miscellaneous	Scientist, University, Discipline, Enzyme, Protein, AstronomicalObject, AcademicJournal, Theory, Award
FabNER	Biomedical	Material, ManufacturingProcess, MachineEquipment, Application, EngineeringFeatures, MechanicalProperties, ProcessCharacterization, ProcessParameters, EnablingTechnology, ConceptPrinciples, ManufacturingStandards
Movie	Year	Actor, Character, Director, Genre, Plot, Rating, RatingsAverage, Review, Song, Tittle, Trailer
Restaurants	Location, Price, Hours	Rating, Amenity, RestaurantName, Dish, Cuisine
E3C	ClinicalEntity	-
MultiNERD	Person, Location, Organization, Biological, Disease, Event, Time, Vehicle	Animal, Celestial, Food, Instrument, Media, Plant, Mythological
$\text{WikiEvents}_{\text{NER}}$	CommercialProduct, Facility, GPE, Location, MedicalHealthIssue, Money, Organization, Person, JobTitle, Numeric, Vehicle, Weapon	Abstract, BodyPart, Information, SideOfConflict
SciERC	PartOf	UsedFor, HyponymOf, Conjuction, FeatureOf, Compare, EvaluateFor
SemEval	ComponentWhole	CauseEffect, InstrumentAgency, ProductProducer, ContentContainer, EntityOrigin, EntityDestination, MemberCollection, MessageTopic
ADECorpus	-	AdverseEffect
CoNLL 2004	LocatedIn, WorkFor, OrganizationBasedIn, LiveIn, Kill	-
NYT11-HRL	FoundedBy, NeighborhoodOf, WorkFor, Founded, PlaceLived, PlaceOfBirth, PlaceOfDeath, Children, Nationality	CaptialOfCountry, LocationContains, AdministrativeDivisionsOfCountry, CountryOfAdministrativeDivisions,
KBP37	FoundedBy, MemberOf, EmployeeOf, StateOrProvinceOfResidence, CountryOfResidence, CityOfResidence, Subsidiary, AlternateName, Origin, TitleOfPerson, Spouse, CountryOfBirth, TopMembersEmployees, CityOfHeadquarters, StateOrProvinceOfHeadquarters, CountryOfHeadquarters	-
GIDS	GraduatedFrom, PlaceOfBirth, PlaceOfDeath	HasDegree
$\text{WikiEvents}_{\text{EE}}$	ConflictEvent, ContactEvent, GenericCrimeEvent, JusticeEvent, MedicalEvent, MovementTransportEvent, PersonnelEvent, TransactionEvent	ArtifactExistanceEvent, CognitiveEvent, ControlEvent, DisasterEvent, LifeEvent
CASIE_{EE}	-	DatabreachAttack, PhisingAttack, RansomAttack, VulnerabilityDiscover, VulnerabilityPatch
PHEE_{EE}	-	PotentialTherapeutic, Adverse
$\text{WikiEvents}_{\text{EAE}}$	killer, victim, place, target, agent, instrument, participants, object, entity, destination, injurer, topic, perpetrator, defendant, judge_court, artifact, detainee, communicator, patient, treater, crime, prosecutor, jailer, giver, recipient, position, employee, organization, transporter, vehicle, origin, investigator, demonstrator	explosive_device, passenger_artifact, body_part, adquired_entity, receptor, payment_barter, impeder
$\text{CASIE}_{\text{EAE}}$	victim, attacker, place, time, price	trusted_entity, tool, purpose, compromised_data, number_of_victim, number_of_data, releaser, vulnerability, patch, pattern, attack_pattern, discoverer, used_for, system_owner, issues_addressed, vulnerable_system, supported_platform, patch_number, system_version, cve, damage_amount, payment_method
PHEE_{EAE}	subject, subject_age	treatment_route, treatment_dosage, treatment, treatment_drug, effect, treatment_disorder, subject_population, treatment_freq, treatment_duration, subject_gender, combination_drug, treatment_time_elapsed, subject_disorder, subject_race