IARPA HIATUS HRS foreground data

12 December 2025

As stated on IARPA's website[1], "The HIATUS program aims to develop novel human-useable systems for attributing authorship and protecting author privacy. Authorship attribution capabilities address many Intelligence Community (IC) needs, including combating sophisticated malicious information campaigns online and identifying counterintelligence risks. Authorship privacy capabilities protect authors whose writing, if attributed, could place them in danger."

IARPA HIATUS data consists of plain-text documents in English, Arabic, Chinese, and Russian, collected for development and testing of software to assist with authorship attribution and authorship privacy. The documents are grouped into data sets which we refer to as **"collections"**. All documents in a given collection are written in the same language, are of similar length, and have topical or other textual similarities; for example, one genre might be a collection of English language news articles 500 words or longer; another might be a collection of Chinese-language social media posts between 30 and 160 characters long.

Two kinds of documents were collected for HIATUS: **"background" documents** were scraped from the Internet, while **"foreground" documents** were authored for the program by subjects in a research study titled "Writing Styles Around the Globe". Not all collections contain background documents, but when a collection does contain them, foreground document authors were asked to write in a way that would resemble the background documents.

HIATUS collections are divided into two groups: **HIATUS Resource Sets (HRS)** and **HIATUS Test Sets (HTS)**. HRS data was released to HIATUS performer teams to be used for development, analysis, and testing of their systems. HTS data has been held back from performer teams and used to evaluate the systems they developed. All HRS collections contain both foreground and background documents.

The present data set includes only foreground documents from HRS collections.

Composition and structure of the data

The data is structured as a collection of JSON Lines files, one file per collection, in which each record represents a document and its associated metadata. The table below lists all HRS collections. Note that Arabic foreground documents have not yet been collected.

[1] https://www.iarpa.gov/research-programs/hiatus

Collection number	Language	Type of documents in collection	Length category	Source of background documents[2]
HRS1.1	English	Board game reviews	Long	Boardgamegeek.com
HRS1.2	English	Instructions	Long	Instructables.com
HRS1.3	English	Citizen journalism	Long	Globalvoices.org
HRS1.4	English	Literature	Long	Several literature-related forums from stackexchange.com (via StackExchange data dump on archive.org)
HRS1.5	English	STEM texts	Long	Several science-related forums from stackexchange.com (via StackExchange data dump on archive.org)
HRS2.1	English	Pet forum posts	Medium	Reddit.com forums r/Pets, r/PetAdvice
HRS2.2	English	Anecdotes about work	Medium	27 Reddit.com forums, including reddit.com/r/CoworkerStories, reddit.com/r/IQuit, reddit.com/TalesFromRetail
HRS2.3	English	Product descriptions	Medium	"Amazon Product Details" dataset from Kaggle.com
HRS2.4	English	Obituaries	Medium	Websites for eight funeral homes from the United States and Canada
HRS2.5	English	Movie and TV reviews	Medium	Letterboxd.com
HRS2.101	Russian	Citizen journalism	Long	Publizist.ru
HRS2.102	Russian	Do-it-yourself instructions	Long	sdelay.tv, diy.ru
HRS2.103	Russian	Movie and TV recommendations	Long	Film.ru, irecommend.ru
HRS2.104	Russian	Pet advice	Medium	Pesikot.org, animals.moe-online.ru
HRS2.105	Russian	Humorous personal anecdotes	Medium	Anekdotov.net, pikabu.ru, anekdot.ru
HRS3.1	English	For free ads	Short	Trashnothing.com
HRS3.2	English	Poetry	Short	Reddit forums r/ocpoetry

[2] Sources of background documents are provided only to indicate the type of document foreground document authors were trying to create. None of the data described in this document is web-scraped.

Collection number	Language	Type of documents in collection	Length category	Source of background documents[2]
HRS3.3	English	StackExchange comments	Short	StackExchange data dump from archive.org
HRS3.4	English	Wikipedia edit summaries	Short	Wikipedia.org
HRS3.5	English	Music album reviews	Short	Albumoftheyear.org
HRS3.201[3]	*Arabic*	*Animal care forum posts*	*Medium*	*2zoo.com*
HRS3.202	*Arabic*	*For sale ads*	*Short*	*Q84sale.com, almobawabah.com*
HRS3.203	*Arabic*	*News articles*	*Long*	*Youm7.com*
HRS3.204	*Arabic*	*Soccer news stories*	*Medium*	*Tunisia-sat.com, al2la.com*
HRS3.205	*Arabic*	*Responses to questions*	*Short*	*Io.hsoub.com forums AskIO and Advice*
HRS3.301	Chinese	Movie reviews	Long	Funscreen.tfai.org.tw, filmcritics.org.hk, 130q.com
HRS3.302	Chinese	Citizen journalism	Medium	Peopo.org
HRS3.303	Chinese	Pet adoption ads	Short	Meetpets.org.tw, bagong.cn, bobocw.com, rcyzgf.com
HRS3.304	Chinese	Test preparation	Medium	Chasedream.com
HRS3.305	Chinese	Hotel reviews	Short	Gckzw.com

Document length categories are defined below. Length was determined before a document underwent removal of personally identifying information (PII); PII removal often caused a small reduction in the overall word count of a document.

Length category	Language	Length range
Short	Arabic	15-90 words
	Chinese	30-160 characters
	English	20-100 words
	Russian	N/A
Medium	Arabic	90-300 words
	Chinese	160-560 characters
	English	100-350 words
	Russian	81-284 words
Long	Arabic	300+ words
	Chinese	560+ characters
	English	350+ words
	Russian	284+ words

[3] Arabic HRS foreground documents have not yet been collected.

The table below describes the structure of records in the dataset.

Field	Category	Format	Example	Required	Notes
documentID	identifier	Unique UUID	“40688002-0fce-4446-a50c-c3030514d638”	Yes	
authorIDs	label	List of unique UUIDs	[“adcf2d30-9973-4fea-9daf-f8695c1022b1”]	No (Required for foreground documents but not for background documents where author is not known)	List to account for multiple authorship. Can be empty list when unknown for background documents.
fullText	text	String	"This is an example."	Yes	
spanAttribution	label	JSON	[{“authorID”:“adcf2d30-9973-4fea-9daf-f8695c1022b1”, “start”: 0, “end”: 500}, {“authorID”:“6973952f-8bc7-4109-b3df-0d2f81de9b2c”, “start”:501,	No (Missing only when author is unknown. If there is a single author, include span comprising the whole document)	Character level spans attributed to each author

Field	Category	Format	Example	Required	Notes
			“end”:750}, [{“authorID”:“adcf2d30-9973-4fea-9daf-f8695c1022b1”, “start”: 751, “end”: 1000}]		
IsForeground	provenance	Boolean	False	Yes	
machineAuthored	provenance	Boolean	False	No (Required for needle documents. Missing for haystack documents)	
dateCollected	provenance	ISO 8601 date	"2022-08-01"	Yes	Date collected by ARLIS
publiclyAvailable	provenance	Boolean	True	Yes	
collectionNum	provenance	String	"HRS1.1"	Yes	
source	provenance	String	“boardgamegeek.com”	Yes	Where the data was collected from
deidentified	provenance	Boolean	True	Yes	Should always be True
languages	text derivative	List of strings	["en"]	Yes	ISO 639-1 codes

Field	Category	Format	Example	Required	Notes
lengthWords	text derivative	Integer	4	Yes	Number of words in the document (counted prior to PII removal)
dateCreated	metadata	ISO 8601 date	"2021-10-07"	No (Omitted when not available)	Date published, posted, or created where available.
timeCreated	metadata	ISO 8601 time	“12:07:22”	No (Omitted when not available)	Time published, posted, or created where available. Omitted when not available.

sourceSpecific	metadata	JSON	{
			"forumID": "1000074"
			}
				No (Omitted when no source specific fields)	Any source specific metadata that is available excluding PII fields

Below is a description of fields that appear within the sourceSpecific field, by collection and document type:

Collection and document type	Field name	Description
All foreground	participantID	ID assigned to the author of the document by ARLIS

	originalFile	Name given to the file by its original author
	processedFile	Name of the text file generated from the original file
Phase 1 foreground	robertaProbability	A measure of the likelihood that the document was authored by a generative AI model, based on RoBERTa (https://huggingface.co/docs/transformers/en/model_doc/roberta)
	luarProbability	A measure of the likelihood that the document was authored by a generative AI model, based on LUAR (https://github.com/LLNL/LUAR)

Collection process

Document collection procedures were reviewed and approved by the University of Maryland Institutional Review Board (protocols 1969504, 1986111, 2091700).

Phase 1

For phase 1, ARLIS recruited English-speaking participants, including both native and non-native speakers of English, to author documents in the same style and on the same topics as the background documents. Recruited authors were allowed to select the collection(s) they authored documents for and were asked to write between one and eight documents per collection. They submitted their documents in Microsoft Word format. Documents were reviewed by ARLIS personnel and had to be on topic and at least 340 words long. Documents suspected of containing plagiarized content or of being authored by generative AI were excluded from the dataset.

Phase 2

For phase 2, ARLIS recruited individuals who spoke either English or Russian (or, in a few cases, both languages), natively or non-natively, to author documents in the same style and on the same topics as the background documents. Authors were assigned specific collections to write in and were assigned to write between four and eight documents per collection. Documents were authored in a web-based data collection platform called DOCENT that was developed at ARLIS for the HIATUS program. The platform stores text as HTML. In addition to storing text, DOCENT also captures authors' typing patterns. This data is used to assess the likelihood that documents were legitimately authored in situ in DOCENT rather than copy-pasted or retyped from another source (e.g., from a generative AI tool). Authors were made aware that data about their typing would be recorded and gave their consent.

Documents were reviewed by ARLIS personnel and had to meet topic and length requirements. All English documents collected in phase 2 were medium-length documents (100-350 words). Russian documents were either medium-length (81-284 words) or long (284+ words), depending on the collection.

Phase 3

For phase 3, ARLIS recruited individuals who spoke one or more of the following languages: English, Arabic, and Chinese. As in phase 2, authors were assigned specific collections to write in. Authors were assigned a specific number of documents to write per collection; for long and medium-length collections, this number was between four and eight, whereas for short collections it was between six and eight. Also as in phase 2, documents were authored in DOCENT and typing data was collected, with the authors' consent.

Documents were reviewed by ARLIS personnel for topic and length. All English documents collected in phase 3 were short documents (20-100 words). Arabic and Chinese documents could be short (15-90 words for Arabic, 30-160 characters for Chinese), medium-length (90-300 words for Arabic, 160-560 characters for Chinese), or long (300+ words for Arabic, 560+ characters for Chinese), depending on the collection.

Document processing

All documents were converted to plain text. For HTML documents, this was done using the Python Inscriptis library (https://pypi.org/project/inscriptis/). Microsoft Word documents were converted to HTML using the Python Mammoth library (https://pypi.org/project/mammoth/), then to plain text using Inscriptis.

After conversion to plain text, curly quotation marks and apostrophes were replaced with equivalent "straight quotes", and contiguous substrings of whitespace characters were replaced with single whitespace characters (a linefeed [U+000A] if the substring contained one or more line breaks, such as a linefeed or carriage return, or a space [U+0020] otherwise).

Proprietary tools provided by Lawrence Livermore National Laboratory (LLNL) were used to evaluate the likelihood that each foreground document was produced by a generative AI language model. The tool developed for phase 1 assigned two scores to each document, one based on RoBERTa (https://huggingface.co/docs/transformers/en/model_doc/roberta) and the other based on LUAR (https://github.com/LLNL/LUAR). These scores were included in the source-specific metadata of each phase 1 foreground document. Documents with a RoBERTa-based score greater than or equal to 0.8 were excluded from the dataset. For phases 2 and 3, outputs from LLNL tools were combined with output from tools developed at ARLIS to evaluate the likelihood that texts were machine generated based on text content and typing data collected by DOCENT. Documents determined to be likely machine generated were discarded.

Personally identifying information (PII) was automatically removed from both background and foreground documents using Microsoft Presidio (https://microsoft.github.io/presidio/); specifically, we redacted names of people, phone numbers, email addresses, and IP addresses. We supplemented detection of phone numbers with custom code using regular expressions; phone numbers found in this way were manually removed.

Note: if you discover personally identifying information in this data, kindly report it to hiatus_data@umd.edu.