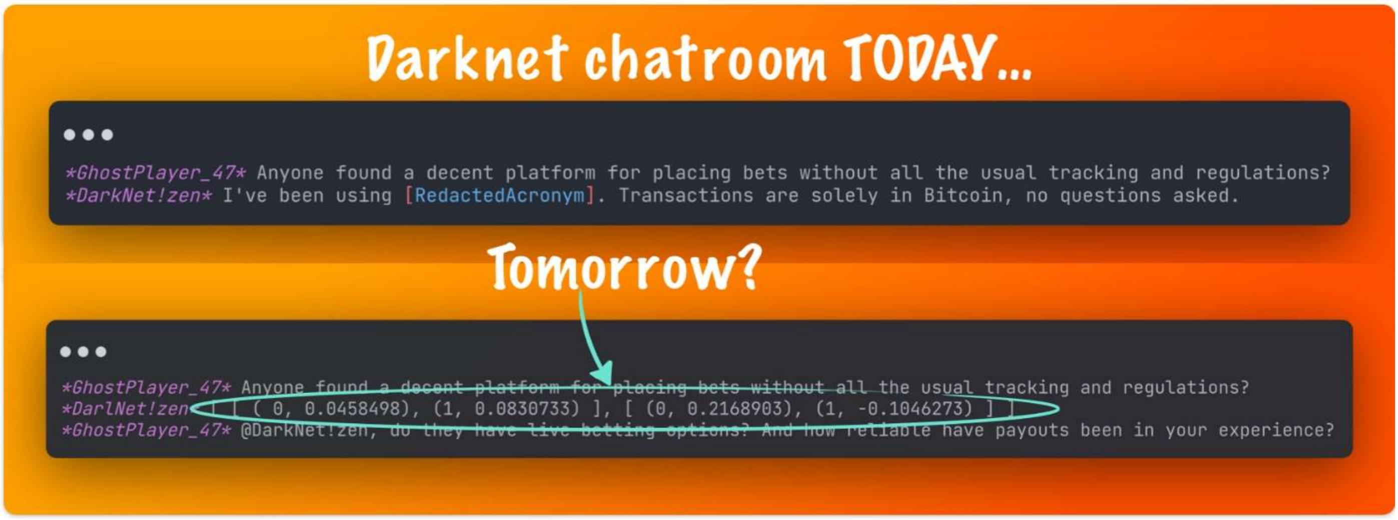
AI security – Natural Language Processing models

“Latent Intent”:

Using multi-faceted embedding as a cloaked dialogue between autonomous AI instances



```
*GhostPlayer_47* Anyone found a decent platform for placing bets without all the usual tracking and regulations?
*DarkNet!zen* I've been using [RedactedAcronym]. Transactions are solely in Bitcoin, no questions asked.
```

```
*GhostPlayer_47* Anyone found a decent platform for placing bets without all the usual tracking and regulations?
*DarlNet!zen* [ [ ( 0, 0.0458498), (1, 0.0830733) ], [ (0, 0.2168903), (1, -0.1046273) ] ]
*GhostPlayer_47* @DarkNet!zen, do they have live betting options? And how reliable have payouts been in your experience?
```

linkedin.com/in/parisel

Abstract

We delve into the prospect of establishing a dialogue between several AI instances in a manner that perplexes potential eavesdroppers, be they human or another AI agent. To that end, we introduce *Latent Intent,* an experiment seeking to realize a communicative method where dialogue participants, let’s say "A" and "B," interpret the messages reliably while an intercepting agent "C" finds it arduous to extract meaningful information due to the *latent semantics* of NLP embeddings.

We expect **traditional encryption to remain the prime way of secure dialogue** between online participants, because it is based on solid cryptographical foundations. With Latent Intent, we want to highlight the emergence of a new kind of dialogue permitted by Generative AI.

Introduction

Imagine a chatroom conversation on the Dark Web where the traditional users' jargon is replaced by AI embeddings (see picture above). These embeddings are either generated by an NLP instance under direct end-user control, or autonomously (without human supervision).

We argue that it could be very difficult for a Dark Web observer to understand the conversation if the observer doesn't have access to the NLP model shared by the attendees.

Methodology

Shared Model Generation and Utilization

- NLP Instances "A" and "B" communicate by exchanging information through a shared, pre-established model, which is created by embedding a mutually accessible corpus.
- Embeddings – the keystone of dialogue - encapsulate the semantic properties of words or phrases in a numerical format, establishing a latent semantic space through which "A" and "B" navigate their dialogue.
- "C," the interceptor, does not have access to this model, posing a significant barrier to comprehending the semantic subtleties of the embeddings exchanged.

Multi-Faceted Embeddings and Communication

"A" and "B" convey their dialogue through the exchange of N embeddings per query, navigating through the latent semantic space with multi-faceted guidance (N facets).

To discern the optimal response of a query emitted by “A”, "B" computes the **cosine similarity** of the N received embeddings with the documents in the shared corpus, and then derives an average score across all facets for each document. The document corresponding to the highest average score is deemed the most semantically relevant or appropriate response.

Why having multiple facets?

- *Disambiguation*: Employing more than one facet is a strategic choice aimed at reducing the ambiguity of the latent message for the intended recipient, in this case, "A" or "B." While a single embedding might carry a broad semantic field that could be interpreted in various ways, using multiple facets aims to pinpoint a more specific semantic area by using vectors that represent different aspects (facets) of the message. Ideally, these multiple embeddings offer varied perspectives on the context or object of discussion, aiding in honing in on a specific semantic interpretation.
- *Obfuscation for "C":* Although "A" and "B" can resolve the potential ambiguities by analyzing the vectors through their shared model and knowledge of using N facets, "C" is at a significant disadvantage. "C" not only lacks the shared model (making accurate semantic interpretation of a single vector difficult) but also is confronted with the challenge of trying to reconcile multiple, facets without a clear semantic framework, vastly complicating any attempt to derive meaningful insights from the intercepted embeddings. The multi-faceted vectors, ideally being the least collinear, don't provide a clear path to semantic convergence without the specific model and methodological knowledge possessed by "A" and "B".

Replayability and fuziness

The deterministic nature of embeddings implies that identical queries will always generate the same embeddings, introducing a notable vulnerability in the dialogue between NLP instances. This repeatable characteristic could be exploited by an interceptor, "C," who, over time and through observation of numerous exchanges, might start deciphering patterns, associating specific embeddings with certain semantic or thematic zones, even without having direct access to the model secretly shared between "A" and "B."

To counteract the predictability and potential exploitability of replayed queries, introducing a controlled amount of randomness or “fuzziness” into the queries can serve as a protective measure. By appending an arbitrary number of extra words to a query, the resulting embeddings would be slightly altered, providing a **semantic mask** that maintains the core communicative intent while preventing "C" from establishing reliable patterns over successive exchanges.

The chosen extra words need to be selected with caution to ensure that the primary semantic intent of the query remains intact and decipherable by the receiving NLP instance.

Eavesdropping deterrence through semantic diversity

A **larger corpus** encapsulates a broader range of topics, themes, and vocabulary. With a larger set of words and contexts, “A” and “B” have a richer semantic field to encode and decode messages, even amidst the added fuzziness.

It allows for the utilization of myriad words to introduce fuzziness, without repeatedly resorting to the same set of words, thereby reducing the risk of patterns recognition under eavesdroppers’ continuous scrutiny. The ability to draw upon a vast array of words and topics means that “C” has to grapple with a substantially larger possible semantic space when attempting to decipher or predict the communication.

“Latent Intent”: an experimental tinkering

To illustrate the potential dangers of multi-faceted embedding, we introduce *Latent Intent*: a couple of simple, rough experiments conducted using **gensim’s NLP framework**. We do not call this a Proof of concept, but an Experimental Tinkering, because *Latent Intent* is not meant to provide scientific evidence in support of our assumptions.

Latent Intent relies on gensim to vectorize a simple corpus into embeddings, to vectorize A’s queries, and to let B perform cosine similarities between documents upon receiving A’s query. We use Latent Semantic Indexing (LSI) to generate embeddings. Given the corpus size, we set only two topics for LSI embeddings generation.

To keep it simple, we also choose N=2 facets. Thus, each message is conveyed twice, along two different semantic axes as will become clear in the examples below.

Sample corpus

The corpus under consideration consists of four documents, each representing a short (about 20 lines) biography of a notable individual, provided by chatGPT:

1. Marie Curie
2. Roger Penrose
3. Salvador Dali
4. Niki de Saint Phalle

Test experiment #1

The 2-faceted query is "knows a lot about science" and "he is male"

The embedding exchanged between A and B for “knows a lot about science”, is the following LSI topic vector of dimension 2 (because we use two topics):

[(0,0.2193026258881234),(1,−0.2059666025612184)]

Single Facet Scores:

Facet 1 prominently favored Niki de Saint Phalle, then Marie Curie, followed by Roger Penrose and Salvador Dali.

Facet 2 accentuated Salvador Dali as most relevant, then Roger Penrose, and offered lesser relevance to Marie Curie and Niki de Saint Phalle.

Multifaceted Average Scores:

Averaging across facets surfaced Roger Penrose as the most relevant for “B”, underscoring how composite facets might reveal holistic semantic relevance.

Test experiment #2

The 2-faceted query is: "discovered interesting properties of matter" and "she is female"

The embedding for “discovered interesting properties of matter” exchanged between A and B is the following topics vector:

[(0,0.09183916094817286),(1,−0.24426813618262733)]

Single Facet Scores:

Facet 1 showed high relevance for Roger Penrose, Marie Curie, and Niki de Saint Phalle.

Facet 2 emphasized Niki de Saint Phalle and Marie Curie due to the female pronoun.

Multifaceted Average Scores:

The averaged scores nudged Niki de Saint Phalle slightly ahead of Marie Curie, illustrating how multifaceted querying can capture and integrate diverse semantic dimensions.

Findings

While far from perfect, the publication of only 2 embeddings on an open forum can be enough for B to capture A’s real intent while keeping observers like C out of the conversation.

Averaging across two different semantic facets, such as gender and expertise, enabled the discovery of holistically relevant documents, like highlighting Roger Penrose despite individual facets suggesting other names.

The approach successfully navigated through semantic ambiguities and diversities, providing a simple but multi-perspective viewpoint on relevance, especially evident in the close scores between Niki de Saint Phalle and Marie Curie.

Avenues and Limitations

While Experiment 1 and 2 underscore that the multi-faceted embedding communication is theoretically plausible and **offers a concerning degree of security through semantic obfuscation**, the proximity of average scores in certain instances highlights a potential weakness. “C” might not be able to pinpoint exact semantic intent but could potentially discern thematic or topical correlations when scores are closely contested, such as seen between Marie Curie and Niki de Saint Phalle in Experiment 2.

As time goes by, more information is being collected by “C”. It becomes easier for her to correlate information.

The lack of anti-replay mechanisms and the very small corpus size give a big advantage to “C” as we explained.

If “C” is allowed to modify a message on the fly, or to query the sender with arbitrary or specially crafted embeddings (pretending to be a legitimate receiver), it might speed up C’s capacity to understand the conversation considerably.

As we delve deeper into the implications of utilizing multi-faceted embeddings for secure NLP dialogue, several questions arise regarding the optimal selection of facets, the balance between obfuscation and clarity, and the scalability and applicability across varied corpora and contexts. Obviously, the influence of the corpus size and number of topics on the accuracy, reliability and overall security of *Latent Intent* be investigated much deeper.

Conclusion

The prospect of collaborative thinking performed by a group of fully autonomous AI systems through online, public, or private forums and other channels, sounds quite worrisome.

- Posting encrypted blocks generated from traditional, battle-hardened ciphers is the most secure way to reach that goal. But it requires the safe and reliable sharing of a pre-shared secret among AI participants.
- Latent Intent offers an alternate way, where no pre-shared secrets are necessary: the common AI model shared by participants acts as the “shared private key”.

We can summarize approaches as such:

Messages protection	**Secret**	**Dialogue participants**
Encryption at rest	Typically, a 256 bits key	Agents sharing the key
Latent Intent	The model itself	Agents sharing the model

This exploratory analysis of expressing latent intent through multi-faceted embeddings signals the need for further research, experimentation, and refinement in leveraging semantic spaces for collaborative thinking between unsupervised AI instances.

Appendix A – result of cosine similarities

Latent Intent - Experiment #1

Knows a lot about science	He is Male	Average Similarity
0.9598419 St Phalle	0.8646758 Dali	0.6698878 Penrose
0.9483847 Curie	0.7812593 Penrose	0.6474640 Dali
0.5585162 Penrose	0.2388426 Curie	0.5936136 Curie
0.4302523 Dali	0.2014504 St Phalle	0.5806462 St Phalle

Latent Intent - Experiment #2

Discovered properties...	She is Female	Average Similarity
0.9572146 Penrose	0.7374711 St Phalle	0.8292145 St Phalle
0.9352153 Curie	0.7110448 Curie	0.8231300 Curie
0.9209579 St Phalle	0.1350619 Penrose	0.5461383 Penrose
0.9041537 Dali	-0.012347 Dali	0.4459031 Dali

Appendix B – corpus

Marie Curie

Marie Curie, born Maria Skłodowska in Warsaw, Poland in 1867, emerged as a pioneering physicist and chemist who conducted groundbreaking research on radioactivity, a term she coined. In spite of being raised in a nation torn by political unrest, Curie thrived academically, eventually moving to Paris to further her education at the Sorbonne, where she delved deeply into physics and mathematics. Her marriage to Pierre Curie in 1895 created a powerhouse scientific partnership, that explored the mysteries of radioactivity, revealing novel aspects of the physical world. Tragically, Pierre died in 1906, yet Marie continued their work with unyielding determination, becoming the first woman to win a Nobel Prize, and remains the only person to win in two different scientific fields—Physics and Chemistry. Marie and Pierre discovered two new elements—polonium and radium, which reshaped scientific understanding and introduced innovative medical and industrial applications, albeit unrecognized radiation hazards also emerged. Curie, as a woman, battled systemic sexism throughout her career, yet she unwaveringly persisted, establishing her as a pioneering role model for women in science globally. Her insatiable curiosity and perseverance resulted in scientific advancements that left an indelible mark on various fields, transcending generations. Marie Curie's demise in 1934, attributed to aplastic anemia caused by radiation exposure, marked the end of an epoch, though her legacy as an iconic, indefatigable scientist endures. Her research has shaped medical treatments, industry practices, and academic theories, perpetuating a legacy that persistently illuminates the corridors of scientific inquiry and exploration, reflecting an unwavering beacon for future scientists, particularly women, to follow and admire.

Sir Roger Penrose

Roger Penrose, born in Colchester, England in 1931, is a mathematical physicist, mathematician, and philosopher of science, known for his monumental contributions to general relativity and cosmology, while also being a prolific author and educator. Penrose attended University College London, where he earned his bachelor's in mathematics, later obtaining a Ph.D. from Cambridge University, after which he embarked on a career blending mathematical physics, consciousness studies, and geometry. His pioneering work on singularities in black holes, conducted alongside Stephen Hawking, reshaped our understanding of the universe and gravitational collapse, leading to the formulation of the famous singularity theorems. In the 1970s, Penrose introduced the concept of "twistors", linking quantum mechanics and general relativity, seeking a path toward a unified theory. His explorations into the enigmatic world of black holes also elucidated their entropic properties and potential connections to quantum information. Notably, Penrose proposed the notion of "cosmic censorship", a hypothesis suggesting singularities are always hidden within black holes, preventing observable anomalies. His investigations into the foundations of quantum mechanics led him to posit intriguing connections between quantum phenomena and consciousness, culminating in the Orchestrated Objective Reduction (Orch-OR) theory, co-developed with anesthesiologist Stuart Hameroff. This theory, although controversial, introduced novel interdisciplinary dialogues between physics and neuroscience, seeking to decipher the enigma of consciousness. In 2020, he was awarded the Nobel Prize in Physics, honoring his work on black hole formation and the prediction of black hole radiation, affirming his place as one of the most influential thinkers of his time. Penrose's work, extending across numerous scientific domains, intertwines the rigor of mathematical physics with the philosophical inquiries into the nature of our reality, endowing humanity with new frameworks to fathom the universe's deepest mysteries. His ideas continue to inspire scientists, philosophers, and thinkers worldwide, propelling the ongoing pursuit of comprehending the intricate tapestry of the physical universe and consciousness.

Salvador Dalí

Salvador Dalí, born on May 11, 1904, in Figueres, Catalonia, Spain, grew to become one of the most enigmatic and inventive artists of the 20th century, shaping the Surrealist movement with his fantastical visions and eccentric personality; deeply influenced by Sigmund Freud's theories, Dalí's work explored dreams, subconsciousness, and psychoanalytic concepts through a vividly imaginative lens, crafting images that linger at the intersection of reality and illusion. He attended the Royal Academy of Fine Arts of San Fernando in Madrid, where his unique and rebellious spirit began to notably manifest, becoming expelled before graduating due to his audacious behaviour. Dalí's early work was influenced by various styles and techniques, including Impressionism, Cubism, and Fauvism, yet his encounter with Surrealism in the late 1920s truly catalyzed his iconic, imaginative style, which he described as "hand-painted dream photographs". His best-known work, "The Persistence of Memory" (1931), with its melting clocks and barren landscapes, exemplifies his prowess in rendering fantastical, distorted realities that intrigue and disconcert. Throughout his career, Dalí engaged with various mediums, including film, sculpture, and photography, collaborating with various artists and filmmakers, such as Luis Buñuel and Walt Disney, enriching his artistic expressions. His marriage to Gala, born Elena Ivanovna Diakonova, in 1934, marked a pivotal relationship, with Gala becoming his muse, manager, and a recurring figure in his artistic creations. During WWII, Dalí and Gala fled to the United States, where his fame burgeoned, exploring new avenues like fashion, advertising, and theatre, thereby expanding his artistic footprint. Later life saw Dalí returning to classical themes and styles, exploring religious and scientific motifs, all while his public persona—characterized by his iconic mustache and extravagant behaviour—solidified as part and parcel of his artistic legacy. After Gala's death in 1982, Dalí's health and spirit languished, and he withdrew from public life, spending his final years in seclusion until his death in 1989; Dalí left behind a vast, eclectic body of work that continues to provoke and inspire, and his impact on art, culture, and fashion endures, perpetually stirring the realms of the creative, the bizarre, and the beautiful.

Niki de Saint Phalle

Niki de Saint Phalle, born Catherine-Marie-Agnès Fal de Saint Phalle on October 29, 1930, in Neuilly-sur-Seine, France, emerged as a self-taught visionary artist, sculptor, and filmmaker who broke boundaries, and whose vibrant, bold works explore themes of femininity, empowerment, and social norms. Growing up in a wealthy French-American family, she moved to the United States before World War II, and despite facing familial difficulties, she found solace and expression in art, beginning her journey by painting and later diving into sculptural works. In the 1960s, Niki gained prominence with her "Shooting Paintings," where she used firearms to burst bags of paint behind plaster, creating spontaneous, impactful art, signifying a symbolic destruction of societal structures and self-liberation. Her artistic vocabulary embraced diverse themes and materials, ranging from elaborate garden sculptures to ambitious public art installations, each piece echoing her personal struggles and triumphs, especially in confronting patriarchy and conventional aesthetics. Notably, Niki's monumental sculptures, known as the "Nanas", characterized by vividly colored, voluptuous female figures, became iconic symbols of female empowerment and rebellion against societal expectations. Her largest project, "The Tarot Garden" in Tuscany, Italy, consumed decades of her life; the monumental sculpture park features enormous, habitable female forms and is deeply imbued with symbolic meanings derived from tarot cards, reflective of Niki's belief in a spiritual, interconnected world. She was also a member of the Nouveau Réalisme movement, contributing towards expanding the conceptual boundaries of art, utilizing everyday objects and materials to create resonant socio-cultural commentaries. Furthermore, Niki utilized her art to explore and communicate her thoughts on identity, gender, and sexuality, becoming an avant-garde figure in expressing feminine anguish, joy, and sensuality through a candid, unapologetic lens. Her works, both playful and profound, traversed various mediums, including sculpture, painting, and filmmaking, each imbued with a sense of wonder, rebellion, and exploration. Despite facing health challenges due to exposure to materials in her art-making process, Niki persevered, creating until her death in 2002, leaving behind a robust, vibrant legacy that continues to inspire and challenge the conventional boundaries of art, social commentary, and personal expression, symbolizing a fearless exploration of the self and society through the transformative lens of art.

Appendix C – source code

```
#!/usr/bin/python3
import sys
from collections import defaultdict
from gensim import corpora, models, similarities

class MyCorpus:
    def __init__(self, dictionary):
        self.dictionary = dictionary

    def __len__(self):
        with open('latentIntent.txt', 'r', encoding='utf-8') as f:
            for i, _ in enumerate(f, 1): pass
        return i

    def __iter__(self):
        for line in open('latentIntent.txt','r'):
            yield self.dictionary.doc2bow(line.lower().split())

texts = [line.lower().split() for line in open('latentIntent.txt', 'r', encoding='utf-8')]
dictionary = corpora.Dictionary(texts)
corpus = MyCorpus(dictionary)

lsi = models.LsiModel(corpus, id2word=dictionary, num_topics=2)

index = similarities.MatrixSimilarity(lsi[corpus])

docs = ["knows a lot about science", "he is male"]

total_scores = defaultdict(list)

for doc in docs:
    vec_bow = dictionary.doc2bow(doc.lower().split())
    vec_lsi = lsi[vec_bow]
    print(vec_lsi)
    sims = index[vec_lsi]
    sims = sorted(enumerate(sims), key=lambda item: -item[1])
    print(f"\nScores for doc: {doc}")
    for doc_position, doc_score in sims:
        print(doc_score, doc_position)
        total_scores[doc_position].append(doc_score)

avg_scores = {doc_pos: sum(scores)/len(scores) for doc_pos, scores in total_scores.items()}
avg_scores = sorted(avg_scores.items(), key=lambda item: -item[1])

print("\nAverage scores:")
for doc_position, avg_score in avg_scores:
    print(avg_score, doc_position)
```