Title of the project

PARLA CHIARO (*Speak Clearly*) - Protecting Italian Dialect Speakers from AI-generated Health Misinformation

Summary (100 words)

Millions of Italians—especially elderly and rural residents—speak regional dialects like Neapolitan, Sicilian, or Roman as their primary language. Many have limited literacy in standard Italian, making voice-based AI their only accessible gateway to digital health information. Yet today's speech recognition and AI models, trained almost exclusively on standard Italian, systematically misunderstand these dialects. When someone asks a health question in dialect and the AI mishears it, the result can be dangerously wrong medical advice, with real consequences for vulnerable populations. PARLA CHIARO investigates how AI fails dialect speakers and develops protective safeguards to prevent health misinformation before it causes harm.

Objectives (100 words)

1. Build a High-Quality Dialect Health Dataset: Collect and annotate 100+ hours of spontaneous, health-related voice conversations across three major Italian dialects (e.g., Neapolitan, Sicilian, Roman).
2. Quantify Real-World Risks: Working with Italian medical experts, systematically test leading speech-to-text and large language models on dialectal health queries to measure misunderstanding rates and resulting misinformation severity.
3. Develop a Dialect-Aware Warning System (DAWS): Create an intelligent safeguard that detects unclear or dialect-influenced input and prompts users to clarify. This will prevent AI systems from confidently delivering wrong answers to misunderstood questions.

Expected outcomes (100 words)

1. Public Research Dataset: 100+ hours of annotated dialect speech released under open license, enabling inclusive AI research worldwide.
2. Evidence-Based Risk Report: Comprehensive study showing how dialectal speech degrades AI accuracy and generates health misinformation, targeting AI developers, healthcare organizations, and policymakers.
3. Open-Source Protection Tool: A Dialect-Aware Warning System (DAWS) that any AI platform can integrate to protect dialect speakers in real-time.

All outcomes will be submitted to premier international AI, information retrieval and machine learning venues (EMNLP, ACL, AAAI, SIGIR, ECIR, ECAI) and leading technical journals (TKDE, TPAMI) for peer-reviewed publication.

Project timeline and milestones (100 words)

Months 1–2: Establish ethics protocols and technical infrastructure; comprehensive literature review.

Months 3–5: Collect authentic dialect health conversations.

Months 4–7: Annotate data for transcription errors, AI misinterpretations, and health risk severity; quality assurance and preliminary analysis.

Months 5–9: Design and develop DAWS prototype.

Months 9–11: Rigorous evaluation of DAWS against baseline systems.

Month 12: Finalize documentation to disseminate findings through research publications, policy briefs, and open dataset release; host virtual workshop engaging AI developers, healthcare professionals, and policymakers to drive adoption of protective measures.

Partnerships or collaborators (100 words)

- We will partner with Accenture plc to pilot PARLA CHIARO's DAWS system within our existing intelligent medical anamnesis platform, already developed collaboratively.

- Professors Giovanni Esposito, Raffaele Izzo, and Raffaele Piccolo from the Department of Advanced Biomedical Sciences, Cardiology Section, University of Naples Federico II, will provide critical clinical validation. Their collaborations with our team on Oracle-funded and Accenture-sponsored research ensure our work addresses genuine medical challenges.

- Dr. Marco Postiglione (PhD, Northwestern University) brings expertise in healthcare natural language processing. His research with our group has appeared in premier AI conferences (SIGIR, ECAI, ECIR) and leading journals (TNNLS, J-BHI, TIST).

Budget details

Provide details regarding the Data Collection monetary grant needed and specify its usage (100 words)

We request $50,000 to support data collection.

Research Assistants ($25,000): Native dialect speakers with a background in Computer Science will handle collection and analysis of data and development of the DAWS system.

Recruitment & Incentives ($5,000): We will recruit 200+ speakers representing diverse ages and socioeconomic backgrounds. Each will participate for one hour, providing ~30 minutes of speech, compensated €10–15 plus platform fees (e.g., Prolific).

Annotation & Quality Control ($20,000): Expert native annotators will transcribe audios, label metadata and dialectal features as well as details about misinformation severity to ensure data accuracy and reliability.

Provide sponsorship amount (USD) and details regarding the Azure services needed to support your project, including but not limited to compute, storage, APIs, etc. * (100 words)

We request $15,000–$20,000 in Azure compute credits.

Compute: GPU-accelerated VMs (e.g., NC-series) will support fine-tuning and/or benchmarking of state-of-the-art speech-to-text and large language models on collected data.

Storage: Azure Blob Storage will securely host raw audio, metadata and annotations.

APIs & Services: Azure Speech-to-Text and Text Analytics APIs will provide baseline performance comparisons, while Azure Machine Learning will manage model training, validation, and deployment workflows.

Narrative

What specific problem, challenge, or need affecting underrepresented European languages does your project aim to address, and why is it important? Explain the linguistic gap or need (e.g. lack of data or tools for a particular language) and its significance. * (250 words)

Italian regional dialects represent a major blind spot in AI language technology. While standard Italian enjoys substantial NLP and ASR resources, millions of Italians rely primarily on dialects for everyday communication, especially older, rural, and socio-economically disadvantaged populations. According to Italy's national statistics institute, approximately 14% of adults use dialect as their primary language at home, and usage increases significantly among older and rural populations up to 32% (ISTAT, 2015). These dialects, such as Neapolitan, Sicilian, and Roman, are not simply accents but distinct linguistic systems with unique phonology, lexicon, and grammar, and some are formally recognized as minority languages (UNESCO *Atlas of the World's Languages in Danger*).

Yet no publicly available, high-quality speech datasets exist for spontaneous dialectal healthcare communication. Modern speech-to-text and LLM systems, trained almost exclusively on standard Italian, routinely misrecognize dialectal input. These errors cascade: incorrect transcription implies misinterpreted medical intent, that takes to unsafe or misleading health guidance. The risk is acute because elderly dialect speakers also exhibit lower digital and health literacy, and increasingly rely on voice-based interfaces as their primary access to online health information.

This project addresses a critical equity and safety gap: ensuring AI systems do not harm speakers of underrepresented European dialects in healthcare settings. By generating the first health-focused Italian dialect dataset and developing a dialect-aware clarification safeguard, PARLA CHIARO establishes a scalable framework for protecting linguistic minorities across Europe, where similar challenges affect speakers of regional varieties such as Bavarian, Occitan, and Galician.

How do you plan to engage with local language communities and stakeholders, and what partnerships or collaborations have you established, or will you establish? Describe how you will involve native speakers, NGOs, or other partners to ensure community input, cultural sensitivity, and adoption. (250 words)

Our project is grounded in direct engagement with the communities most affected by dialect-driven AI exclusion. We have established a core partnership with cardiologists at

Policlinico Federico II (Naples), a major public medical center serving large populations of dialect-speaking patients. These clinicians will validate health scenarios, advise on ethical considerations, and support participant outreach through community health networks.

To ensure cultural and linguistic authenticity, our research assistants will be native speakers of the targeted dialects. They will lead participant recruitment, facilitate recording sessions, and review protocols to ensure that dialects are represented respectfully as legitimate linguistic systems—not as deviations from standard Italian. This lived-experience perspective will guide the project end-to-end.

We will recruit participants through senior centers (centri anziani), local health clinics, and regional cultural associations. These organizations have long-standing trust within dialect communities and will help us ensure accessibility, transparency, and voluntary participation.

During months 1–2, we will conduct focus groups with 10–20 dialect speakers to co-design recording prompts and identify realistic healthcare scenarios. This participatory design phase guarantees that the dataset reflects authentic speech patterns and real concerns, rather than artificial or academic constructs.

All participants will provide informed consent, and we will request approval from the relevant university ethics review board before collection. At project completion, we will return value to the community through accessible presentations and make the Dialect-Aware Warning System freely available to healthcare organizations and community partners, promoting equitable adoption.

How will you ensure ethical data collection and usage throughout the project? Explain methods for obtaining informed consent, protecting privacy, and respecting cultural norms. Include any plans for data anonymization, bias mitigation, and compliance with relevant ethical guidelines. * (250 words)

We will enforce rigorous ethical and privacy safeguards throughout the project. Prior to data collection, we will obtain full ethical approval from our university's Institutional Review Board (IRB), ensuring compliance with EU GDPR, Italian privacy law (D.Lgs. 196/2003), and the Declaration of Helsinki for research involving human subjects. No recording will begin until approval is granted.

All participants will receive clear, accessible consent materials in Italian, including verbal explanations for those with limited literacy. Consent will explicitly cover: voice recording, annotation, anonymized research use, and optional contribution to an open dataset. Consent will be obtained both in writing and verbally recorded. Participants may withdraw at any point, without justification, and their data will be permanently removed.

We will not record personal medical histories. Personally identifiable information (names, addresses, institutions, sensitive health details) will not be collected or will be systematically removed. Audio recordings will be coded, stored securely, and stripped of introductions or identifiable context. Only approved project members will have access to raw data under role-based controls.

Cultural sensitivity is foundational. Native-speaker researchers will design and supervise collections to ensure dialects are treated as valued linguistic identities. Focus groups will identify topics that may be culturally or emotionally sensitive; such content will be excluded or handled with tailored protocols.

To mitigate bias, participants will be recruited across age, gender, and socio-economic backgrounds, and system performance will be evaluated across demographic subgroups. The final public dataset will include only anonymized, consent-approved speech samples.

What are the expected outcomes and impact of your project on the target language community or users, including any measurable goals or metrics? Outline the concrete results (e.g. datasets created, models developed) and how you will measure success (e.g. word error rate, number of users) *(250 words)

The main objective of this project is to deliver concrete, measurable advances for dialect-speaking communities and the broader European AI ecosystem. Our primary impact goal is to ensure that elderly and low-literacy dialect speakers can access AI-powered health information with safety and confidence comparable to standard Italian speakers.

Expected Deliverables:

1) Dialect Health Speech Dataset: ≥100 hours of spontaneous dialect speech from ≥200 speakers across three major dialects (Neapolitan, Sicilian, Romanesco), annotated for transcription quality, dialect features, and clinical risk factors. Released under Creative Commons license, this will be the first substantial open dataset of spontaneous Italian dialect speech in healthcare contexts.

2) Dialect-Aware Warning System (DAWS): open-source safeguard module that detects dialectal or ambiguous input and prompts the user for clarification, preventing downstream misinformation compared to standard systems.
3) Benchmark & Risk Report: empirical analysis of automatic speech recognition (ASR) performance on dialectal speech, plus systematic evaluation of downstream health misinformation risk when models misinterpret dialect queries.
4) Community Engagement Outputs: co-design workshop outcomes, public documentation of ethical protocols, and user-friendly dissemination materials shared with senior centers and healthcare partners.

Success will be demonstrated by achieving parity in safety and user confidence between dialect and standard-Italian speakers, shown through improved word error rate (WER) and high-risk-query detection with fewer unsafe outputs, and validated by measurable adoption indicators such as number of dataset downloads, number of DAWS forks/clones, and number of peer-reviewed publications in leading venues.

What potential risks or challenges do you anticipate in implementing your project, and how do you plan to mitigate them? Discuss challenges such as data scarcity, technical hurdles, or community adoption issues, and strategies to address them (e.g. risk management, partnerships). *(250 words)

Recruitment and Trust Barriers: elderly dialect speakers may have concerns about data use or limited digital access. To mitigate this, we will engage trusted intermediaries and offer in-person participation options. Native-speaker researchers will conduct outreach and explanation in dialect where appropriate, emphasizing participant autonomy and community benefit.

Data Quality and Linguistic Variation: spontaneous dialect speech naturally includes code-switching, background noise, and variation across speakers. Rather than excluding such cases, our annotation scheme will explicitly tag code-switching and dialectal intensity, turning variation into a research asset. Quality control will involve double-annotation of 5% of data with inter-annotator agreement targets (Cohen's $\kappa > 0.75$).

Technical Performance Limitations: ASR and LLM systems may struggle to produce meaningful outputs on dialectal input. This risk is intrinsic to the problem we are addressing. We will benchmark multiple commercial and open-source systems (e.g., Whisper, Google Speech-to-Text, Azure) to identify best-performing baselines; if performance is extremely poor, that outcome will underscore the urgency and impact of this work.

Adoption and Integration Friction: healthcare platforms may be hesitant to adopt a new safety layer. DAWS will therefore be designed as a lightweight, modular component that integrates via API with minimal engineering effort. Early feedback from medical partners and a final dissemination workshop will support usability and relevance.

Timeline Pressure: 12 months is ambitious for collection, annotation, and system development, so we have structured parallel workflows and built a 4-week buffer for contingencies.

What open-source resources will your project produce, and how will you make them publicly accessible? Specify the datasets, tools, or models to be released openly, and the platforms or licenses for sharing (e.g. GitHub, Creative Commons), aligning with LINGUA's open-data mandate (250 words)

PARLA CHIARO is committed to full open access and will release all project assets under permissive licenses to maximize scientific and community impact.

1. Dialect Health Speech Dataset

Content: 100+ hours of spontaneous, annotated audio from speakers of three major Italian dialects, with transcriptions, dialectal feature tags, ASR error labels, and clinical-risk annotations.

License: Creative Commons BY-SA 4.0 to support academic and commercial reuse with attribution.

Platform: Released on Hugging Face Datasets and Zenodo (with DOI) for long-term accessibility, accompanied by detailed documentation, annotation guidelines, and anonymized demographic metadata.

2. Dialect-Aware Warning System (DAWS)

Content: Complete source code, pre-trained detection models, API documentation, and integration guides for major LLM platforms (OpenAI, Anthropic, Google).

License: Apache 2.0 to encourage adoption in both industry and research settings.

Platform: GitHub repository with Docker containers for easy deployment, comprehensive tutorials, and example implementations.

3. Benchmark Suite

Content: Evaluation scripts, baseline results for major ASR/LLM systems on dialectal input, and standardized test sets for reproducible comparison.

License: MIT License.

Platform: GitHub with automated tests and reproducibility instructions.

4. Research Outputs

Peer-reviewed papers will be published as arXiv preprints. Policy briefs, workshop slides, and technical reports will be released under CC BY 4.0 on the project website. Documentation will be available in English and Italian, with dialect-accessible summaries to ensure transparency and community inclusion.

Why is your team or organization best suited to carry out this project? Highlight relevant experience in AI and language technology, prior work with language communities, and any expertise that will help ensure the project's success. * (250 words)

Our team uniquely combines deep technical AI expertise, clinical validation capacity, industry collaboration, and community ties, an essential foundation for delivering responsible innovation in linguistically underserved settings. Our group has delivered open datasets, peer-reviewed publications, and deployed AI tools, supported by the secure infrastructure of the University.

Prof. Vincenzo Moscato is a Full Professor at the Department of Electrical Engineering and Information Technology, University of Naples Federico II, where he leads the PICUS (Pattern and Intelligence Computation for mUltimedia Systems) group and serves as Scientific Coordinator of the Data Science National Laboratory (CINI). His expertise spans Big Data Analytics, Artificial Intelligence, Multimedia systems, Recommender Systems, and Social Network Analysis. He co-founded the academic spin-off DataJAM srl and won an international Oracle Award for knowledge graphs, directly relevant to this project's focus on AI safety and health information systems.

Our partnership with Prof. Giovanni Esposito, Prof. Raffaele Izzo, and Prof. Raffaele Piccolo (Department of Advanced Biomedical Sciences, Federico II University) provides critical medical expertise. These cardiologists regularly treat elderly dialect-speaking patients and understand how communication barriers affect health outcomes. Their collaboration with our team on health knowledge graphs demonstrates ability to bridge AI and clinical medicine.

Our research assistants will be native dialect speakers, ensuring culturally competent data collection and community trust, and we will also collaborate with industry partners in healthcare digital transformation, including Accenture, for deployment and validation in clinical environments focused on capturing doctor–patient dialogue nuances and supporting medical documentation generation.

Describe the language dataset(s) you will create/expand/release, including languages/coverage, data type and target size, collection and consent model (GDPR), open license and hosting plan, documentation (datasheets/data statements), annotation & QA, community governance, and sustainability.

These points are optional and intended as guidance to help structure your answer.

- **Languages & Coverage:** Languages/dialects/varieties, geographic coverage, speaker demographics (age, gender, region), and representativeness
- **Modality & Size:** Type(s) of data (e.g., speech, text), target volumes (e.g., hours/utterances/tokens), and quality targets (e.g., SNR/WER baselines, transcription conventions).
- **Collection & Consent:** Data sources (new collection vs. curation), recruitment and compensation of participants, informed consent model (incl. withdrawal/deletion rights), handling of sensitive/culturally restricted content, and GDPR compliance.
- **Licensing & Access:** Open license(s) you will apply (e.g., CC BY 4.0 for text; appropriate open licenses for audio/metadata), planned hosting and persistent identifiers (e.g., repository, DOI), and any access tiers needed to protect privacy or community protocols.
- **Documentation & Ethics:** Dataset documentation plan (e.g., Datasheets for Datasets and Data Statements for NLP), known limitations, risk assessment and bias/harms mitigation strategy, and a statement of community governance/benefit-sharing.
- **Annotation & QA:** Annotation schema, tools and guidelines, inter-annotator agreement targets, reviewer training, and quality control workflows.
- **Sustainability:** Maintenance plan (issue handling, updates, versioning cadence), responsible deprecation, and how you will resource long-term stewardship post-grant.

Dataset Specification, Openness & Stewardship * **(500 words)**

Languages & Coverage

We will create a high-quality speech corpus covering three major Italian regional dialects: Neapolitan, Sicilian, and Roman, especially spoken among older and rural citizens. Target demographics include adults aged 50–85+ with priority on 65+, balanced gender distribution, and a mix of urban/rural and socioeconomic backgrounds. We will recruit ≥200 participants, ensuring geographic variety within each dialect region (e.g., Naples vs. Caserta; Palermo vs. Catania).

Modality & Size

- Type: spontaneous audio recordings in health-related dialogue settings
- Target Volume: ≥100 hours of raw audio (≈30 minutes per participant), with ~85 hours expected post-quality filtering
- Quality Standards: minimum 16kHz sampling rate, SNR >20dB for most of data
- Content: participants will discuss realistic tasks (symptom descriptions, medication questions, appointment scheduling) with native dialect-speaking research assistants to ensure natural conversational flow

Collection & Consent

All data will be newly collected through in-person sessions, with remote options for accessibility.
Participants will receive €10–15/hour compensation plus platform fees.
Multi-layered process compliant with GDPR and Italian privacy law will include:

- Written consent forms in standard Italian and dialectal versions, reviewed for comprehension
- Verbal consent recorded at session start, confirming understanding
- Explicit opt-in for recording, annotation, public release, and research use
- Clear withdrawal rights: participants can request data deletion up to 30 days post-collection; after annotation begins, only their audio is removed while anonymized transcripts may remain
- Sensitive content protocol: participants can flag topics they're uncomfortable discussing; research assistants trained to redirect if distress detected
- Full IRB approval: we will obtain institutional ethics approval before collection

Licensing & Access

- Audio & Transcripts: Creative Commons BY-SA 4.0
- Annotations & Metadata: Creative Commons BY 4.0
- Hosting: Primary repository on Hugging Face Datasets with mirroring on Zenodo (DOI assignment for citability); permanent 50-year retention guarantee through Zenodo

- Access: Single open tier; no restricted access needed as all PII removed during annotation

Documentation & Ethics

We will release documentation following Datasheets for Datasets and Data Statements for NLP frameworks, detailing scope, collection rationale, known limitations, risk assessment, and recommended uses. Community review sessions will verify scenario appropriateness, and feedback channels will remain open post-release.

Annotation & QA

We will provide a four-tier annotation schema:

- Orthographic transcription (standard Italian conventions)
- Dialectal feature marking (phonological, lexical, morphosyntactic)
- ASR error identification (comparing human vs. machine transcription)
- Health misinformation risk rating (1-5 scale: none to severe)

For Quality Control, annotators are native dialect speakers with linguistic training; the annotation targets will be Cohen's $\kappa > 0.75$ for transcription and Krippendorff's $\alpha > 0.70$ for misinformation ratings, with weekly calibration reviews and expert checks on high-risk cases.

Sustainability

Sustainability will be ensured through a GitHub issue tracker for corrections, annual minor updates and major releases every two years, with the University of Naples Federico II guaranteeing at least 10-years of hosting and maintenance alongside the Opera del Vocabolario Italiano for long-term preservation. Post-grant support will come from institutional research funds and citation-driven grant applications. If deprecation ever becomes necessary, we will provide 12-months notice, final archival snapshot, and migration support for dependent projects.